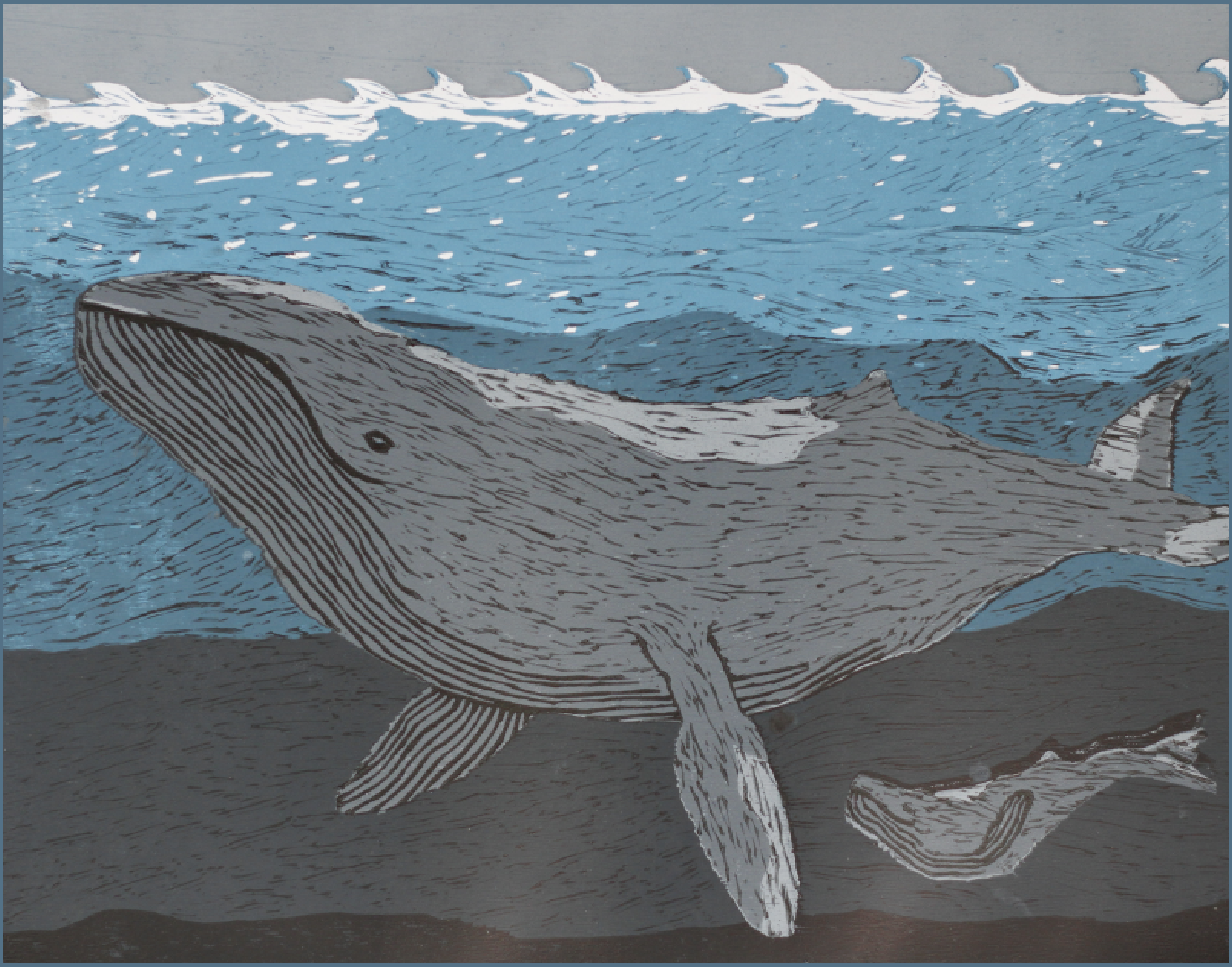
Annals of Computer Science and Information Systems
Volume 5

# Proceedings of the 2015 Federated Conference on Computer Science and Information Systems

September 13–16, 2015. Łódź, Poland



Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki
(eds.)

# Annals of Computer Science and Information Systems, Volume 5

# Proceedings of the 2015 Federated Conference on Computer Science and Information Systems

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki
(eds.)

2015, Warszawa,
Polskie Towarzystwo
Informatyczne

2015, New York City,
Institute of Electrical and
Electronics Engineers

Annals of Computer Science and Information Systems, Volume 5

Proceedings of the 2015 Federated Conference on Computer Science and Information Systems

ART: ISBN 978-83-60810-65-1, IEEE Catalog Number CFP1485N-ART
USB: ISBN 978-83-60810-67-5, IEEE Catalog Number CFP1485N-USB
WEB: ISBN 978-83-60810-66-8

ISSN 2300-5963
DOI 10.15439/978-83-60810-66-8

Al. Solidarności 82A m. 5
01-003 Warsaw
Poland

10662 Los Vaqueros Circle
Los Alamitos, CA 90720
USA

**Contact:** secretariat@fedcsis.org
`http://annals-csis.org/`

**Cover:**
Jana Waleria Denisiuk,
*Elbląg, Poland*

**Also in this series:**

Volume 7: Proceedings of the LQMR Workshop, **ISBN WEB: 978-83-60810-78-1, ISBN USB: 978-83-60810-79-8**

Volume 6: Position Papers of the 2015 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83- 60810-76-7, ISBN USB: 978-83-60810-77-4**

Volume 4: Proceedings of the E2LP Workshop, **ISBN WEB: 978-83-60810-64-4, ISBN USB: 978-83-60810-63-7**

Volume 3: Position Papers of the 2014 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-60-6, ISBN USB: 978-83-60810-59-0**

Volume 2: Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, **WEB: ISBN 978-83-60810-58-3, USB: ISBN 978-83-60810-57-6, ART: ISBN 978-83-60810-61-3**

Volume 1: Position Papers of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS), **ISBN WEB: 978-83-60810-55-2, ISBN USB: 978-83-60810-56-9**

DEAR Reader, it is our pleasure to present to you Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), which took place in Łódź, Poland, on September 13–16, 2015. Each of the papers, found in this volume, was refereed by at least two referees and the acceptance rate of full (regular) papers was 24.01% (91 papers out of 379 submissions).

FedCSIS 2015 was organized by the Polish Information Processing Society (Mazovia Chapter), Wrocław University of Economics, Systems Research Institute Polish Academy of Sciences and Łódź University of Technology. FedCSIS was organized in technical cooperation with: IEEE Computer Society, IEEE Region 8, IEEE SMC Technical Committee on Computational Collective Intelligence, Computer Society Chapter Poland, Gdańsk Computer Society Chapter, Poland, Polish Chapter of the IEEE Computational Intelligence Society (CIS), ACM Special Interest Group on Applied Computing, European Alliance for Innovation (EAI), Łódź ACM Chapter, Committee of the Computer Science of the Polish Academy of Sciences, Polish Operational and Systems Research Society, Mazovia Cluster ICT and Eastern Cluster ICT Poland. Furthermore, the 10th International Symposium Advances in Artificial Intelligence and Applications (AAIA'15) was organized in technical cooperation with: International Fuzzy Systems Association, European Society for Fuzzy Logic and Technology, International Rough Set Society and Polish Neural Networks Society.

FedCSIS 2015 consisted of the following events:

- AAIA'15—10th International Symposium Advances in Artificial Intelligence and Applications
  - AIMaVIG'15—1st International Workshop on Artificial Intelligence in Machine Vision and Graphics
  - AIMA'15—5th International Workshop on Artificial Intelligence in Medical Applications
  - ASIR'15—5th International Workshop on Advances in Semantic Information Retrieval
  - LQMR'15—1st Workshop on Logics for Qualitative Modelling and Reasoning
  - WCO'15—8th Workshop on Computational Optimization
- CSS—Computer Science & Systems
  - BCPC'15—1st International Workshop on Biological, Chemical and Physical Computations
  - CANA'15—8th Computer Aspects of Numerical Algorithms
  - IWCPS'15—2nd International Workshop on Cyber-Physical Systems
  - MMAP'15—8th International Symposium on Multimedia Applications and Processing
  - WAPL'15—5th Workshop on Advances in Programming Languages
- ECRM—Education, Curricula & Research Methods
  - DS-RAIT'15—2nd Doctoral Symposium on Recent Advances in Information Technology
- iNetSApp'15—3rd International Conference on Innovative Network Systems and Applications
  - EAIS'15—2nd Workshop on Emerging Aspects in Information Security
  - SoFAST-WS'15—4th International Symposium on Frontiers in Network Applications, Network Systems and Web Services
  - WSN'15—4th International Conference on Wireless Sensor Networks
- IT4MBS—Information Technology for Management, Business & Society
  - ABICT'15—6th International Workshop on Advances in Business ICT
  - AITM'15—13th Conference on Advanced Information Technologies for Management
  - ISM'15—10th Conference on Information Systems Management
  - IT4L'15—4th Workshop on Information Technologies for Logistics
  - KAM'15—21st Conference on Knowledge Acquisition and Management
- JAWS—Joint Agent-oriented Workshops in Synergy
  - ABC:MI'15—10th Workshop on Agent Based Computing: from Model to Implementation
  - MAS&S'15—9th International Workshop on Multi-Agent Systems and Simulations
  - SEN-MAS'15—4th International Workshop on Smart Energy Networks & Multi-Agent Systems

Furthermore, an AAIA'15 Data Mining Competition, focused on "Recognition of activities carried out by first responders at a fire scene based on body sensor network readings" has been organized. Its results constitute a separate section in these proceedings. Awards for the winners of the contest were sponsored by: Mazovia Chapter of the Polish Information Processing Society.

Each event constituting FedCSIS had its own Organizing and Program Committee. We would like to express our warmest gratitude to members of all of them for their hard work attracting and later refereeing 379 submissions.

FedCSIS 2015 was organized under the auspices of Prof. Lena Kolarska-Bobińska, Minister of Science and Higher Education, Andrzej Halicki, Minister of Administration and Digitization, Prof. Michał Kleiber, President of the Polish Academy of Sciences, Witold Stępień, Marshal of Łódź Province, Hanna Zdanowska, Mayor of the City of Łódź, and Prof. Stanisław Bielecki, Rector of Łódź University of Technology.

FedCSIS was sponsored by the Ministry of Science and Higher Eduction, Intel and Samsung.

***Maria Ganzha,*** *Co-Chair of the FedCSIS Conference Series, Systems Research Institute Polish Academy of Sciences, Warsaw, Poland, and Warsaw University of Technology, Poland*
***Leszek Maciaszek,*** *Co-Chair of the FedCSIS Conference Series, Wrocław University of Economics, Wrocław, Poland and Macquarie University, Sydney, Australia*
***Marcin Paprzycki,*** *Co-Chair of the FedCSIS Conference Series, Systems Research Institute Polish Academy of Sciences, Warsaw and Management Academy, Warsaw, Poland*

Annals of Computer Science and Information Systems, Volume 5

# Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (FedCSIS)

**September 13–16, 2015. Łódź, Poland**

## TABLE OF CONTENTS

### Conference Keynote Papers

### 10$^{TH}$ International Symposium Advances in Artificial Intelligence and Applications

## 1$^{\text{ST}}$ International Workshop on Artificial Intelligence in Machine Vision and Graphics

## 5$^{\text{TH}}$ International Workshop on Artificial Intelligence in Medical Applications

## 5<sup>TH</sup> International Workshop on Advances in Semantic Information Retrieval

## Data Mining Competition: Tagging Firefighter Activities at a Fire Scene

## 8$^{\text{TH}}$ Workshop on Computational Optimization

## Computer Science & Systems

## 1ST International Workshop on Biological, Chemical and Physical Computations

## 8TH Computer Aspects of Numerical Algorithms

## 2$^{\text{ND}}$ International Workshop on Cyber-Physical Systems

## 8TH International Symposium on Multimedia Applications and Processing

## 5$^{TH}$ Workshop on Advances in Programming Languages

## Education, Curricula & Research Methods

## 2$^{ND}$ Doctoral Symposium on Recent Advances in Information Technology

## Innovative Network Systems and Applications

## 2nd Workshop on Emerging Aspects in Information Security

## 4th International Symposium on Frontiers in Network Applications, Network Systems and Web Services

## $4^{TH}$ International Conference on Wireless Sensor Networks

## Information Technology for Management, Business & Society

## 6TH INTERNATIONAL WORKSHOP ON ADVANCES IN BUSINESS ICT

## 13TH CONFERENCE ON ADVANCED INFORMATION TECHNOLOGIES FOR MANAGEMENT

## 10TH Conference on Information Systems Management

## 4TH Workshop on Information Technologies for Logistics

## 21TH Conference on Knowledge Acquisition and Management

## 4TH Joint Agent-oriented Workshops in Synergy

## 10TH Workshop on Agent Based Computing: from Model to Implementation

## 9TH International Workshop on Multi-Agent Systems and Simulation

## 4TH International Workshop on Smart Energy Networks & Multi-Agent Systems

DOI: 10.15439/2015F002
ACSIS, Vol. 5

# On the Routing in Flying Ad hoc Networks

Md. Hasan Tareque, Md. Shohrab Hossain
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
Email: hasantareque07@gmail.com, mshohrabhossain@cse.buet.ac.bd

Mohammed Atiquzzaman
School of Computer Science
The University of Oklahoma
Norman, OK 73019, USA
Email: atiq@ou.edu

***Abstract*—The usage of Unmanned Aerial Vehicles (UAVs) is increasing day by day. In recent years, UAVs are being used in increasing number of civil applications, such as policing, fire-fighting, etc in addition to military applications. Instead of using one large UAV, multiple UAVs are nowadays used for higher coverage area and accuracy. Therefore, networking models are required to allow two or more UAV nodes to communicate directly or via relay node(s). Flying Ad-Hoc Networks (FANETs) are formed which is basically an ad hoc network for UAVs. This is relatively a new technology in network family where requirements vary largely from traditional networking model, such as Mobile Ad-hoc Networks and Vehicular Ad-hoc Networks. In this paper, Flying Ad-Hoc Networks are surveyed along with its challenges compared to traditional ad hoc networks. The existing routing protocols for FANETs are then classified into six major categories which are critically analyzed and compared based on various performance criteria. Our comparative analysis will help network engineers in choosing appropriate routing protocols based on the specific scenario where the FANET will be deployed.**

***Index Terms*—UAV networks, MANET, VANET, FANET, Routing protocols**

## I. Introduction

UNMANNED aerial vehicle (UAV) systems can fly independently or can be operated distantly. The usage of UAVs is increasing day by day. Earlier, UAVs were simple remotely piloted aircrafts and mostly used for military operations / applications. However, in recent years, UAVs are being used in increasing number of civil applications, such as policing and fire-fighting, non-military security work, etc.

The use of single-UAV system is very common, but using a group of small UAVs has become advantageous. Nonetheless, multi-UAV systems have some exclusive challenges and one of the most important design issues is the communication. There are many advantages of multi-UAV systems, such as

- Economical: The installation and maintenance cost of small UAVs are much less than that of a large UAV [1].
- Flexibility: Single UAV have limited coverage area, hence coverage rate is low [2]. However, multi-UAV systems can adapt to the situation easily.
- Continuity: If the UAV operation (operated by one UAV) fails in a mission, it cannot proceed. However, if a UAV goes off in a multi-UAV system, the operation can be survived through other UAVs.
- Faster: It has been shown that the missions can be completed faster with a higher number of UAVs [3].
- Higher accuracy: Instead of one large radar cross-section, multi-UAV systems produce very small radar cross-sections which are more accurate and crucial for military applications [4].
- Sustainable: Multi UAVs are more sustainable than single UAV system.
- Easy to solve: Multi-UAVs sometime can be solved recursively, which is much easier than single UAV system.

Multi-UAV systems have several issues. In a single-UAV system, a ground base station or a satellite is used for communication. Sometimes, communication link is established between the UAV and an airborne control system. In every case, single-UAV communication link is established between the UAV and the infrastructure. When the number of UAVs increases in the unmanned aerial systems, designing effective network architectures becomes a crucial issue.

There are some UAVs, those connect with a ground base station; others can connect to satellites, thereby realizing the UAV-to-UAV communication through the infrastructure. However, there are several design limitations with the infrastructure-based approach. First of all, each UAV must be equipped with an exclusive and complex hardware in order to communicate with a ground base station or a satellite. Reliability of the communication is the second issue. Another problem is the range restriction among the UAVs and the ground base station. If a UAV is outside the coverage area of the ground station, it becomes disconnected.

To resolve all the above mentioned issues, an alternative solution for multi-UAV systems is required to create an ad-hoc network among the UAVs, which is called FANET. In FANET, only a subset of UAVs can interconnect with the ground station or the satellite and all UAVs constitute an ad-hoc network. In this way, the UAVs can communicate with one another in addition to the ground station.

FANET is basically a special form of MANET/VANET. There are also certain differences between FANET and the traditional ad-hoc networks. Mobility degree of FANET nodes is much higher than that of MANET or VANET nodes. While typical MANET and VANET nodes are walking human beings or vehicles, respectively, FANET nodes fly in the sky. Due to high mobility of FANET nodes, the topology changes more frequently than the network topology of a typical MANET or even VANET. FANET needs peer-to-peer connections for

978-83-60810-66-8/$25.00©2015, IEEE

synchronization and relationship of UAVs. It is required to collect data from the environment and to transmit to the command & control center, as in wireless sensor networks [5]. Hence, FANET must support both peer-to-peer communication and converge cast traffic at the equivalent time. The distances among FANET nodes are much higher than in MANETs or VANETs [6]; so higher range of communication is needed. Multi-UAV systems may include different types of sensors, and each sensor may require different data distribution approaches.

There exist a few studies on UAV networks [7]–[11]. In [7], authors discussed Unmanned Aircraft System (UAS) where wireless communication is performed through IEEE 802.11b/g and Dynamic Source Routing (DSR) protocol was used as the routing protocol. In [8], a net-centric communication process is described with full command and control architecture for a heterogeneous unmanned aircraft system (in a small topology). In [9], authors discussed several networking issues related to delay-tolerant mobile ad-hoc network architecture. However, none of these works have provided a comprehensive survey of the routing issues of FANET networks. A survey [10] was performed on Flying ad-hoc network where FANET application scenario are discussed. It also discusses the FANET Communication protocols that consist of Physical, MAC, Network, Transport and Cross-layer architectures. FANET network layer has been discussed briefly in [11] where it is proposed that with a small modification on the routing protocols (used for VANET and MANET), they can be used in FANET architecture. However, none of these works [10], [11] have provided a comprehensive survey of the routing issues of FANET networks.

The main *objective* of this paper is to explain FANET as a distinct ad hoc network family and to introduce unique challenges, design constraints and routing issues in FANETs.

The *contributions* of this paper are (i) presenting different challenges and issues of FANET design, (ii) classifying existing routing protocols for FANET, and (iii) critically analyzing and comparing them based on various performance criteria.

Our comparative analysis will help network engineers in choosing appropriate routing protocols based on the specific scenario where the FANET will be deployed.

The rest of the paper is organized as follows. In Section II, we present several FANET designing issues. In Section III, we provide an extensive evaluation of the existing FANET routing protocols. In Section IV, we provide a comparative study among the six basic protocols of FANETs, followed by a discussion of open problems for FANET research in Section V. Finally, Section VI has the concluding remarks.

## II. FANET Design Issues

FANET is a new form of MANET where the nodes are UAVs. So a single-UAV systems cannot form a FANET, and valid only for multi-UAV systems. Again all multi-UAV systems do not form a FANET. UAV communication should create an ad-hoc network between UAVs to create FANET. Therefore, if the communication between UAVs fully relies on UAV-to-infrastructure links, it cannot be classified as a FANET [10].

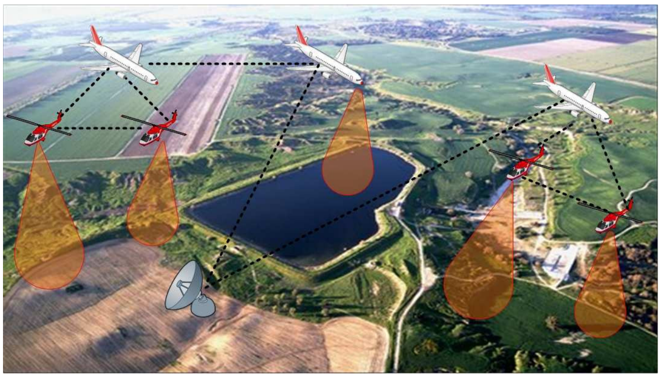


Fig. 1. A FANET scenery of multi-UAV systems.

In Fig. 1, a detailed multi-UAV system is shown. There are several area where FANET linked researches are studied under dissimilar names. Like, aerial robot team [12], aerial sensor network [13]–[15], but exact FANET base study is less interest in this topics. UAV ad hoc network [16] is totally a unique topic, which is thoroughly associated to FANETs. In fact, there is no major change between the existing UAV ad-hoc network researches and the above FANET definition. However, FANET term instantly prompts that it is a specialized form of MANET and VANET. This is why it is called Flying Ad-Hoc Network, FANET.

*FANET vs. traditional ad-hoc networks:*
Wireless ad hoc networks are categorized permitting to their application, positioning, communication and assignment intentions. By characterization, FANET is a form of MANET, and there are many mutual design thoughts for MANET and FANET. FANET can also be classified as a subset of VANET, which is also a subgroup of MANET.

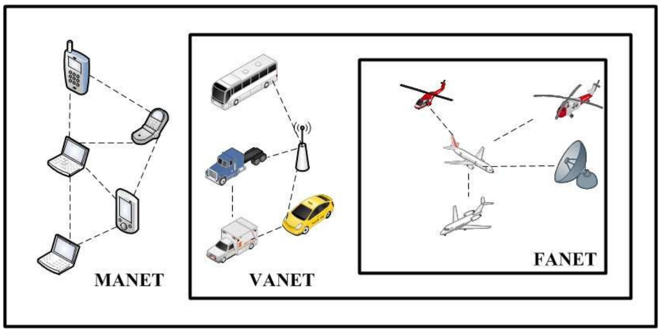


Fig. 2. MANET, VANET and FANET.

This affiliation is shown in Fig. 2. FANET shares common characteristics with these networks, and it also has some unique design challenges. Table I presents a comparison among FANET, VANET and MANET. In the following subsections, the differences among FANET and the existing wireless ad hoc networks are explained in details.

TABLE I
COMPARISON AMONG FANET, VANET AND MANET.

| Ad-Hoc network Types<br>Criteria | FANET | VANET | MANET |
|---|---|---|---|
| Node mobility | High compactness | Medium compactness | Low compactness |
| Mobility model | Usually predetermined, but special mobility models for independent multi-UAV systems | Steady | Arbitrary |
| Node density | Low thickness | Medium thickness | Low thickness |
| Topology change | Rapid and speedy | Average speed | Slow and steady |
| Radio propagation model | High above the ground level,LoS (Line of Sight) is accessible for most of the cases | Close to ground, LoS is not accessible for all cases | Very close to ground, LoS is not accessible for all cases |
| Power consumption and network lifetime | Needed for mini UAVs, but not needed for small UAVs | Not needed | Need of energy efcient protocols |
| Computational power | Very big | Average | Limited |
| Localization | GPS, AGPS, DGPS, IMU | GPS, AGPS, DGPS | GPS |

### A. Node mobility

Node mobility issues are the most significant difference between FANET and the other ad hoc networks. MANET node movement is comparatively slow when it is compared to VANET. In FANET, the nodes mobility degree is much higher than in the VANET and MANET. According to [6], a UAV has a speed of 30–460 km/h, and this situation results in several challenging communication design problems [17].

### B. Mobility model

MANET nodes move on a definite territory, VANET nodes move on the highways, and FANET nodes fly in the sky. In multi-UAV systems, the flight plan is not fixed, if a multi-UAV system uses predefined flight plans it may not be successful, because of the environmental deviations or operation updates, the flight plan may need to be recalculated.

### C. Node density

Node density is defined as the average number of nodes in a unit area. FANET nodes are normally spread in the sky, and the distance between UAVs can be several kilometers even for small multi-UAV systems. As a result of this, FANET node density is much lower than in the MANET and VANET.

### D. Topology change

Due to higher mobility degree, FANET topology changes more regularly than MANET and VANET topology. When a UAV fails, the links that the UAV has been involved in also failed and it results in a topology update. Another factor that affects the FANET topology is the link outages. Because of the UAV schedules and variations of FANET node distances, link quality changes very quickly, and it also causes link outages and topology changes [18].

### E. Radio propagation model

FANET and the other ad hoc network operating environments affect the radio propagation characteristics. MANET and VANET nodes are very close to the ground, and in many cases, there is no line of-sight between the sender and the receiver. Radio signals are mostly affected by the geographic structure. Again, FANET nodes those are away from the ground can be driven remotely and in maximum case; there is a line-of sight between UAVs [10].

### F. Power consumption and network lifetime

Developing energy efficient communication protocols is a major part to increase the network lifetime. Particularly, while the battery-powered computing devices in MANETs; system developers have to pay extra attention to the energy efficient communication protocols. However, FANET communication hardware is powered by the energy source of the UAV. This means FANET communication hardware has no power resource problem as like in MANET.

### G. Computational power

MANET nodes are battery powered small computers such as laptops, PDAs and smart phones. Because of the size and energy constraints, the nodes have only limited computational power. On the other hand VANETs and FANETs support devices with high computational power.

### H. Localization

In MANET, GPS is generally used to receive the coordinates of a mobile communication terminal, and maximum time, GPS is enough to regulate the location of the nodes. In VANET, for a navigation-grade GPS receiver, there is about 10–15 m accuracy, which can be satisfactory for route guidance. Because of the high velocity and dissimilar mobility models of multi-UAV systems, FANET needs highly accurate localization data with smaller time intervals. GPS provides position information at one second interval, and it may not be adequate for certain FANET protocols.

## III. FANET NETWORKING PROTOCOLS

There exists many routing protocols for wireless and ad-hoc networks, such as pre-computed routing, dynamic source routing, on demand routing, cluster based routing, flooding, etc. FANET is a sub-class of VANET and MANET networks. Therefore, MANET routing protocols are initially chosen and tested for FANET. Due to the UAV-specific issues, such as, rapid changes in link quality, most of these protocols are not directly applicable for FANET networks. Hence, in order to implement this new networking model, some definite adhoc networking protocols have been implemented and some previous ones have been adapted. These protocols can be classified into six major categories:

- **Static protocols**, having fixed routing tables (no need to refresh these tables).
- **Proactive protocols**, have periodically refreshed routing tables.
- **Reactive protocols** (also called on-demand protocols) discover paths for messages on demand.
- **Hybrid protocols** that use both proactive and reactive protocols.
- **Position/Geographic Based protocols** that use position or area coverage.
- **Hierarchical protocols** that use hierarchy model for routing.

With the help of all these routing protocols, a FANET can actively discover new paths among the communicating nodes.

## A. Static Routing Protocols

In static routing protocol, a routing table is calculated and uploaded to UAV nodes before an assignment, and cannot be updated during the operation; this is why it is called static. UAVs in this protocol have a constant topology [19]. Here every node communicates with a limited numbers of UAVs or ground stations, and it only preserves their information. In case of a failure (of a UAV or ground station), for updating the tables, it is essential to wait until the assignment is finished. As a result there are no fault tolerant approach for dynamic environments in static routing protocols.

*1) Data Centric Routing:* UAVs wireless communication support one to many data transmission which is similar to one-to-one data transmission [20], [21]. This method is selected when the data is requested by a number of nodes, and data distribution is done by on-demand algorithms. Data-centric routing is a favorable model of routing mechanism and can be adjusted for FANET [22], [23].

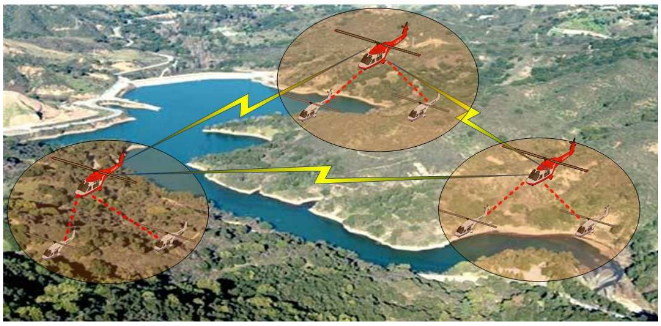


Fig. 3. Data centric routing model in FANET.

Data demand and gathering are done by data attributes instead of sender and receiver nodes' IDs. As shown in Fig. 3, this model is usually skillful with cluster infrastructure.

In this model, the consumer node (either ground node or a UAV) broadcasts queries (such as ”get video of area X if there is a change of more than % 3”) as contribution message in order to collect particular data from a precise area. Routing is done with respect to the content of data. Data aggregation algorithms may be used for energy efficient data broadcasting. This routing executes three scopes of decoupling:

- *Space decoupling*: Communicating parties can be anywhere.
- *Time decoupling*: Data can be transmitted to the subscribers instantly or later.
- *Flow decoupling*: Delivery can be accomplished constantly.

This model can be chosen when the system contains a small number of UAVs on a planned path, which involves minimum assistance.

*2) Multi-Level Hierarchical Routing:* Organizing UAV networks hierarchically a number of clusters needs to operate in different mission areas, as shown in Fig. 4. Each cluster has a cluster head (CH), which will represent the whole cluster; this separate cluster can perform different activities. Each CH is in connection with the upper/lower layers (ground stations, UAVs, satellites, etc.) directly or indirectly. To broadcast data and control info to other UAVs in the cluster, CH should be in direct communication range of other UAVs in cluster.

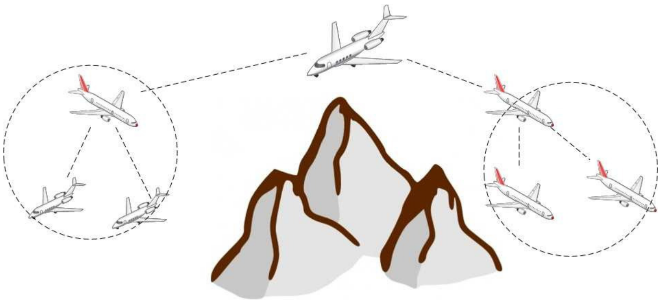


Fig. 4. Multi-Level Hierarchical routing model in FANET

This model is better if UAVs are controlled in changed swarms, the mission area is huge, and several UAVs are used in the network.

*3) Load Carry and Deliver Routing:* In this model, a UAV loads data from a ground node; then the data is being carried to the destination by flying; and at the end the data reached to the destination ground node.

The main objectives of load carry and deliver routing is to maximize throughput and increase the security. But the main drawback of this protocol is whenever the distance of communicating parts growth, the transmission delay becomes tremendously huge and unendurable. To solve this problem multi-UAVs system can be developed so that it decreases transmission delay as well as the distance among UAVs.

## B. Proactive Routing Protocols

Proactive routing protocols (PRP) use tables to store all the routing in the network. The main advantage of proactive routing is that it stores the latest information of the routes; therefore, it is easy to choose a route from the sender to the receiver, as a result transmission delay can be minimized. But, there are some disadvantages of this protocol.

$First$, there are lot of messages are being exchange between nodes, therefore bandwidth optimization is not possible. For this reason PRPs are not suitable for highly mobile and/or larger networks. $Second$, when the topology change or connection failure occurs, PRP shows a slow reaction.

*1) Directional Optimized Link State Routing:* This protocol is based on the well known Optimized Link State Routing Protocol (OLSR) [24]. One of the most important factors that affect the OLSR performance is to select multipoint relay (MPR) nodes. The sender node selects a set of MPR nodes so that the MPR nodes can cover two hop neighbors. One of the most crucial design issues for OLSR is the number of MPRs, which effects the delay dramatically. Simulation studies [25] showed that DOLSR can reduce the number of MPRs with directional antennas.

*2) Destination Sequenced Distance Vector:* This protocol mainly uses the Bellman Ford algorithm with slight modifications for ad hoc networks. In DSDV, each node saves a routing table (with sequence number) for all other nodes, not just for the neighbor nodes [26]. Whenever the topology of the network changes, these changes are circulated by the protocol to update devices. To eliminate routing-loops and to identify the latest route, DSDV uses sequence numbers, which are allocated by destination nodes. The route which has higher sequence number is selected.

The main advantages of DSDV are easy algorithm and the usage of sequence numbers, which guaranteed the protocol to be loop free. Again, it has some drawbacks. For an upto date routing table, each node periodically broadcast routing table updates, which brings overhead to the network.

*3) Topology Broadcast based on Reverse-Path Forwarding:* This protocol use Dynamic Source Routing (DSR) [27]. The main advantage to choose DSR is its reactive configuration. The source tries to find a path to a destination, only if it has data to send. Main drawback of this protocol is the topology is unstable when the nodes are highly mobile.

### C. Reactive Routing Protocols

Reactive Routing Protocol (RRP) can be referred as on demand routing protocol. If there is no connection between two nodes, there is no need to calculate a route between them. The concept RRP is came to overcome the overhead problem of PRP.

There are two different messages in this protocol: Route_Request messages and Route_Reply messages. Route_Request messages are created and transmitted by flooding to the network by the source node, and the destination node responses to this message with a Route_Reply message. RRP is bandwidth efficient, because there is no periodic messaging. Main drawback is it takes long time to find the route; as a result high latency may occur.

*1) Dynamic Source Routing:* Dynamic Source Routing (DSR) is designed for wireless mesh networks [28]. In DSR, the source node broadcasts a route request message to its neighbor nodes. In the entire communication route, there can be many route request messages. So, to avoid mix-up, the source node added a distinctive request_id. If the source node is not capable to use its present route (changes in the network topology), then the route repairs mechanism is triggered. This routing protocol was implemented by Brown et al. in [29] and they found finding a new route in UAV network with DSR can be frustrating.

*2) Ad-hoc On-demand Distance Vector:* Ad-hoc On-demand Distance Vector (AODV) has almost the similar on-demand features with DSR. The only difference is routing table maintenance [30].

In DSR each node can store multiple entries in the routing table for each destination while in AODV; there is a single record for each destination. Another difference in DSR, the data packets transfer the complete path between source and the destination nodes. But in AODV, the source node only stores the next-hop information consistent to each data communication.

AODV routing protocol consists of $three$ phases: route discovery, packet transmitting and route maintaining.

*3) Time-slotted on-demand Routing:* Time-slotted on-demand routing protocol is proposed in [31] for FANETs. Basically it is time-slotted version of AODV. Time-slotted on-demand protocol uses dedicated time slots in which only one node can send data packet. Although it increases the use of network bandwidth but mitigates the packet collisions and ensure packet delivery.

### D. Hybrid Routing Protocols

To overcome the limitations of previous protocol Hybrid routing protocol (HRP) is introduced. Reactive routing protocols needs extra time to discover route and proactive routing protocols has huge overhead of control messages both can mitigate in HRP. HRP is appropriate for large networks. A network can be divided into a number of zones where intra-zone routing used proactive method while inner-zone routing uses reactive method [11].

*1) Zone Routing Protocol:* Zone Routing Protocol (ZRP) is based on the concept of zones [32]. In this protocol, each node has a different zone. The zone is defined as the set of nodes whose minimum distance is predefined radius $R$. So, the zones of neighboring nodes intersect. The routing inside the zone is called as intra-zone routing, and it uses proactive method. If the source and destination nodes are in the same zone, the source node can start data communication instantly. When the data packets need to send outside the zone the inter-zone routing is used and reactive method is applied.

*2) Temporarily Ordered Routing Algorithm:* Temporarily Ordered Routing Algorithm (TORA) routers only preserve info about adjacent routers [33]. TORA mainly uses a reactive routing protocol but it also use some proactive protocol. It constructs and preserves a Directed Acyclic Graph (DAG) from the source node to the destination. TORA does not use a shortest path solution, sometime longer routes used to reduce network overhead. Each node has a parameter value termed as "height" in DAG, which is unique for each node. Data flow as a fluid from the higher nodes to lower. It is structurally loop-free because data cannot flow to the node that has higher value [11].

### *E. Position/Geographic Based Routing Protocols*

Position-based routing needs information about the physical position of the contributing nodes in the network. Generally, each node calculates its own location through the use of GPS or some other type of positioning facilities. Position based routing is primarily motivated by two subjects, $(i)$ A position facility is used by the sender of a packet to decide the position of the destination and $(ii)$ A forwarding approach used to forward the packets.

*1) Greedy Perimeter Stateless Routing:* Greedy Perimeter Stateless Routing (GPSR) is a position based protocol, have several advantages over proactive and reactive routing protocols. Shirani et al. developed a simulation framework to study the position-based routing protocols for FANETs [34]. The outcome of the study is that GPSR can be used for compactly positioned FANETs. But, reliability is the major issue of this protocol. For this another method, like face routing, can be used to achieve more reliability.

*2) Geographic Position Mobility Oriented Routing:* The traditional position-based protocols only depend on the location information of the nodes. But, geographic position mobility oriented routing predicts the movement of UAVs with $Gaussian - Markov$ mobility model, and uses this information to guess the next hop.

### *F. Hierarchical Routing Protocols*

In hierarchical routing protocols the choice of proactive and of reactive routing depends on the hierarchic level. The routing is primarily established with some proactive planned routes and then helps the request from by triggered nodes through reactive protocol at the lower levels. The main drawbacks of this protocol are: complexity and addressing scheme which response to traffic request as a result it hang the interconnecting factors.

*1) Mobility prediction clustering:* It operates on the dictionary of Trie-structure calculation algorithm and link termination time mobility model to guess network topology updates. In this way, it can build more constant cluster formations [35].

*2) Clustering algorithm of UAV networking:* It constructs the clusters on the ground, and then updates the clusters through the mission in the multi-UAV system [36].

## IV. Comparative study

As mentioned earlier, there exist six basic protocols for FANET. In this section, we critically analyze and compare these basic FANET protocols. Table II presents the comparative study among these six FANET routing protocols: static, proactive, reactive, hybrid, position/geographic based and hierarchical protocols. We explain each of the comparison criteria in more details in this section:

### *A. Main Idea*

The main idea for static protocol is routing information is fixed for a specific mission and loaded into the UAVs before the mission. Proactive protocol stores the current route information into the table. Reactive protocol is on demand protocol; when the source asks for destination route, it calculates the route. Hybrid protocol is a combination of both proactive and reactive protocols. Position/geographic protocol uses GPS or other location service to calculate the route. Hierarchical protocols uses hierarchy model to find route.

### *B. Complexity*

For static protocol, complexity is relatively low because destination is fixed. However, for proactive, reactive and hybrid protocol, complexity is medium. In case of topology change, route finding becomes more complex in proactive protocol. For position-based protocol, finding route becomes difficult if the location service is poor. In an urban area, hierarchical protocol is useful but its setup is not so simple.

### *C. Route*

In case of static protocol, route is fixed throughout the mission. For all other protocols, routes are dynamic.

### *D. Topology size*

Static protocol is used for fixed mission. As a result, if the topology size is large, there is chance of topology change. Hence, static protocol is suitable for small networks. Proactive protocol is a table-driven protocol; hence, if the number of UAVs increases, their corresponding routing table entries also increases. Thus, proactive protocol is suited for small networks. For hybrid protocols, intra-zone routing is usually fixed and small sized. Position based and hierarchical protocols are used in larger network.

### *E. Memory size*

In static protocols, the whole routing information is uploaded into the UAVs before the mission. As a result, it requires large memory space. If the number of nodes increases, the table size grows larger. Thus, proactive protocol requires larger memory. Reactive protocol is source driven; hence, when source is required to find route, it is activated, requiring less memory. Position-based protocol caches the coordinates of each UAV, thereby requireing large memory space. Hierarchical protocol uses hierarchical structure whose memory requirement is low.

### *F. Fault tolerant*

In FANETs, mission route or topology change is a very common scenario. However, static protocols do not support this scenario. Therefore, fault tolerance is absent in this protocol. However, every other protocol has some fault tolerance.

TABLE II
COMPARISONS AMONG THE BASIC ROUTING PROTOCOLS IN FANETS.

| Criteria \ Different Protocol Types | Static Protocols | Proactive Protocols | Reactive Protocols | Hybrid Protocols | Position/Geographic Based Protocols | Hierarchical Protocols |
|---|---|---|---|---|---|---|
| Main Idea | Static routing table | Table driven protocols | On demand protocol | Combination of proactive and reactive protocols | Position-based protocol | Protocol maintained through hierarchy |
| Complexity | Low | Medium | Average | Average | High | High |
| Route | Static | Dynamic | Dynamic | Dynamic | Dynamic | Dynamic |
| Topology size | Small network | Small network | Large network | Both small and large network | Large network | Large network |
| Memory size | High | High | Low | Medium | High | Low |
| Fault tolerant | Absent | Present | Present | Mostly present | Present | Present |
| Bandwidth Utilization | Maximum | Minimum | Maximum | Medium | Minimum | Maximum |
| Convergence Time | Fast | Slow | Mostly fast | Average | Average | Average |
| Signalling Overhead | Absent | Present | Present | Present | Present | Present |
| Communication Latency | Low | Low | High | High | Low | High |
| Mission Failure Rate | High | Low | Low | Very low | Very low | Very low |
| Popularity | Less | Medium | Medium | High | Less | High |
| Application | Fixed mission | Dynamic mission | Dynamic mission | Dynamic mission | Dynamic mission | Dynamic mission |

### G. Bandwidth utilization

Static protocols are used in small network where topology is fixed; as a result, bandwidth utilization is high in this protocol. Proactive protocols have to send hello messages periodically in the network. Therefore, this protocol requires more bandwidth. Reactive protocols are source driven, requiring less bandwidth. For hybrid protocols, bandwidth utilization is medium. Position based protocol send source location as extra information; hence, bandwidth consumption rate is higher. Hierarchical protocols use limited bandwidth as each UAV is connected to upper level UAVs.

### H. Convergence time

In the static protocol, destination is predetermined. As a result route finding time is minimal. Proactive protocol searches the destination node after each topology change, resulting in larger converge time. Reactive protocols usually find route much faster but if topology changes, this protocol takes more time than normal case. Hybrid, position based and hierarchical protocols usually take average time to converge.

### I. Signalling overhead

Other than static protocols, each protocol (proactive, reactive, hybrid, position based and hierarchical) have signaling overheads, such as hello message in proactive protocols, route request and route reply message in reactive protocols, etc.

### J. Communication latency

Static, proactive and position based protocols have low communication latency since the distance between the UAVs in these protocols is small. Reactive, hybrid and hierarchical protocols have higher latency because UAV-to-UAV and UAV-to-ground station distance is much higher in these protocols.

### K. Mission failure rate

Topology and route change are common phenomena in FANETs.Other than static protocols, each protocol has backup strategy for topology change. Only static protocols do not have any strategy when topology or route changes, as a result mission failure rate is very high in this protocol.

### L. Popularity

Static protocols is not fault tolerant and position based protocols need extra mechanism to find the positions of the UAVs. This is why, these two protocols are least popular. Rest of the protocols are much more popular.

### M. Applications

Static protocols are used in missions where mission objective and topology are fixed. Hierarchical protocols are mostly used in military operations where communication is difficult. Previously, most of the protocols were used in military operations. However, use of UAVs have increased day by day. As a result, many civil operations are now conducted by multi-UAVs systems. For this reason, all the protocols are being modified so that these protocols can be used in normal and civil operations.

## V. OPEN RESEARCH PROBLEMS

Existing MANET and VANET routing protocols cannot satisfy all the FANET routing requirements. Therefore, routing is one of the most important and challenging issues for FANETs. In this section, we list a few open research issues regarding routing in flying ad hoc networks.

### A. P2P UAV communications

In FANET, movements of UAVs are very fast, resulting in very rapid network topology change. Hence, data routings among the UAVs are challenging. The routing protocols should be capable of updating routing tables dynamically. Peer-to-peer communication is crucial for cooperative synchronization and collision anticipation of multi-UAV systems. FANET can collect information from the environment as in wireless sensor networks, which is a different traffic configuration. Developing a peer-to-peer communication and converge cast traffic can be an attractive topic in FANETs. Data centric routing for FANETs is another encouraging approach which is still unexplored.

### B. Regulations for civilian UAVs

The uses of UAVs are increasing day by day and now it has become a part of most of countries national airspace system. However, most of the existing air principles do not allow UAV operations in civil airspace. This is the biggest obstruction to the development of Unmanned Arial Systems in civilian areas. As a result, distinctive rules and guidelines to integrate UAV flights into the national airspace need to be deployed urgently.

### C. Robust FANET algorithms

In a large area mission and multi-UAV operations, dynamic changes (such as addition / deletion of UAVs, fixed and dynamic threats, etc.) can occur. Therefore, robust algorithms with dynamic route adjustments are compulsory to coordinate the fleets of UAVs. It is essential for FANET to support qualities of services (QoS) so that it can protect against some predetermined service performance constraints, such as delay, bandwidth, jitter, packet loss, etc.

### D. UAV placement

The sizes of mini-UAVs are small and they carry limited payloads, for example single radar, infrared camera, thermal camera, image sensor, etc. Therefore, different sensors can be merging-up with different UAVs; or one UAV can be integrated with a thermal camera and another with image sensor. Regarding this, UAV placement to reduce energy consumption is still an open issue.

### E. FANET standardization

FANET uses various wireless communication bands, such as, VHF, UHF, L-Band, C-band, Ku-Band, etc. [11] which are also used in different application areas, such as GSM networks, satellite communications. For reducing congestion problem, FANETs require standardization. FANET should connect to integrate with a Global Information Grid (GIG) as one of the main information platforms to increase its efficiency.

### F. Coordination of UAVs with Manned aircraft

In the future, flights of UAVs with other manned aircraft are likely to increase. Coordination of these two will ensure the destruction of opponent aircraft with minimal losses. Therefore, the association of UAVs and manned aircraft should be in a networked environment.

## VI. Conclusion

Unmanned Aerial Vehicles have promising role in a large operation zone with complicated missions. For the region that are reasonably isolated from the ground and to accomplish complex tasks, UAVs require cooperation with one another and need a quick and easy deploying network system. Multi-UAV system reduces the operation accomplishment time and increases reliability of the system for airborne operations when compared to a single-UAV system. To apply networking in non-LOS, urban, aggressive, and noisy environment, multi-UAV system is very effective and accurate.

Communication is one of the most challenging issues for multi-UAV systems. In this paper, ad hoc networks among the UAVs, i.e, FANETs are surveyed along with its key challenges compared to traditional ad hoc networks. The existing routing protocols for FANETs are classified into six major categories which are then critically analyzed and compared based on various performance criteria. Finally, we list several open research issues related to FANET routing protocols to inspire researchers work on these open problems.

## References

[1] H. Chao, Y. Cao, and Y. Chen, “Autopilots for small fixed-wing unmanned air vehicles: a survey,” *International Conference on Mechatronics and Automation, 2007 (ICMA 2007)*, pp. 3144–3149, 2007.

[2] B. Morse, C. Engh, and M. Goodrich, “Uav video coverage quality maps and prioritized indexing for wilderness search and rescue,” *Proceedings of the 5th ACM/IEEE International Conference on HumanRobot Interaction, HRI 10, Piscataway, NJ, USA*, vol. 3, pp. 227–234, 2010.

[3] E. Yanmaz, C. Costanzo, C. Bettstetter, and W. Elmenreich, “A discrete stochastic process for coverage analysis of autonomous uav networks,” *Proceedings of IEEE Globecom-WiUAV, IEEE*, 2010.

[4] L. To, A. Bati, and D. Hilliard, “Radar cross-section measurements of small unmanned air vehicle systems in non-cooperative field environments,” *3rd European Conference on Antennas and Propagation, 2009 (EuCAP 2009), IEEE*, pp. 3637–3641, 2009.

[5] M. Rieke, T. Foerster, and A. Broering, “Unmanned aerial vehicles as mobile multi-platforms,” *The 14th AGILE International Conference on Geographic Information Science,Utrecht, Netherlands*, 18-21 April 2011.

[6] J. Clapper, J. Young, J. Cartwright, and J. Grimes, “Unmanned systems roadmap,” *Tech. rep., Dept. of Defense*, pp. 2007–2032.

[7] T. Brown, B. Argrow, E. Frew, C. Dixon, D. Henkel, J. Elston, and H. Gates, “Experiments Using Small Unmanned Aircraft to Augment a Mobile Ad Hoc Network,” ISBN-13: 9780521895842, pp. 179–199.

[8] J. Elston, E. Frew, D. Lawrence, P. Gray, and B. Argrow, “Net-centric communication and control for a heterogeneous unmanned aircraft system,” *Journal of Intelligent and Robotic Systems*, vol. 56(1-2), pp. 199–232, 2009.

[9] E. Frew and T. Brown, “Networking issues for small unmanned aircraft systems,” *Journal of Intelligent and Robotics Systems*, vol. 54 (1-3), pp. 21–37, 2009.

[10] I. Bekmezci, O. K. Sahingoz, and S. Temel, “Flying ad-hoc networks (FANETs): A survey,” *Elsevier, Ad Hoc Networks 11*, pp. 1254–1270, 2013.

[11] O. K. Sahingoz, “(FANETs): Concepts and challenges,” *Springer J Intell Robot System*, vol. 74, pp. 513–527, 2014.

[12] S. Cameron, S. Hailes, S. Julier, S. McClean, G. Parr, N. Trigoni, M. Ahmed, G. McPhillips, R. de Nardi, J. Nie, A. Symington, L. Teacy, and S.Waharte, “SUAAVE: Combining aerial robots and wireless networking,” *25th Bristol International UAV Systems Conference*, 2010.

[13] A. Purohit and P. Zhang, “SensorFly: a controlled-mobile aerial sensor network,” in *ACM,7th ACM Conference on Embedded Networked Sensor Systems, SenSys ’09*, New York, NY, USA, 2009, pp. 327–328.

[14] M. Akbas and D. Turgut, “APAWSAN: actor positioning for aerial wireless sensor and actor networks,” in *36th Conference on Local Computer Networks, LCN ’11, IEEE Computer Society*, Washington, DC, USA, 2011, pp. 563–570.

[15] J. Allred, A. Hasan, S. Panichsakul, W. Pisano, P. Gray, J. Huang, R. Han, D. Lawrence, and K. Mohseni, “Sensorock: an airborne wireless sensor network of micro-air vehicles,” *ACM, 5 th International Conference on Embedded Networked Sensor Systems*, pp. 117–119, 2007.

[16] T. Brown, S. Doshi, S. Jadhav, and J. Himmelstein, “Test bed for a wireless network on small UAVs,” *AIAA 3rd Unmanned Unlimited Technical Conference*, pp. 20–23, 2004.

[17] Z. Han, A. Swindlehurst, and K. Liu, “Optimization of MANET connectivity via smart deployment/movement of unmanned air vehicle,” *IEEE Transactions on Vehicular Technology*, vol. 58, pp. 3533–3546, 2009.

[18] E. Yanmaz, R. Kuschnig, and C. Bettstetter, “Channel measurements over 802.11a-based UAV-to-ground links,” *GLOBECOM Wi-UAV Workshop*, pp. 1280–1284, 2011.

[19] A.Franchi, C.Secchi, M.Ryll, H. Bulthoff, and P.R.Giordano, “Shared control: Balancing autonomy and human assistance with a group of quad rotor UAVs,” *IEEE Robot. Auto Mag*, vol. 19 (3), pp. 57–58, 2012.

[20] J.Ko, A.Mahajan, and R.Sengupta, “A network-centric UAV organization for search and pursuit operations,” *IEEE Aerospace Conference*, pp. 2697–2713, 2002.

[21] J.Lopez, P.Royo, E.Pastor, C.Barrado, and E. maria, “A middleware architecture for unmanned aircraft avionics,” *ACM/IFIP/ USENIX International Conference on Middleware companion (MC 07)*, 2007.

[22] E. D. Jong, “Flexible data-centric UAV platform eases mission adaptation,” *White paper:http://www.rti.com/whitepapers/RTI-Data-Driven-Approach-to-UAV.pdf*, 3 Aug 2013.

[23] A. A. Koller and E.N.Johnson, “Design, implementation, and integration of a publish/subscribe-like multi- UAV communication architecture,” *AIAA Modelling and Simulation Technologies Conference and Exhibit*, pp. 1–17, 2005.

[24] T. Clausen and P. Jacquet, “Optimized link state routing protocol (OLSR),” RFC 3626 (Experimental), October 2003.

[25] A. Alshabtat, L. Dong, J. Li, and F. Yang, “Low latency routing algorithm for unmanned aerial vehicles ad-hoc networks,” *International Journal of Electrical and Computer Engineering*, vol. 6 (1), pp. 48–54, 2010.

[26] D. Jung and P. Tsiotras, “Inertial attitude and position reference system development for a small uav,” *26th AIAA Aeroacoustics Conference*, 2007.

[27] D. Johnson and D. Maltz, “Dynamic source routing in ad hoc wireless networks,” *Mobile Computing, The Kluwer International Series in Engineering and Computer Science, Springer, US*, vol. 353, pp. 153–181, 1996.

[28] D.B.Johnson and D.A.Maltz, “Dynamic source routing in ad hoc wireless networks,” *Kluwer Academic Publishers*, pp. 153–181, 1996.

[29] T.X.Brown, B.Argrow, C.Dixon, S.Doshi, R.G.Thekkekunel, and D.Henkel, “Ad-hoc UAV ground network (AUGNet),” *3rd AIAA Unmanned Unlimited Technical Conference*, pp. 29–39, 2004.

[30] S.Murthy and J. L. Aceves, “An efficient routing protocol for wireless networks,” *ACM Mobile Networks and Applications*, pp. 183–197, 1996.

[31] J. Forsmann, R. Hiromoto, and J. Svoboda, “A time-slotted on-demand routing protocol for mobile ad-hoc unmanned vehicle systems,” SPIE 6561, 2007.

[32] Z.J.Haas and M.R.Pearlman, *Zone Routing Protocol (ZRP) a hybrid framework for routing in ad-hoc networks*. Addison-Wesley, 2001, vol. 1.

[33] V. Park and S.Corson, “Temporarily-ordered routing algorithm (TORA),” *Version 1. Internet draft:IETF MANET working group. http://tools.ietf.org/html/ draft-ietf-manet-tora-spec-04*, 3 Aug 2013.

[34] R. Shirani, M. St-Hilaire, T. Kunz, Y. Zhou, J. Li, and L. Lamont, “The performance of greedy geographic forwarding in unmanned aeronautical ad-hoc networks,” in *2011 Ninth Annual Communication Networks and Services Research Conference, CNSR ’11, IEEE Computer Society*, Washington, DC, USA, 2011, pp. 161–166.

[35] C. Zang and S. Zang, “Mobility prediction clustering algorithm for UAV networking,” in *GLOBECOM Workshops, IEEE*, 2011, pp. 1158–1161.

[36] L. Kesheng, Z. Jun, and Z. Tao, “The clustering algorithm of UAV networking in near-space,” in *8th International Symposium on Antennas, Propagation and EM Theory,(ISAPE 2008)*, 2008, pp. 1550–1553.

DOI: 10.15439/2015F003
ACSIS, Vol. 5

# The Cognitive Cycle

John F. Sowa

*Abstract*—**In the twenty years from first grade to a PhD, students never learn any subject by the methods for which machine-learning algorithms have been designed. Those algorithms are useful for analyzing large volumes of data. But they don't enable a computer system to learn a language as quickly and accurately as a three-year-old child. They're not even as effective as a mother raccoon teaching her babies how to find the best garbage cans. For all animals, learning is integrated with the cognitive cycle from perception to purposeful action. Many algorithms are needed to support that cycle. But an intelligent system must be more than a collection of algorithms. It must integrate them in a cognitive cycle of perception, learning, reasoning, and action. That cycle is key to designing intelligent systems.**

## I. Theories of Learning and Reasoning

THE nature of the knowledge stored in our heads has major implications for educating children and for designing intelligent systems. Both fields organize knowledge in teachable modules that are presented in textbooks and stored in well structured databases and knowledge bases. A systematic organization makes knowledge easier to teach and to implement in computer systems. But as every student discovers upon entering the workforce, "book learning" is limited by the inevitable complexities, exceptions, and ambiguities of engineering, business, politics, and life. Although precise definitions and specifications are essential for solving problems in mathematics, science, and engineering, most problems aren't well defined. As Shakespeare observed, "There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy."

During the past half century, neuroscience has discovered a great deal about the organization of the human brain and its inner workings. But each discovery has led to far more questions than answers. Meanwhile, artificial intelligence developed theories, tools, and algorithms that have been successfully applied to practical applications. Both neuroscience and AI have been guided by the centuries of research in other branches of cognitive science: philosophy, psychology, linguistics, and anthropology.

One explanation of learning that has been invented and reinvented in various guises is the *apperceptive mass* or dominant system of ideas. It was suggested by Leibniz, elaborated by Johann Friederich Herbart [5], and had a strong influence on educational psychology [31]:

> There is a unity of consciousness — attention, as one might call it — so that one cannot attend to two ideas at once except in so far as they will unite into a single complex idea. When one idea is at the focus of the consciousness it forces incongruous ideas into the background or out of consciousness altogether. Combined ideas form wholes and a combination of related ideas form an apperceptive mass, into which relevant ideas are welcomed but irrelevant ones are excluded… If information is to be acquired as easily and as rapidly as possible, it follows that in teaching one should introduce new material by building upon the apperceptive mass of already familiar ideas. Relevant ideas, then, will be most easily assimilated to the apperceptive mass, while irrelevant ideas will tend to be resisted and, consequently, will not be assimilated as readily.

In AI, versions of semantic networks from the 1960s [20] to the 1990s [8] resemble Herbart's apperceptive mass. In fact, the implementations exhibit many of the properties claimed by educational psychologists [26]. Piaget and his colleagues in Geneva observed children and analyzed the *schemata* they used at various ages [19]. They showed that learning progresses by stages, not by a series of database updates. At each stage, the brain *assimilates* new information to its existing structures. When the complexity of the information grows beyond the capacity of the structures at one stage, a minor revolution occurs, and a new schema is created to reorganize the mental structures. The later, more abstract *conceptual* schemata are derived by generalizing, building upon, and reorganizing the *sensorimotor* schemata formed in infancy and early childhood.

Rumelhart and Norman [23] used the term *accretion* for Piaget's assimilation of new information to the old schemata. They split Piaget's notion of *accommodation* in two stages of *tuning* and *restructuring*:

- *Accretion.* New knowledge may be added to episodic memory without changing semantic memory. It corresponds to database updates that state new facts in terms of existing categories and schemata.
- *Schema tuning.* Minor changes to semantic memory may add new features and categories, generalize old schemata to more general supertypes, and revise priorities or defaults.
- *Restructuring.* As episodic memory becomes more complex, a major reorganization of semantic memory may be needed. Completely new schemata and concept types may represent old information more efficiently, support complex deductions in fewer steps, and make drastic revisions to old ways of thinking and acting.

Restructuring is responsible for the *plateau effect*: people quickly learn a new skill up to a modest level of proficiency; then they go through a period of stagnation when study and practice fail to show a noticeable improvement; suddenly,

978-83-60810-66-8/$25.00©2015, IEEE

they break through to a new plateau where they again progress at a rapid rate — until they reach the next plateau. Restructuring is the creative discovery. When a person attains a new insight, a sudden revolution reorganizes the old information in the new categories and schemata.

The paradigms that psychologists have proposed for human learning have their counterparts in AI. In 1983, a review of machine learning [3], distinguished three historical periods, each characterized by its own paradigm for learning:

- *Stimulus-response*. In the late 1950s and early 60s, *neural nets* and *self-organizing systems* were designed to start from a neutral state and build up internal connections solely by reinforcement and conditioning.
- *Induction of concepts*. A higher form of learning is the induction of new types of categories from the data. It started in the 1960s with clustering and concept learning techniques, and it can be integrated with every learning system, formal or informal.
- *Knowledge-intensive learning*. Before a system can learn anything new, it must already have a great deal of initial knowledge. Most methods of restructuring are compatible with knowledge-intensive learning.

One of the first steps in restructuring is categorization: selecting new concept types, mapping percepts to those types, and assigning the types to appropriate levels of the type hierarchy. A more radical alteration of schemata is a leap into the unknown. It requires someone to abandon comfortable old ways of thought before there is any reassurance that a new schema is better. To explain how such learning is possible, Charles Sanders Peirce [17] proposed the reasoning method of *abduction*, which operates in a *cognitive cycle* with observation, induction, abduction, deduction, testing, action — and repeat:

- Deduction is logical inference by formal reasoning or plausible reasoning.
- Induction involves gathering and generalizing new data according to existing types and schemata.
- Abduction consists of a tentative guess that introduces a new type or schema, followed by deduction for exploring its consequences and induction for testing it with reality.

An abduction is a hunch that may be a brilliant breakthrough or a dead end. The challenge is to develop a computable method for generating insightful guesses. Beforehand, a guess cannot be justified logically or empirically. Afterwards, its implications can be derived by deduction and be tested against the facts by induction. Peirce's logic of pragmatism integrates the reasoning methods with the twin gates of perception and purposeful action.

## II. Deep Learning

Educational psychologists, who were strongly influenced by Herbart, Piaget, and William James, distinguish *deep learning* as a search for meaning from *surface learning* as the memorization of facts. Although their definitions are not sufficiently precise to be implemented in AI systems, surface learning corresponds to accretion, and deep learning corresponds to schema tuning and restructuring.

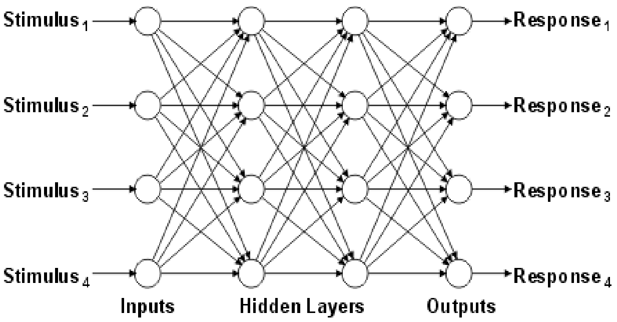


Fig 1. A neural net for connecting stimuli to responses

Educational psychologists considered behaviorism *rat psychology*. But Thorndike [29], a former student of William James, used animal experiments to develop a stimulus-response theory, which he called *connectionism*: rewards strengthen the S-R connections, and punishments weaken them. In 1943, McCulloch and Pitts [13], a neurophysiologist collaborating with a logician, designed a theoretical model of neural nets that were capable of learning and computing any Boolean function of their inputs. To implement a version, Rosenblatt used potentiometers to change weights on the links of a net called a *perceptron* [22]. Later neural nets were implemented by programs on a digital computer.

Behaviorist methods of operant conditioning suggested neural-net methods of learning by *backpropagation*. Figure 1 shows a neural net with stimuli entering the layer of nodes at the left. The nodes represent neurons, and the links represent axons and dendrites that propagate signals from left to right. In computer simulations, each node computes its output values as a function of its inputs. Whenever a network generates an incorrect response to the training stimuli, it is, in effect, "punished" by adjusting weights on the links, starting from the response layer and propagating the adjustments backwards toward the stimulus layer.

Figure 1 is an example of a multilayer feedforward neural network. The original perceptrons had just a single layer of nodes that connected inputs to outputs. With more layers and the option of cycles that provide feedback from later to earlier nodes, more complex patterns can be learned and recognized. After a network has been trained, it can rapidly recognize and classify patterns. But the training time for backpropagation increases rapidly with the number of inputs and the number of hidden layers.

Many variations of algorithms and network topologies have been explored, but the training time for a network grew exponentially with the number of layers. A major breakthrough for *deep belief networks* was a strategy that reduced the training time by orders of magnitude. According to Hinton [6], the key innovation was "an efficient, layer by layer procedure" for determining how the variables at one layer depend on the variables at the previous layer.

> Deep belief nets are learned one layer at a time by treating the values of the latent variables in one layer, when they are being inferred from data, as the data for training the next layer. This efficient, greedy learning can be followed by, or combined with, other learning procedures that fine-tune all of the weights to improve the generative or discriminative performance of the whole network.

With this strategy, the time for training a network with *N* layers depends on the sum of the times for each layer, not their product. As a result, deep belief nets can quickly learn and recognize individual objects, even human faces, in complex street scenes.

The adjective *deep* in front of *belief network* introduced an ambiguity. The term *deep learning*, which the educational psychologists had used for years, implies a human level of understanding that goes far beyond pattern recognition. Enthusiastic partisans of the new methods for training neural networks created an ambiguity by adopting that term. To avoid confusion, others use the more precise term *deep neural network* (DNN). The remainder of this article will use the acronym DNN for methods that use only the neural networks. For hybrid systems that combine a DNN with other technologies, the additional methods and the roles they play are cited.

The Stanford NLP group led by Christopher Manning has been in the forefront of applying statistical methods to language analysis. By combining DNNs for scene recognition with statistical NLP methods, they relate objects in a scene to sentences that describe the scene — for example, the sentence *A small crowd quietly enters the historic church* and the corresponding parts of the scene [25]. But to derive implications of the sentences or the scenes, they use a variation of logic-based methods that have been used in AI for over 40 years [9]. Some of those methods are as old as Aristotle and Euclid. Others were developed by Peirce and Polya [28].

DNNs are highly successful for static pattern recognition. Other techniques, such as hidden Markov models (HMMs), are widely used for recognizing time-varying sequences. For a dissertation under Hinton's supervision, Navdeep Jaitly developed a DNN-HMM hybrid for speech recognition [7]. His office mate, Volodymyr Mnih, combined DNNs with a technique, called *Q-learning*, which uses patterns from two or more time steps of a neural net to make HMM-like predictions [30]. By using DNNs with Q-learning, Mnih and his colleagues at DeepMind Technologies designed the DQN system, which learned to play seven video games for the Atari 2600: Pong, Breakout, Space Invaders, Seaquest, Beamrider, Enduro, and Q*bert [15]. DQN had no prior information about the objects, actions, features, or rules of the games. For each game, the input layer receives a sequence of screen shots: 210×160 pixels and the game score at each step. Each layer learns features, which represent the input data for the next layer. The final layer determines which move to make for the next time step based on patterns in the current and previous time steps.

When compared with other computer systems, DQN outperformed all other machine learning methods on 6 of the 7 games. Its performance was better than a human expert on Breakout, Enduro, and Pong. Its score was close to human performance on Beamrider. But Q*bert, Seaquest, and Space Invaders require long-term strategy. On those games, DQN performed much worse than human experts. Yet the results were good enough for Google to buy the DeepMind company for $400 million [21].

In a critique of DNNs [11], the psycholinguist Gary Marcus wrote "There is good reason to be excited about deep learning, a sophisticated machine learning algorithm that far exceeds many of its predecessors in its abilities to recognize syllables and images. But there's also good reason to be skeptical... deep learning takes us, at best, only a small step toward the creation of truly intelligent machines."

Advocates of DNNs claim that they embody "a unified theory of the human cortex" that holds the key to all aspects of intelligence. But Marcus observed "They have no obvious ways of performing logical inferences, and they are also still a long way from integrating abstract knowledge, such as information about what objects are, what they are for, and how they are typically used. The most powerful AI systems, like Watson, the machine that beat humans in Jeopardy!, use techniques like deep learning as just one element in a very complicated ensemble of techniques." After a discussion with Peter Norvig, director of Google Research, Marcus reported that "Norvig didn't see how you could build a machine that could understand stories using deep learning alone."

## III. Models and Reality

Language, thought, and reality are systems of signs. They are related to the world (or reality), but in different ways. Humans and other animals relate their internal signs (thoughts and feelings) to the world by perception and purposeful action. They also use their sensorimotor organs for communicating with other animals by a wide range of signs, of which the most elaborate are human languages, natural or artificial. Cognitive scientists — philosophers, psychologists, linguists, anthropologists, neuroscientists, and AI researchers — have used these languages to construct an open-ended variety of models and theories about these issues.

As Charles Sanders Peirce observed, all the models, formal and informal, have only two things in common: first, they consist of signs about signs about signs...; second, they are, at best, fallible approximations to reality. As engineers say, all models are wrong, but some are useful. To illustrate the relationships, Figure 2 shows a model as a Janus-like structure, with an engineering side facing the world and an abstract side facing a theory.

On the left is a picture of the physical world, which contains more detail and complexity than any humanly conceivable model or theory could represent. In the middle is a mathematical model that represents a domain of individuals **D** and a set of relations **R** over individuals in **D**. If the world had a unique decomposition into discrete objects and relations, the world itself would be a universal model, of which all correct models would be subsets. But the selection of a domain and its decomposition into objects depend on

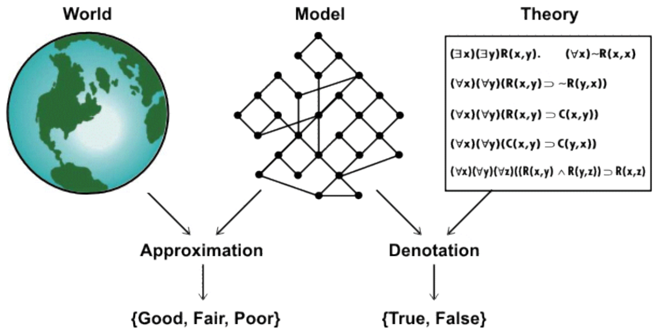

World
Model
Theory
(∃x)(∃y)R(x,y).  (∀x)~R(x,x)
(∀x)(∀y)(R(x,y) ⊃ ~R(y,x))
(∀x)(∀y)(R(x,y) ⊃ C(x,y))
(∀x)(∀y)(C(x,y) ⊃ C(y,x))
(∀x)(∀y)(∀z)((R(x,y) ∧ R(y,z)) ⊃ R(x,z)
Approximation
Denotation
{Good, Fair, Poor}
{True, False}

Fig 2. A model that relates a theory to the world

the intentions of some agent and the limitations of the agent's measuring instruments. Even the best models are approximations to a limited aspect of the world for a specific purpose.

The two-stage mapping from theories to models to the world can reconcile a Tarski-style model theory with the fuzzy methods of Lotfi Zadeh [34]. In Tarski's models, each sentence has only two possible truth values: {T,F}. In fuzzy logic, a sentence can have a continuous range of values from 0.0 for certainly false to 1.0 for certainly true. Hedging terms, such as likely, unlikely, very nearly true, or almost certainly false, represent intermediate values. The two-stage mapping of Figure 2 can accommodate an open-ended variety of models and methods of reasoning. In addition to the Tarski-style models and the fuzzy approximations, it can represent engineering models that use any mathematical, computational, or physical methods for simulating a physical system. For robot guidance, the model may represent the robot's current environment or its future goals. For mental models, it could represent the virtual reality that people have in their heads.

No discrete model can be an exact representation of a physical system. But discrete models can exactly represent a digital computation or a mathematical game such as chess, checkers, or go. In fact, the game of checkers has a far deeper strategy than any of those Atari games. But in 1959, Art Samuel wrote a program that learned to play checkers better than he could [24]. He ran it on an IBM 704, a vacuum-tube computer that had only 144K bytes of RAM and a CPU that was much slower than the Atari. Yet it won one game in a match with the Connecticut state checkers champion.

For the learning method, Samuel used a weighted sum of features to evaluate each position in a game. His algorithm was equivalent to a one-layer neural net. But the program also tested each evaluation by looking ahead several moves. During a game, it maintained an exact model of each state of the game. By searching a few moves ahead, it would determine exact sequences, not the probabilities predicted by an HMM. In the match with the human expert, Samuel's program found the winning move when the human made a mistake.

Samuel's program was a hybrid of a statistical learning method with an exact model for positions in the game. Many chess-playing programs use a similar hybrid, but the programs that model game positions are completely different. Some AI researchers have claimed that to exhibit a human level of intelligence, a *general game playing system* should be be able to learn multiple games without special programming for each one [4]. But the criterion of "special programming" is unclear:

- The rules for games of strategy, such as bridge, chess, and go can be learned in a day. But mastery requires years.
- Many people have become good amateur players of all three games, but no one has ever reached a world-class mastery of more than one. Grandmasters in those games begin early in life, preferably before puberty, and devote many hours per week to reach a world-class level.
- Early learning is also necessary for native mastery of languages, musical performance, gymnastics, and other complex skills. Mentoring by a master is also critical. That concentrated study could be considered a kind of special programming.

## IV. The Cognitive Cycle

The human brain is a complex hybrid of multiple components with different kinds of representations interacting at various stages during the processes of learning and reasoning. After many years of research in the design and implementation of AI systems, Minsky [14] argued that no single mechanism, by itself, can adequately support the full range of functions required for a human level of intelligence:

> What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle. Our species has evolved many effective although imperfect methods, and each of us individually develops more on our own. Eventually, very few of our actions and decisions come to depend on any single mechanism. Instead, they emerge from conflicts and negotiations among societies of processes that constantly challenge one another.

Evidence from fMRI scans supports Minsky's hypothesis. Mason and Just [12] studied 14 participants who were learning the internal mechanisms of four devices: a bathroom scale, a fire extinguisher, an automobile braking system, and a trumpet. For all 14 participants, the brain regions involved in learning how each device worked progressed through the same stages:

> (1) initially, the representation was primarily visual (occipital cortex); (2) it subsequently included a large parietal component; (3) it eventually became cortically diverse (frontal, parietal, temporal, and medial frontal regions); and (4) at the end, it demonstrated a strong frontal cortex weighting (frontal and motor regions). At each stage of knowledge, it was possible for a classifier to identify which one of four mechanical systems a participant was thinking about, based on their brain activation patterns.

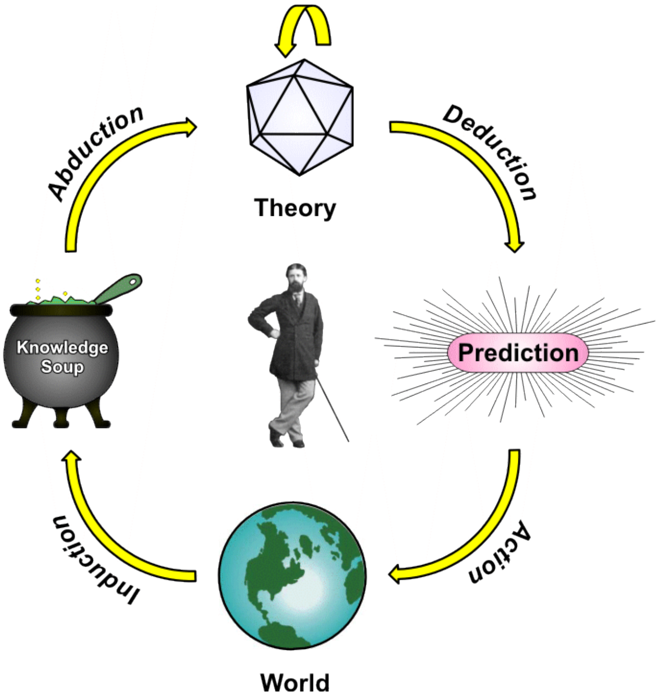


Fig 3. Peirce's cycle of pragmatism

In the first stage, the visual cortex recognized and encoded the visual forms and details of each device. In the second stage, the parietal lobes, which represent cognitive maps or schemata, involved "imagining the components moving." The third stage involved all lobes of the cortex. The medial frontal cortex, which has sensorimotor connections to all parts of the body, suggests that the participants were "generating causal hypotheses associated with mental animation." Finally, the frontal cortex seemed to be "determining how a person (probably oneself) would interact with the system."

An educational psychologist would not find the studies by Mason and Just to be surprising. But a partisan of DNNs might find them disappointing. DNNs may be useful for simulating some aspects of intelligence, but AI and cognitive science have developed many other useful components. In the book *Deep Learning*, the psychologist Stellan Ohlsson reviewed those developments and systematic ways of integrating them [16]. Ohlsson defined learning as "nonmonotonic cognitive change," which "overrides experience" by

- creating creating novel structures that are incompatible with previous versions,
- adapting cognitive skills to changing circumstances, and
- testing those skills by acting upon the environment.

Ohlsson cites Peirce's logic of pragmatism [17] and adopts Peirce's version of abduction as the key to nonmonotonic reasoning. The result is similar to Peirce's cycle of pragmatism (Figure 3).

In Figure 3, the pot of knowledge soup represents the highly fluid, loosely organized accumulation of memories, thoughts, fantasies, and fears in the mind. The arrow of induction represents new observations and generalizations that are tossed in the pot. The crystal at the top symbolizes the elegant, but fragile theories that are constructed by abduction from chunks in the knowledge soup. The arrow above the crystal indicates the process of belief revision, which uses repeated abductions to modify the theories. At the right is a prediction derived from a theory by deduction. That prediction leads to actions whose observable effects may confirm or refute the theory. Those observations are the basis for new inductions, and the cycle continues.

Learning is the process of accumulating chunks of knowledge in the soup and organizing them into theories — collections of consistent beliefs that prove their value by making predictions that lead to successful actions. Learning by any agent — human, animal, or robot — involves a constant cycling from data to models to theories and back to a reinterpretation of the old data in terms of new models and theories. Beneath it all, there is a real world, which the entire community of inquirers learns to approximate through repeated cycles of induction, abduction, deduction, and action.

The cognitive cycle is not a specific technology, such as a DNN, an HMM, or an inference engine. It's a framework for designing hybrid systems that can accommodate multiple components of any kind. Ohlsson, for example, is a psychologist who had collaborated with AI researchers, and he was inspired by Peirce's writings to design systems based on a cognitive cycle that is similar to Figure 3. James Albus was an engineer who studied neuroscience to get ideas for designing robots [1]. Although he had not studied Peirce, he converged on a similar cycle. Majumdar and Sowa [10] did study Peirce, and they adopted Figure 3 as their foundation. Each of the five arrows in Figure 1 may be implemented in a variety of different methods, formal or informal, crisp or fuzzy, statistical or symbolic, declarative or procedural.

In the conclusion of an article about statistical modeling, the statistician Leo Breiman stated an important warning [2]:

> Oddly, we are in a period where there has never been such a wealth of new statistical problems and sources of data. The danger is that if we define the boundaries of our field in terms of familiar tools and familiar problems, we will fail to grasp the new opportunities.

Another statistician, Martin Wilk [33], observed "The hallmark of good science is that it uses models and 'theory' but never believes them." The logician Alfred North Whitehead stated a similar warning about grand theories of any kind [2]: "Systems, scientific and philosophic, come and go. Each method of limited understanding is at length exhausted. In its prime each system is a triumphant success: in its decay it is an obstructive nuisance."

The cognitive cycle is a metalevel framework. It relates the methods of reasoning that people use in everyday life and scientific research. The cycle is self correcting. Every prediction derived in one cycle is tested in the next cycle. Although perfect knowledge is an unattainable goal, repeated cycles can converge to knowledge that is adequate for the purpose. That is the logic of pragmatism.

## References

[1] Albus, James S. (2010) A model of computation and representation in the brain, *Information Sciences* **180**, 1519–1554. http://www.james-albus.org/docs/ModelofComputation.pdf

[2] Breiman, Leo (2001) Statistical Modeling: The Two Cultures, *Statistical Science* **16:3**, 199-231.

[3] Carbonell, Jaime G., Ryszard S. Michalski, & Tom M. Mitchell (1983) An overview of machine learning, in Michalski, Carbonell, & Mitchell, *Machine Learning*, Palo Alto: Tioga, pp. 3-23.

[4] Genesereth, Michael, Nathaniel Love, & Barney Pell (2005) General Game Playing: Overview of the AAAI Competition, *AI Magazine* **26:2**, 62-72.

[5] Herbart, Johann Friedrich (1816) *A Textbook in Psychology*, translated by M. K. Smith, New York: Appleton, 1891.

[6] Hinton, Geoffrey E. (2009) Deep belief networks, *Scholarpedia*, 4(5):5947, http://www.scholarpedia.org/article/Deep_belief_networks

[7] Jaitly, Navdeep (2014) Exploring deep learning methods for discovering features in speech recognition, PhD Dissertation, University of Toronto. http://www.cs.toronto.edu/~ndjaitly/Jaitly_Navdeep_201411_PhD_thesis.pdf

[8] Lamb, Sydney M. (1999) *Pathways of the Brain: The Neurocognitive Basis of Language*, Amsterdam: John Benjamins.

[9] MacCartney, Bill, & Christopher D. Manning (2014) Natural logic and natural language inference, in Harry Bunt et al., *Computing Meaning* **4**, Berlin: Springer, pp. 129–147.

[10] Majumdar, Arun K., & John F. Sowa (2009) Two paradigms are better than one and multiple paradigms are even better, in S. Rudolph, F. Dau, and S.O. Kuznetsov, eds., *Proceedings of ICCS'09*, LNAI 5662, Springer, pp. 32-47. http://www.jfsowa.com/pubs/paradigm.pdf

[11] Marcus, Gary F. (2012) Is "deep learning" a revolution in artificial intelligence? *New Yorker*, 25 November 2012. http://www.Newyorker.com/news/news-desk/is-deep-learning-a-revolution-in-artificial-intelligence

[12] Mason, Robert A., & Marcel Adam Just (2015) Physics instruction induces changes in neural knowledge representation during successive stages of learning, *NeuroImage* **111**, 36-48.

[13] McCulloch, Warren S., & Walter Pitts (1943) A logical calculus of the ideas immanent in nervous activity, *Bulletin of Mathematical Biophysics* **5**, 115-133.

[14] Minsky, Marvin (1986) *The Society of Mind*, New York: Simon & Schuster, Section 30.8.

[15] Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmille (2013) Playing Atari with deep reinforcement learning, NIPS Deep Learning Workshop, https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf

[16] Ohlsson, Stellan (2011) *Deep Learning: How the Mind Overrides Experience*, Cambridge: University Press.

[17] Peirce, Charles Sanders (1903) *Pragmatism as a Principle and Method of Right Thinking*, The 1903 Lectures on Pragmatism, ed. by P. A. Turrisi, SUNY Press, Albany, 1997. Also in [18], pp. 131-241.

[18] Peirce, Charles Sanders (EP) *The Essential Peirce*, ed. by N. Houser, C. Kloesel, and members of the Peirce Edition Project, 2 vols., Indiana University Press, Bloomington, 1991-1998. Peirce, Charles Sanders (CP) *Collected Papers of C. S. Peirce*, ed. by C. Hartshorne, P. Weiss, & A. Burks, 8 vols., Cambridge, MA: Harvard University Press, 1931-1958.

[19] Piaget, Jean (1968) *On the Development of Memory and Identity*, Barre, MA: Clark University Press.

[20] Quillian, M. Ross (1968) Semantic memory, in M. Minsky, *Semantic Information Processing*, Cambridge, MA: MIT Press, pp. 227-270.

[21] Regalado, Antonio (2014) Is Google cornering the market on deep learning? *Technology Review*, 29 January 2014. http://www.technologyreview.com/news/524026/is-google-cornering-the-market-on-deep-learning/

[22] Rosenblatt, Frank (1958) The perceptron: a probabilistic model for information storage and organization in the brain, *Psychological Review* **65:6**, 386-408.

[23] Rumelhart, David E., & Donald A. Norman (1978) Accretion, tuning, and restructuring: three modes of learning, in J. W. Cotton & R. L. Klatzky, eds., *Semantic Factors in Cognition*, Hillsdale, NJ: Lawrence Erlbaum, pp. 37-54.

[24] Samuel, Arthur L. (1959) Some studies in machine learning using the game of checkers, *IBM Journal of Research and Development* **3**, 211-229.

[25] Socher, Richard, Cliff Chiung-Yu Lin, Andrew Y. Ng, & Christopher D. Manning (2011) Parsing natural scenes and natural language with recursive neural networks, ICML 2011. http://www.socher.org/uploads/Main/SocherLinNgManning_ICML2011.pdf

[26] Sowa, John F. (1992) Semantic networks, *Encyclopedia of Artificial Intelligence,* Second Edition, edited by S. C. Shapiro, Wiley, New York. Updated version at http://www.jfsowa.com/pubs/semnet.pdf

[27] Sowa, John F. (2006) The challenge of knowledge soup, in J. Ramadas & S. Chunawala, *Research Trends in Science, Technology, and Mathematics Education*, Mumbai: Homi Bhabha Centre, pp. 55-90. http://www.jfsowa.com/pubs/challenge.pdf

[28] Sowa, John F. (2015) Peirce, Polya, and Euclid: integrating logic, heuristics, and geometry, lecture presented at the American Philosophical Association conference, Vancouver, April 2015. Slides at http://www.jfsowa.com/talks/ppe.pdf

[29] Thorndike, Edward Lee (1932) *The Fundamentals of Learning*, New York: Teachers College Press.

[30] Watkins, Christopher J.C.H., & Peter Dayan (1992) Q-learning, *Machine Learning* **8:3-4**, 279–292.

[31] Watson, Robert I. (1963) *The Great Psychologists from Aristotle to Freud*, New York: Lippencot, pp. 209-210.

[32] Whitehead, Alfred North (1933) *Adventures of Ideas*, New York: Macmillan.

[33] Wilk, Martin, as quoted by John W. Tukey (1962) The future of data analysis, *Annals of Mathematical Statistics* **33**:1, 1-67.

[34] Zadeh, Lotfi A. (1965) Fuzzy sets, *Information and Control* **8**, 338-353.

# 10th International Symposium Advances in Artificial Intelligence and Applications

THE AAIA'15 will bring researchers, developers, practitioners, and users to present their latest research, results, and ideas in all areas of artificial intelligence. We hope that theory and successful applications presented at the AAIA'15 will be of interest to researchers and practitioners who want to know about both theoretical advances and latest applied developments in Artificial Intelligence. As such AAIA'15 will provide a forum for the exchange of ideas between theoreticians and practitioners to address the important issues.

## Topics

Papers related to theories, methodologies, and applications in science and technology in this theme are especially solicited. Topics covering industrial issues/applications and academic research are included, but not limited to:

- Knowledge Management
- Decision Support Systems
- Approximate Reasoning
- Fuzzy Modeling and Control
- Data Mining
- Web Mining
- Machine Learning
- Combining Multiple Knowledge Sources in an Integrated Intelligent System
- Neural Networks
- Evolutionary Computation
- Nature Inspired Methods
- Natural Language Processing
- Image Processing and Interpreting
- Applications in Bioinformatics
- Hybrid Intelligent Systems
- Granular Computing
- Architectures of Intelligent Systems
- Robotics
- Real-world Applications of Intelligent Systems
- Rough Sets

## Professor Zdzislaw Pawlak Best Paper Awards

We are proud to announce that we will continue the tradition started during the AAIA'06 Symposium and award two "Professor Zdzislaw Pawlak Best Paper Awards" for contributions which are outstanding in their scientific quality. The two award categories are:

- Best Student Paper - for graduate or PhD students. Papers qualifying for this award must be marked as "Student full paper" to be eligible for consideration.
- Best Paper Award for the authors of the best paper appearing at the Symposium.

Candidates for the awards can come from AAiA and all workshops organized within its framework (i.e. AIMaViG, AIMA, ASIR, CEIM, LQMR, WCO).

In addition to a certificate, each award carries a prize of 300 EUR provided by the Mazowsze Chapter of the Polish Information Processing Society.

## IFSA Award for Young Scientist

During the Advances in Artificial Intelligence and Applications (AAIA) Symposium, the International Fuzzy Systems Association (IFSA) Best Paper Award for Young Scientist, will be presented.

Candidates for the awards can come from AAiA and all workshops organized within its framework (i.e. AIMaViG, AIMA, ASIR, CEIM, LQMR, WCO).

## Event Chairs

**Janusz, Andrzej,** University of Warsaw, Poland

**Ślęzak, Dominik,** University of Warsaw & Infobright Inc., PolandEvent Chairs

## Advisory Board

**Kacprzyk, Janusz,** Systems Research Institute, Warsaw, Poland

**Kwaśnicka, Halina,** Wroclaw University of Technology, Poland

**Markowska-Kaczmar, Urszula,** Wroclaw University of Technology, Poland

**Skowron, Andrzej,** University of Warsaw, Poland

## Program Committee

**Artiemjew, Piotr,** University of Warmia and Mazury, Poland

**Bartkowiak, Anna,** Wroclaw University, Poland

**Bazan, Jan,** University of Rzeszów, Poland

**Bodyanskiy, Yevgeniy,** Kharkiv National University of Radio Electronics, Ukraine

**Błaszczyński, Jerzy,** Poznan University of Technology, Poland

**Cetnarowicz, Krzysztof,** AGH University of Science and Technology, Poland

**Chakraverty, Shampa,** Netaji Subhas Institute of Technology, India

**Cheung, William,** Hong Kong Baptist University, Hong Kong S.A.R., China

**Cyganek, Boguslaw,** AGH University of Science and Technology, Poland

**Czarnowski, Ireneusz,** Gdynia Maritime University, Poland

**Dardzinska, Agnieszka,** Bialystok University of Technology, Poland

**Dey, Lipika,** Tata Consulting Services, India

**Duentsch, Ivo,** Computer Science Department, Brock University, Canada

**Froelich, Wojciech,** Institute of Computer Science, University of Silesia, Poland

**Girardi, Rosario,** Federal University of Maranhão, Brazil

**Hassanien, Aboul Ella,** Cairo University, Egypt

**Herrera, Francisco,** University of Granada, Spain

**Holzinger, Andreas,** Graz University of Technology, Austria

**Jaromczyk, Jerzy W.,** University of Kentucky, United States

**Jin, Xiaolong,** Chinese Academy of Sciences, China

**Jin, Peng,** Leshan Normal University, China

**Kayakutlu, Gulgun,** Istanbul Technical University, Turkey

**Korbicz, Józef,** University of Zielona Gora, Poland

**Krasuski, Adam,** The Main School of Fire Service (SGSP), Poland

**Kuznetsov, Sergei,** National Research University - Higher School of Economics, Russia

**Lewis, Rory,** University of Colorado at Colorado Springs, United States

**Loukanova, Roussanka,** Department of Mathematics, Stockholm University, Sweden

**Marek, Victor,** University of Kentucky, United States

**Matson, Eric T.,** Purdue University, United States

**Menasalvas, Ernestina,** Universidad Politécnica de Madrid, Spain

**Mercier-Laurent, Eunika,** IAE Lyon3, France

**Mihálydeák, Tamás,** University of Debrecen, Hungary

**Miroslaw, Lukasz,** University of Applied Science Rapperswil & Wroclaw University of Technology, Switzerland

**Miyamoto, Sadaaki,** University of Tsukuba, Japan

**Moshkov, Mikhail,** King Abdullah University of Science and Technology, Saudi Arabia

**Myszkowski, Pawel,** Wroclaw University of Technology, Poland

**Ngan, Ben C. K.,** The Pennsylvania State University, United States

**Nourani, Cyrus F.,** Akdmkrd-DAI TU Berlin, CBS Copenhagen-TansMedia GmbH, Munich, and SFU Burnaby, Canada

**Nowostawski, Mariusz,** Gjovik University College, Norway

**Pancerz, Krzysztof,** University of Management and Administration in Zamość, Poland

**Paradowski, Mariusz,** Wroclaw University of Technology, Poland

**Peters, Georg,** Munich University of Applied Sciences, Germany

**Porta, Marco,** University of Pavia, Italy

**Przybyła-Kasperek, Małgorzata,** University of Silesia, Poland

**Ramanna, Sheela,** University of Winnipeg, Canada

**Ras, Zbigniew,** University of North Carolina at Charlotte, United States

**Reformat, Marek,** University of Alberta, Canada

**Santos Jr., Eugene,** Dartmouth College, United States

**Sas, Jerzy,** Wroclaw University of Technology, Poland

**Schaefer, Gerald,** Loughborough University, United Kingdom

**Sikora, Marek,** Silesian University of Technology, Poland

**Snasel, Vaclav,** VSB -Technical University of Ostrava, Czech Republic

**Sydow, Marcin,** Polish Academy of Sciences and Polish-Japanese Acad. of IT, Poland

**Szczęch, Izabela,** Poznan University of Technology, Poland

**Szpakowicz, Stan,** University of Ottawa, Canada

**Szwed, Piotr,** AGH University of Science and Technology, Poland

**Tsay, Li-Shiang,** North Carolina A&T State University, United States

**Unland, Rainer,** Universität Duisburg-Essen, Germany

**Unold, Olgierd,** Wroclaw University of Technology, Poland

**Wang, Xin,** University of Calgary, Canada

**Wieczorkowska, Alicja,** Polish Japanese Academy of Information Technology, Poland

**Wiśniewski, Piotr,** Nicolaus Copernicus University, Poland

**Wozniak, Michal,** Wroclaw University of Technology, Poland

**Wysocki, Marian,** Rzeszow University of Technology, Poland

**Zadrozny, Slawomir,** Systems Research Institute, Poland

**Zaharie, Daniela,** West University of Timisoara, Romania

**Zakrzewska, Danuta,** Lodz University of Technology, Poland

**Zielosko, Beata,** University of Silesia, Poland

**Zighed, Djamel Abdelkader,** University of Lyon, Lyon 2, France

**Ziolko, Bartosz,** AGH University of Science and Technology, Poland

DOI: 10.15439/2015F378
ACSIS, Vol. 5

# Recurrent drifts: applying fuzzy logic to concept similarity function

Abad, Miguel Ángel
Facultad de Informática,
Universidad Politécnica de Madrid,
Campus de Montegancedo,
28660 - Boadilla del Monte,
Madrid - Spain
Email: miguel.abad.arranz@alumnos.upm.es

Menasalvas, Ernestina
Facultad de Informática,
Universidad Politécnica de Madrid,
Campus de Montegancedo,
28660 - Boadilla del Monte,
Madrid - Spain
Email: emenasalvas@fi.upm.es

***Abstract*—Recurrent drift, as a specific type of concept drift, is characterised by the appearance of previously seen concepts. Therefore, in those cases the learning process could be saved or at least minimized by applying an already trained classification model. In this paper we propose Fuzzy-Rec, a framework that is able to deal with recurrent concept drifts by means of a repository of classification models and a similarity function.**

**Fuzzy logic is used in the framework to implement the similarity function needed to compare different classification models. This is a crucial aspect when dealing with drift recurrence, as long as some measure must be implemented to determine which model better fits a previously seen context. As it can be seen in the experimentation results of this paper, this fuzzy similarity function provides excellent results both in synthetic and real datasets. As a conclusion, we can state that the introduction of fuzzy logic comparisons between models could lead to a better efficient reuse of previously seen concepts, saving computational resources by applying not just equal models, but also similar ones.**

## I. Introduction

TRADITIONAL data stream classification [1] aims to learn a classification model from a stream of training records in order to use it later to predict the class of unlabeled records with high accuracy. Most of these kinds of classification models lack an efficient adaptation to the environment where they are implemented which, in most cases, is constantly changing. For this reason, coping with the improvement and adaptation of classification algorithms on data streams is still a great challenge, as long as data stream mining imposes some requirements that have to be accomplished, namely: maintaining an efficient behaviour in the system, i.e. stable computational and memory load; while providing suitable quality in the classification process, i.e. high accuracy of predictions.

Concept drift is known as the intrinsic changes that occur on the data being processed during data-mining tasks. These changes might be caused by data distribution alterations or by the appearance of a new context that alters the relations among the data attributes. Keeping this scenario in mind, different concept drift techniques have been extensively applied to cope with changes in the underlying distribution of records over time, allowing classification models to be able to adapt their behaviour when needed [2, 3, 4].

Moreover, it is common in real-world data streams for previously seen concepts to reappear [5]. This represents a particular case of concept drift [6], known as recurring concepts [7]; [8]; [5]; [9]; [10]. An adequate management of recurrent concept drifts would lead to a better overall data stream learning and classification processes efficiency and efficacy.

Some real cases where concept recurrence is likely to appear are:

1) Product recommendation systems. Drift in these kind of systems is usually related to fashion trends, economy fluctuations or other hidden context. Anyway, in the first two causes recurrence it is likely to reappear. This is due to the fact that fashion and economy trends reappear during time. A system able to deal with concept recurrence could save some precious training time by means of reusing previously seen recommendation models.
2) Weather prediction. The changes that occur in weather predictions are usually recurrent according to the seasons. Therefore, prediction models that deal well with a specific season could be reused latter on time.
3) Intrusion detection systems. An intrusion detection system (IDS) is a typical monitoring problem which aims to detect cyber incidents. In this case, a trained classification model could send alerts to the operator when a malfunction in the system occurs. A concept drift in an IDS means that the system is behaving in a different way from that expected. But that different behaviour may be caused by a new kind of intrusion that is probably taking place, or because the system monitored is changing in a controlled environment (no intrusion is taking place). If we were able to store all the patterns that represent the different situations of the system monitored (its concepts), we could reuse previously seen models easily.
4) Fraud detection. A similar situation like the one explained in the case of IDS, would be the case of a set of systems dealing with fraud detection. Fraud detection systems are able to detect misbehaviours on

978-83-60810-66-8/$25.00©2015, IEEE

a big amount of data. In this kind of scenarios drift is usually related to economy variations or seasons. As in the previous cases, the detection of drift recurrence could aid in the performance and precision values of the prediction mechanism.

Therefore, there are situations in which a new training is not needed, as the new concept is equal or similar to a previous one. In those cases we could reuse a previously-trained model, saving computation costs and thus providing an efficient method to undertake this new context.

Extending this idea, in this work we propose Fuzzy-Rec as a novel data-stream learning system to help in the process of recurrent concept drift management. In situations where concepts reappear we propose to use a fuzzy similarity function to help in the process of getting the most similar model in a specific context. Some approaches already exist for that goal, but they refer to crisp logic based on true/false values. We assume that a similarity function based on fuzzy logic [11] would improve the similarity process, also allowing us to obtain a better knowledge of what is happening in that process. Furthermore a fuzzy logic similarity function could be adapted for each situation, depending on the feature space of the data stream or on the computational capabilities of the system.

This is a crucial aspect to improve the storing process in the repository effectively. When a model has to be stored in the repository, it is required to know if the concept that the model is representing is recurrent or not. In case of a recurrent concept, it should not be stored, as there are already previous models representing it.

In order to calculate the fuzzy similarity level between models, two main features have been used in this work: i) the level of accuracy of the different models involved regarding a specific set of instances; ii) the number of instances used to train the different models, as a measure of maturity and stability of the model. This process helps the system as a whole to save memory consumption, as long as just the models needed are stored in the repository. In this way, the proposed mechanism makes it feasible to work with complex data-stream environments where an overloaded repository would make it difficult to achieve a suitable quality of the system.

To the best of our knowledge, this is the first work to deal with concept drift by means of the use of fuzzy logic to predict similar previously seen concepts.

Experiments performed with different real and synthetic datasets show that Fuzzy-Rec provides similar precision results when comparing it with other approaches.

The rest of the paper is organized as follows. In Section II, we summarize related work on concept drift and fuzzy logic, which is followed in Section III by the preliminaries of the approach where the motivation, challenges and problem definition are stated. Furthermore, in Section IV, we propose Fuzzy-Rec as a solution to work in recurring concept drift environments, with a detailed description of its components and the algorithm used. Section V presents the results obtained by the experimentation phase. Finally Section VI presents the main conclusions and discussion of future lines of research.

## II. Related Work

The approach that we propose in this paper relies on the storage of previously learnt concepts. A fuzzy similarity function is used to retrieve a previously built model which is similar to the current one.

Consequently, in this section we review and compare our proposal with methods that address the problems of: recurring concepts, change detection and conceptual equivalence.

A recent review of the literature related to the problem of concept drift can be found in [4]. Moreover, a review on the challenges for adaptative learning systems have been published in [12].

### A. Recurring Concepts

There have been several techniques developed to achieve the challenge that arises when dealing with concept drift, be they algorithms adaptations or wrapper mechanisms. New algorithms have recently appeared [1, 2, 3, 4, 5, 13, 14], but some other related challenges have received far less attention. Such is the case of situations where the same concept or a similar one reappears, and a previous model could be reused to enhance the learning process in terms of accuracy and processing time [7, 8, 10, 15, 16].

In this way, most existing proposals do not exploit this and have to learn new concepts from scratch even if they are recurrent. However, there are some solutions that deal with concept recurrence, as is the case of the work presented by Ramamurthy and Bhatnagar [15]. In this research, the authors present an ensemble approach that exploits concept recurrence, using a global set of classifiers learned from sequential data chunks. If no classifier in the ensemble performs better than the error threshold, a new classifier is learned and stored to represent the current concept. The classifiers with better performance on the most recent data form part of the ensemble for labeling new records. In [17] and [18] an ensemble mechanism is used to deal with concept drift. Similarly, in [8] an ensemble is also used, but incremental clustering is performed to maintain information on historical concepts. In this way, the proposed framework captures batches of examples from the stream into conceptual vectors. Conceptual vectors are clustered incrementally according to their distance and for each cluster a new classifier is learnt. Classifiers in the ensemble are then learnt using the clusters. Recently [19] proposed Learn++.NSE, an extension of [20] for nonstationary environments, Learn++.NSE is also an ensemble approach that learns from consecutive batches of data without making any assumptions on the nature or rate of drift. The classifiers are combined using dynamic weight majority and the major novelty is on the weighting function that uses the classifiers time-adjusted accuracy on current and past environments. To deal with resource constraints [21] proposes a novel algorithm to manage a pool of classifiers when learning recurring concepts. The main drawback of these methods, apart from the computational process time needed, is the need of constantly train the models used being them recurrent or not.

More sophisticated approaches that use drift detection [3] have also been proposed to address concept recurrence, such as [7, 10]. These approaches store learned models and reuse them when a similar concept reappears in the stream, thus avoiding the effort necessary to relearn a previously observed concept. The method proposed by Yang et al. [10] consists of using a proactive approach to recurring concepts, which means reusing a concept from the concept history. This concept history is represented as a Markov chain which allows the most probable concept to be selected according to a given transition matrix. The approach proposed by Gama and Kosina [7] uses the drift detection method presented in [3] to identify stable concepts and it also memorizes learned classifiers that represent these concepts. After a change is detected in situations of recurrence, referees are used to choose the most appropriate classifier to be reused (i.e., the referee prediction on the applicability of the classifier is greater than a pre-defined threshold). [22] is a recent work on drift detection which uses a control chart to monitor the misclassification rate of the data stream classifier. RCD [23] is a recent recurring concept drift framework that uses a non-parametric multivariate statistical tests to check for recurrence. [24] proposes a semi-supervised recurring concept learning algorithm that takes advantage of unlabelled data.

In the approach proposed in [16] context-concept relationships are learnt from the concept history. A model from a previously learnt concept associated with a particular context is reused in situations of recurrence. Moreover, the proposed method does not require the partition of the dataset into small batches. The concept representations are learnt by a base learner algorithm from an arbitrary number of records. These concept boundaries are determined when a drift detection method signals a change/drift. To improve [16], which relies on a single classifier to deal with recurring concepts, the use of ensembles has been proposed in [25].

### B. Change Detection

Although online learning systems are able to adapt to evolving data without any additional change detection mechanism, the advantage of explicit change detection is providing information about the intrinsic dynamics of the process generating data. In this way, the change detection module characterizes the techniques and mechanisms for drift detection. One advantage of detection models, is that they can provide a meaningful description (indicating the change-points or small time-windows where the change occurs) and the quantification of changes. They may be divided into two different approaches:

- Monitoring the evolution of performance indicators [5]. Some indicators (e.g., performance measures, properties of the data, etc.) are monitored over time. In the work presented in [26] the monitoring of three performance indicators (accuracy, recall, and precision) has been proposed. Furthermore, a highly referenced work that uses this approach is the FLORA family of algorithms developed by [5].
- Monitoring distributions on two different time-windows. A reference window, which usually summarizes past information, and a window over the most recent records. The work proposed by [27] uses statistical tests based on Chernoff bound to determine if the samples drawn from two probability distributions are different and then decide if a concept change occurred. Also [28, 29, 30] approaches are based on monitoring two different time-windows.

[3] and [9] approaches monitor the error-rate of the learning algorithm to find drift events. In [3], when the learning process error-rate increases above certain pre-defined levels, the method signals that the underlying concept has changed. Alternatively [31] uses the distribution of the distances between classification errors to signal drift. If the distance, which results from more consecutive errors is above pre-defined threshold, the underlying concept must be changing and an event is triggered. The basic adaptation strategy after drift is detected is to discard the old model and learn a new one to represent the new underlying concept [3, 31].

### C. Conceptual equivalence

To determine whether a certain model represents a new concept or a reappearing one, a similarity measure is required. The current work is an improvement of the *Conceptual equivalence* measure proposed by Yang et al. [9] where a fuzzy logic function [32] is used to better represent the relationship between different concepts.

## III. Problem Definition and Preliminaries

This section provides the necessary background to understand Fuzzy-Rec system. We start by motivating and defining the problem, including some basic definitions to understand the basics of the solution proposed as well as the main challenges we are dealing with in this paper.

From now on we assume that the data streams that are used as input in the Fuzzy-Rec model are already preprocessed and adapted to work well with incremental data stream classification processes. In this way, we can then assume that we do not need to preprocess the data streams, this work being out of the scope of this paper.

### A. Motivation

Data stream mining algorithms must come up with the problem of having to keep in memory just a limited number of records to train their models. That is why these algorithms have to cope with the task of processing each training record only once, while maintaining a suitable quality on the resulting model. This is the main difference from traditional data mining algorithms, where multiple passes over data are common.

In particular, that leads to classification techniques on data streams where models have to be learned incrementally with each incoming record. With the availability of these kinds of models, it is feasible to predict the class of unlabeled records anytime from an early stage. Obviously, the more training records the better precision values obtained.

However, in scenarios where the data distribution changes the accuracy of the classification is expected to decrease. In these cases, to continuously maintain the quality of the models, it is also important for them to be able also to detect and adapt anytime to changes in the underlying concept that they represent [2].

One of the causes of changes in the underlying concept may be recurrence, as a particular case of concept drift [5, 7, 8, 10]. In those cases, a previously learned concept is expected to reappear.

We envisage that recognizing and predicting already learned concepts might help the system to better adapt to future changes where these concepts reappear. With that recognition task in place, it would be possible for the algorithms to avoid relearning something from scratch that has been already learned [5, 7, 8, 10, 16]. This same idea has been already explored in [16], where concepts are able to be saved in a repository.

However in our approach we propose a fuzzy based mechanism to decide about similarity of models, improving the storage of the models. As a result, by means of a good similarity selection, the number of instances needed for the training process is expected to decrease.

### B. Preliminaries

*1) Learning with Concept Drift:* Let $X$ be the space of attributes with its possible values; $Y$ the set of possible discrete class values. Let $D$ be the data stream of training records arriving sequentially $X_i = (\vec{x_i}\ , y_i)$ with $x_i \in X$ (feature space) and $y_i \in Y$, where $\vec{x_i}$ is a vector of the attribute values and $y_i$ is the (discrete) class label for the $i^{th}$ record in the stream. In order to train a base learner based on a classification model $m$ incrementally, these records are processed by $m$ with the goal of predicting the class label of a new record $\vec{x} \in X$, so that $m(\vec{x}) = y \in Y$.

As stated in [9], the concept term is more subjective than objective. That is why in the scope of this paper a concept is represented by the learning results of the classification algorithm used as a base learner, such a Hoeffding Tree [33].

In this field, we consider that a stable concept has been learned when the records used during a given period $k$ are independently and identically distributed according to a probability distribution $P_k(x, y)$. In these situations where concept change, $P_k(x, y) \neq P_{k+1}(x, y)$.

*2) Recurring Concepts :* A recurring concept change can be detected when the input records during a period $k$ are generated based on the same distribution as a previously observed period, in a way that $P_k(x, y) = P_{k-j}(x, y)$. To deal with these kinds of situation, the model $m_k$ learned from a certain period $k$ could be saved to be reused later if it is needed. This would avoid the need to learn a new model representing the same concept as $m_k$. With this solution the continuous learning process improves its behaviour, not requiring a previously learned concept to be learnt from scratch. In addition this approach needs fewer training records to be processed than other approaches that do not deal with recurrent concepts. However, to better calculate whether a concept is recurrent or not, a similarity function is usually used. This is the case of the similarity function proposed in [9], which is the starting point in developing the new fuzzy similarity method proposed in this paper.

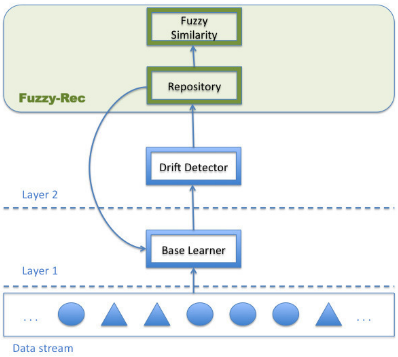


Fig. 1. Fuzzy-Rec Components

### C. Challenges

The main challenges we deal with in this paper are:

- Arranging a fuzzy similarity function in order to better calculate the level of similarity between different concepts.
- Improving the precision values of similar recurrent methods.
- Reducing the number of instances needed to train the classification model.

Fuzzy-Rec faces the aforementioned challenges by means of a fuzzy function to deal with concept similarities evaluation.

### D. Fuzzy-Rec

The main elements of Fuzzy-Rec, as depicted in figure 1, are:

- A repository of previously seen models.
- A concept similarity function based on fuzzy logic. This function is used to determine the level of similarity between concepts. This fuzzy similarity function is crucial to solve the problem of deciding not just which is the most suitable model, but also if the storage of a specific model in the repository is required.

This is therefore a substantial improvement in the work presented in [16], where the problem of concept drift in recurring scenarios was solved in a similar way also by using a repository of concepts and a crisp similarity function.

The proposed Fuzzy-Rec system allows us to better deal with recurrent situations in a classification problem in data streams, helping the evolving base learner to adapt to drifts. Hence the Fuzzy-Rec system is a feasible tool to be used in a wide range of real application scenarios.

We propose a concept similarity function that uses fuzzy logic and it is based on that presented in [9]. This function is defined by the following parameters:

- A conceptual degree of equivalence based on the matching of two different models classifying many instances,

even when their classifications are both wrong. It is not therefore an accuracy equivalence, but a measure of the level of both models to classify in the same way.
- A measure that represents the difference in the number of records used to train each model. This parameter is intended to provide a measure on the maturity and stability of the model.

From the aforementioned two parameters, a fuzzy [34] similarity value is estimated from a previously defined set of rules, making it possible to obtain the poor, average or high values.

There are two situations where a similarity function is required:

1) A model must be stored in the repository of previously trained concepts: in this case the fuzzy similarity function is used to assess the need to store a new model. If there is a similar model in the repository, storing a similar one would not improve the quality of the classification process while unnecessarily increasing the memory consumption.
2) A drift is taking place and it is time to decide whether the new concept is recurrent or not: in this case, the system has been training two different models in parallel to adapt to the drift in a recurrent way (the current model and a new one).

## IV. Implementation of Fuzzy-Rec

As in the case of the MRec system proposed in [16], Fuzzy-Rec can be seen as a two-layer framework:

1) A basic layer where an incremental learning algorithm is able to represent the underlying concept by means of a classification model.
2) An extended layer in which detection and adaptation to concept changes takes place. The detection of recurrent concepts is implemented in this layer. It is also at this level where Fuzzy-Rec implements its fuzzy similarity mechanism.

To provide an in depth knowledge of the implementation of the Fuzzy-Rec system proposed in this paper, first the learning process is described in section IV-A, whereas the description of the fuzzy similarity concepts is presented in section IV-C.

### A. The Learning Process

The on-line learning process for the proposed learning system, as well as the method to detect and adapt to recurrent concepts are detailed in Algorithm 1. The process proceeds as follows:

- It continuously processes the records $X_i = \{\vec{x}, y\}$ with $\vec{x} \in X$ as they appear in the Data Stream.
- In line 3, *currentClassifier* represents the base learner classifier that is currently being used to classify unlabeled records. Its prediction $y$, being right or wrong, on $X_i$ is passed to the drift detection method used to identify the suitable drift level (*stable, warning or drift* ), as explained in IV-B.
- If the process is at the normal level (line 7), the base learner represented by the *currentClassifier* is updated with the new training record. This is the same behaviour as in any other traditional data mining model ready to work with data streams.
- In the case of a warning level (line 8), if the repository does not have the *currentClassifier*, or a similar way as referred to in IV-C, the *currentClassifier* is stored. Still at this level (lines 12 and 13), a $newLearner$ is updated with the training record; the training record is also added to a $warningWindow$. The $warningWindow$ contains the latest records (which should belong to the most recent concept), and will also be used to calculate the conceptual equivalence and estimate the accuracy of models stored with the current concept.
- When drift is signalled (line 14), until there are enough records (i.e., stability period) in the $warningWindow$ the $newLearner$ is updated. When the stability period is over (line 18) it is compared with repository models in terms of conceptual equivalence as stated in IV-C. If the current underlying concept is recurrent a stored model from the repository is used to represent the recurring underlying concept, otherwise the $newLearner$ is used. It is important to remark that the benefit of implementing a previously seen model is that it does not need to be trained again, as it is supposed to be a stable model. When using the newLearner, it needs to be constantly trained during the learning process as it is still an immature model. Therefore, if newLearner is used there is not a decrease in the number of training instances needed. However, the risk of reusing a not suitable recurrent model is still latent. In those cases, the accuracy of the classification base learner would drop.
- A false alarm (line 24) case is used when a warning is signaled but then the learner returns back to normal without achieving drift. In those cases, both the $warningWindow$ and the $newLearner$ are cleared.

In short, the Fuzzy-Rec system can be seen as a continuous learning process with the following steps:

1) The base learner processes the incoming records from data streams by means of an incremental learning algorithm to generate a decision model $m$ representing the underlying concept. The model $m$ will be used to classify unlabelled records.
2) A drift detection method is continuously monitoring the error-rate of the learning algorithm [3]. When the error-rate goes beyond some predefined levels, the drift detection method signals a *warning* (possible drift) or a *drift*.
3) Throughout the life cycle of the system, two different cases may be used to adapt to changes in the underlying concept: i) the concept similarity method detects that the underlying concept is new, and the base learner has to learn it by processing the current incoming labelled records in an incremental way. ii) the (fuzzy) concept

**Algorithm 1** Data Stream Learning Process

```
Require: Data stream $DS$, ModelRepository $MR$
repeat
    Get next record $X_i$ from $DS$;
    prediction = currentClassifier.classify($X_i$);
    DriftDetection.update(prediction);
    switch DriftDetection.level
    case Normal
    currentClassifier.train($X_i$);
    case Warning
    if $\neg MR.containsSimilar(currentClassifier)$ then
        MR.store(currentClassifier);
    end if
    WarningWindow.add($X_i$);
    newLearner.train($X_i$);
    case Drift
    repeat
        WarningWindow.add($X_i$);
        newLearner.train($X_i$);
    until $WarningWindow.size > \tau$ //Stability Period
    if $\neg MR.containsSimilar(newLearner)$ then
        currentClassifier = newLearner;
    else
        currentClassifier = MR.getEquivalent(newLearner);
    end if
    case FalseAlarm
      WarningWindow.clear();
      newLearner.delete();
    end switch
until END OF STREAM
```

similarity method detailed in IV-C detects that the underlying concept is recurrent, and a previous model is applied.

### *B. Drift Detection Mechanism*

The Fuzzy-Rec system needs to know when a concept drift is taking place from the behaviour of a base learner. For this purpose Fuzzy-Rec uses the method proposed by Gama el al. [3]. From this method, it is important to remark the following characteristics:

- The system assumes the observation of periods of stable concepts followed by changes that lead to new stable periods with different underlying concepts.
- The error-rate of the base learning algorithm is considered as a random variable from a sequence of Bernoulli trials.
- The general form of the probability of detecting an error is given by means of a binomial distribution.
- Three different drift levels are defined to manage concept changes: stable or at a control level, warning level and drift or out of control level. These levels represent the confidence of the mechanism of having detected a concept drift.

It also important to note that other similar methods can be used to detect change detections in concepts. Since the Fuzzy-Rec system has been developed as a wrapper mechanism, the specific method used for it is transparent, so it is not necessary to change the learning process.

Having detected a change in the underlying concept, there are some situations in which a concept recurrence appears. In these cases it is worth anticipating to the reappearing concept, in order to improve the learning process efficiency [16]. In order to do so, a concept similarity method must be used.

### *C. Concept Similarity*

To determine whether a certain model represents a new concept or a reappearing one, a similarity measure is required. In this paper, the *Conceptual equivalence* measure is developed by means of a fuzzy logic system [35] where two variables are used to calculate the similarity between two models.

The term “fuzzy logic” was introduced in [34], and is a way of representing many-valued logic, allowing approximate reasoning to be applied through the definition of variables with several truth ranges (from 0 to 1) and rule sets. A rule set determines which fuzzy operator must be used in each case.

By means of using fuzzy logic, it is easy to deal with the concept of partial truth, where a truth value may range from completely true to completely false. In fuzzy logic applications it is common to use linguistic variables to facilitate the implementation of rules and truth values. In this way, a linguistic variable may have several truth values in the same system. These truth values can be seen as subranges of a continuous variable.

In the proposed Fuzzy-Rec system, three linguistic variables are defined:

- The variable $equal\_classified$, used to represent the similarity in the classification precision behaviour of two different models, may take the values: poor, good and excellent.
- The variable $diff\_training$, used to represent the difference that exist in the number of training records used between two different models, may take the values: small and big.
- The variable $similarity$, a variable use to calculate the output of the fuzzy system based on the aforementioned variables, may take the values: poor, average and high.

Therefore, it is assumed to get a value of “high” when calculating the $similarity$ variable, although in some cases the range could be lowered to “average” values, depending on the characteristics of the dataset used.

The variable $equal\_classified$ is based on the method proposed by Yang et al. [9] to calculate its conceptual equivalence. In our case, as it has been outlined, the equivalence between two models when dealing with classification similarity is just one parameter of the global fuzzy function. This parameter is calculated as follows:

1) Given two classification models $m_1$,$m_2$ and a sample dataset $D_n$ of $n$ records, it is possible to calculate for each instance $X_i$=($\vec{x_i}$ , $y_i$) a score, score($D_n$) = +1 if (prediction($m_1(\vec{x_i})$) = prediction($m_2(\vec{x_i})$))

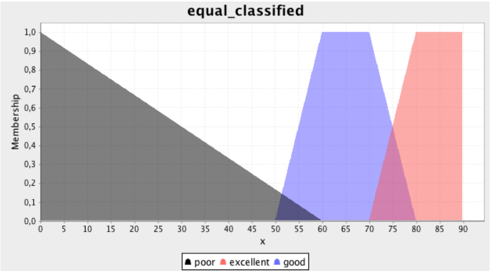


Fig. 2. Membership function of "equal_classified"

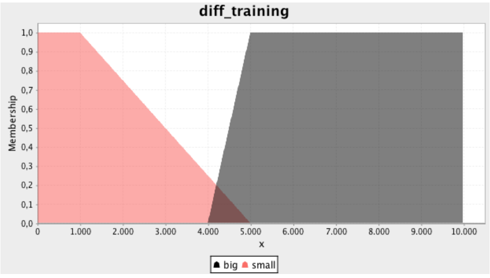


Fig. 3. Membership function of "diff_training"

2) score($D_n$) is used to represent the degree of equivalence in the classification process between $m_1$ and $m_2$.
3) The final classification equivalence $ce$ value, that is a continuous value score with range [0,1], is calculated by

$$ce = \frac{score(D_n)}{N}$$

Depending on the value of $ce$, $equal\_classified$ will take one or another membership value, as represented in the figure 2, where we can see the values this variable may take. The larger the output value of $ce$, the higher the degree of classification equivalence. For the records in $D_n$ it compares how $m_1$ and $m_2$ classify the records. As in [9], the similarity in the classification processes is not necessarily related to the accuracy attribute. This means that two models that present low accuracy for a set of records will have a high $ce$ value, and therefore a high $equal\_classified$ value.

As regards the variable $diff\_training$, its value represents the difference in the number of instances used to train each model we are trying to compare. In figure 3 we can see the values this variable may take.

The rule set implemented to develop the fuzzy logic inference is defined as follows:

1) IF $equal\_classified$ IS $poor$ OR $diff\_training$ IS $big$ THEN $similarity$ IS $poor$;
2) IF $equal\_classified$ IS $good$ AND $diff\_training$ IS $big$ THEN $similarity$ IS $poor$;
3) IF $equal\_classified$ IS $good$ AND $diff\_training$ IS $small$ THEN $similarity$ IS $average$;
4) IF $equal\_classified$ IS $excellent$ AND $diff\_training$ is $big$ THEN $similarity$ is $average$;
5) IF $equal\_classified$ IS $excellent$ AND $diff\_training$ is $small$ THEN $similarity$ is $high$;

Finally, a defuzzification method [35] is needed to get a crisp value of the variable measured. In the case of Fuzzy-Rec, "Center Of Gravity" is the method used to calculate the final value of the $similarity$ variable representing the conceptual equivalence, it being a very popular method in which the "center of mass" of the result provides the crisp value.

From the crisp value returned by the defuzzification method, we evaluate if it is above a predefined threshold. In that case, we assume that the models are similar and thus represent the same underlying concept.

It is important to highlight that without the existence of such a fuzzy method the similarity process should be more naive, being able to integrate just one of the aforementioned variables in the process. As a consequence, a biased classification model with a short life cycle but a high performance could be selected as similar to a more mature and stable model. The fuzzy similarity function implemented in Fuzzy-Rec avoids these kind of behaviours, strengthening the similarity process.

## V. Experiments

In order to validate the Fuzzy-Rec method, and taking into account that Fuzzy-Rec is an extension of the MRec method cited in [16], two different experiments have been developed:

1) Experiment 1: The goal of this experiment is to prove that the precision of Fuzzy-Rec is similar to the MRec method, and no worse than other methods able to deal with concept drift. In order to do so, accuracy and kappa statistic measures are evaluated.
2) Experiment 2: The goal of this experiment is to prove that the training instances needed by Fuzzy-Rec when drifts appear are fewer than the ones needed when using MRec.

In addition a statistical analysis has been developed to validate the results provided by the execution of experiments 1 and 2.

To sum up, the main goal of this experimentation phase is to test the feasibility of using a fuzzy similarity procedure to determine similar previously seen concept drifts.

### A. Parameters setting

To develop the aforementioned experiments, both synthetic and real datasets have been used. A description of the different datasets applied is presented below. Regarding the similarity threshold values needed both for the MRec and Fuzzy-Rec methods, the experiments were developed setting high similarity values; in the specifica case of MRec, a threshold of 0.9 value was used; in the case of Fuzzy-Rec, a similarity value of 0.9 after defuzzification was used. This similarity threshold must be established to afford the comparison process between models. This is important because we must assure that the reused models really fit the context of the data during the learning process. Hence, lower values of the similarity threshold would lead to reuse models that may be not appropriate to the new concept in course. In contrast, higher values would make MRec and Fuzzy-Rec to look for previously seen models that really fit the concept represented by data. In situations where noise could be present in the data, it is important to set higher similarity threshold values to avoid misconceptions.

Regarding the number of classifiers stored in the repository, 10 was the value set for both experiments.

Below a description of the different datasets used during the experimentation phase is made.

### B. Datasets

*1) SEA dataset:* This synthetic dataset is made up of 1.8M instances, representing two drifts repeated for three times. Specifically this dataset was created by means of the following methodology:

1) Create a dataset with 2 concept drifts, changing from SEA function 1 to function 4.
2) Create a dataset with 2 concept drifts, changing from SEA function 2 to function 3.
3) Merge the previous files for three times in a global dataset to ensure that concepts are mixed and may appear in any time during the execution of the dataset.

*2) Hyperplane dataset:* A different synthetic dataset with gradual drifting concepts was created based on a moving hyperplane. A hyperplane in d-dimensional space is denoted by equation: $\sum_{i=1}^{d} a_i x_i = a_0$. Instances are labeled as positive if $\sum_{i=1}^{d} a_i x_i \geq a_0$, and as negative if $\sum_{i=1}^{d} a_i x_i < a_0$. Hyperplanes have been used to simulate time-changing concepts because the orientation and the position of the hyperplane can be changed in a smooth manner by changing the magnitude of the weights [13]. This dataset contains 170,000 instances and it represent different recurrent drifts. Taking into account that some noise has been introduced to the dataset, a threshold value of 0.9 is set to ensure that the reused models really fit the concept represented by the data when drifts happen.

*3) Electricity dataset:* The Electricity Market Dataset (Elec2) [36] is a real dataset that uses data collected from the Australian New South Wales Electricity Market, where the electricity prices are not stationary and are affected by the market supply and demand. The market demand is influenced by context such as season, weather, time of the day and central business district population density. In addition, the supply is influenced primarily by the number of on-line generators, whereas an influencing factor for the price evolution of the electricity market is time. During the time period described in the dataset, the electricity market was expanded with the inclusion of adjacent areas (Victoria state), which led to more elaborate management of the supply as oversupply in one area could be sold interstate.

The Elec2 dataset contains 45,312 records obtained from 7th May 1996 to 5th December 1998, with one record for each half hour (i.e., there are 48 instances for each time period of one day). The class label identifies the change in the price related to a moving average of the last 24 hours. As shown in [36], the dataset exhibits substantial seasonality and is influenced by changes in context. Taking into account that this dataset is expected to have gradual or soft drifts, a similarity threshold of 0.9 is used for this dataset in order to force both MRec and Fuzzy-Rec to reuse just the models associated to concepts really similar to the new appearing one in case of a drift.

*4) Sensor dataset:* Sensor stream [37] is a real dataset that contains information (temperature, humidity, light, and sensor voltage) collected from 54 sensors deployed in Intel Berkeley Research Lab. The whole stream contains consecutive information recorded over a 2 months period (1 reading per 1-3 minutes) which makes a total of 2,219,803 instances. The learning task of the stream is to correctly identify which of the 54 sensors is associated to the sensor information read. The goal of this experiment is to effectively detect and adapt to the multiple concept drifts that this dataset contains.

Taking into account that recurrent drifts are expected to appear in this dataset, a similarity threshold of 0.9 is set in order to force both MRec and Fuzzy-Rec to use previously seen models just in case there were a high level of certainty of equivalence between concepts.

### C. Environment

The implementation of the Fuzzy-Rec learning system was developed in Java, using the MOA [38] environment as a test-bed. The specific fuzzy similarity function implemented in Fuzzy-Rec was developed using jFuzzyLogic [39].

During the execution of the different experiments, the following MOA evaluation features were established:

1) The *Prequential-error* method [38] as the main evaluation technique. When using this evaluator, each individual example can be used to test the model before it is used for training, and from this the accuracy can be incrementally updated. Therefore, the results presented in tables I and II have been gathered in this way.
2) The *Naive Bayes* [40] class as base learner.
3) The *SingleClassifierDrift* class as the method in charge of detecting drifts. This class implements the drift detection method of [3] and adapts to drift by learning a new classifier (i.e., discards previous concept representations).

It is important to note that no distributed environment has been available for the execution of the experimentation phase.

In order to develop the statistical analysis R [41] software was used with the “coin” and “multcomp” packages. Taking into account that when comparing several methods over multiple datasets a post-hoc analysis is desired, in this case the post-hoc tests have been developed using the Wilcoxon-Nemenyi-McDonald-Thompson test [42], using the code of [43].

### D. Results

A description of the results obtained during the execution of the different experiments presented in the beginning of section V is made below. All the experiments have been executed on the datasets presented in section V-B, and comparisons are made with the following methods:

1) MRec with Naive Bayes class as base learner.
2) The RCD method presented in [23] also using Naive Bayes.

*1) Experiment 1:* The goal of this experiment was to prove that the precision values (accuracy and kappa statistic) provided by Fuzzy-Rec were better to the ones provided by the MRec and RCD methods.

As it can be seen in table I, when testing the electricity dataset MRec and Fuzzy-Rec improve the RCD precision values, both behaving in a similar way. This is caused by an application of the same recurrent concepts, so in this case the fuzzy similarity function does not improve the precision values. However, in this case it is demonstrates that in the worst case Fuzzy-Rec provides similar precision values to MRec.

In the case of the sensor dataset, Fuzzy-Rec improves the precision values of MRec by applying a better selection of previously seen models. As long as the difference of training instances between models is one of the variables used when comparing models in Fuzzy-Rec, this allows a more precise choice. Due to the fact that MRec does not make use of such variable, it is using in some cases possible similar models that do not count with enough training records, and therefore the similarity process is biased. When comparing Fuzzy-Rec to RCD in this dataset, similar results are obtained.

When using the SEA dataset there is a big difference on the precision values of Fuzzy-Rec when comparing it with MRec method. The reason of this behaviour is the same than before: the fuzzy similarity function allows to make more precise choices of previously seen models. Also in this case the values obtained when using RCD are similar to Fuzzy-Rec.

Lastly, in the case of the hyperplane dataset Fuzzy-Rec is the method that provides better precision results, improving the behaviour of both RCD and MRec methods.

To sum up, we can conclude in this experiment that Fuzzy-Rec provides similar or even better precision values than MRec or RCD methods in all cases. There are no situation in which Fuzzy-Rec behaves worst than the other methods assessed.

*2) Experiment 2:* The goal of this experiment is to prove that the training instances needed by Fuzzy-Rec when drifts appear are fewer than the ones needed when using MRec and RCD.

The results of this experiments are shown in table II. We can see that except in the case of SEA dataset, Fuzzy-Rec needs fewer training instances than RCD. Comparing these results to the one presented in the previous experiment, we can state that Fuzzy-Rec makes a more efficient use of training instances, improving the precision values of RCD. The reason why Fuzzy-Rec needs more instances when using SEA dataset is due to the high threshold value established to determine similarity. This threshold forces the method to reuse models with a high similarity, which is not reached for this dataset. However, as it has been shown in table I, the precision values are similar to RCD.

Lastly, when comparing the training instances needed by Fuzzy-Rec with the ones needed by MRec, we can see that there are just some slightly differences. As in the previous case, the unique exception is when using the SEA dataset, because of the aforementioned reasons. However, although the instances usage is similar, when comparing these results with the precision values of the previous experiment (see table I), we can conclude that Fuzzy-Rec makes an optimal selection of previously seen models. It is important to note the case of MRec when dealing with SEA dataset, because comparing it with Fuzzy-Rec we can see that although the former makes a lower use of training instances, the precision values obtained drop significantly; in contrast, Fuzzy-Rec while needing more instances provides the better precision results among all the methods assessed. The behaviour of MRec and Fuzzy-Rec with SEA dataset reinforce the idea that the latter provides a more appropriate selection of similar models.

## VI. Conclusions and Future Lines of Research

In this paper the Fuzzy-Rec system, has been described as a mechanism to deal with concept drift in recurring situations.

The main contributions of Fuzzy-Rec are:

1) The implementation of a new similarity concept function using fuzzy logic techniques, which helps in the assessment of similarity between concepts in an improved way.
2) The development of Fuzzy-Rec as a wrapper mechanism, allowing it to be used in an easy way with different base learners and drift detector methods. Furthermore, this wrapper mechanism allows the behaviour of the similarity concept function to be parametrized depending on the needs of each dataset or real-world environment.

Fuzzy-Rec has been tested on different synthetic and real datasets, and comparisons have been made with other similar context-aware algorithms able to deal with drift recurrence. The main conclusions obtained from those experiments are that:

- Fuzzy-Rec is the method that provides the better balance between training instances needed and precision values obtained.
- Fuzzy-Rec needs a low rate of training instances as long as it reuses previously seen models.
- The fuzzy similarity function helps to find the most appropriate model without loosing precision.
- Fuzzy-Rec does not decrease the precision values when comparing it with similar methods.

Future lines of research are: i) analysis of a loss function to penalize bad similarity calculus; ii) implementation of a common fuzzy similarity function in distributed environments.

In both cases, an improvement on precision values is foreseen.

## References

[1] M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "A survey of classification methods in data streams," *Data Streams*, pp. 39–59, 2007.

[2] A. Tsymbal, "The problem of concept drift: definitions and related work," *Computer Science*

TABLE I
DATASETS PRECISION

| | RCD | | MRec | | Fuzzy-Rec | |
|---|---|---|---|---|---|---|
| Dataset | Acc. | Kappa | Acc. | Kappa | Acc. | Kappa |
| Elec2 | 79.2 ± 7.18 | 56.22 ± 15.43 | 81.87 ± 5.66 | 61.71 ± 12.46 | 81.87 ± 5.66 | 61.71 ± 12.46 |
| Sensor | 85.29 ± 13.92 | 84.87 ± 14.32 | 80.99 ± 19.74 | 80.46 ± 20.32 | 85.25 ± 14.64 | 84.82 ± 15.15 |
| SEA | 85.06 ± 2.09 | 67.77 ± 4.41 | 64.63 ± 22.9 | 37.31 ± 33.52 | 85.08 ± 2.04 | 67.85 ± 4.34 |
| Hyperplane | 68.46 ± 10.97 | 36.91 ± 21.92 | 80.41 ± 9.99 | 60.79 ± 19.98 | 81.61 ± 9.1 | 63.19 ± 18.18 |

TABLE II
INSTANCES USED

| Dataset | RCD | MRec | Fuzzy-Rec |
|---|---|---|---|
| Elec2 | 57.41% | 16.72% | 16.69% |
| Sensor | 21.11% | 16.62% | 14.23% |
| SEA | 48.6% | 6.81% | 84.92% |
| Hyperplane | 68.46% | 19.41% | 20.61% |

*Department, Trinity College Dublin*, 2004. [Online]. Available: http://www.cs.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-15.pdf

[3] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," *Lecture Notes in Computer Science*, pp. 286–295, 2004.

[4] I. Žliobaitė, "Learning under concept drift: an overview," *Technical Report. Faculty of Mathematics and Informatics, Vilnius University: Vilnius, Lithuania.*, 2010. [Online]. Available: http://arxiv.org/abs/1010.4784

[5] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine learning*, vol. 23, no. 1, pp. 69–101, 1996.

[6] J. Gama, *Knowledge Discovery from Data Streams*, 1st ed. Chapman & Hall/CRC, 2010.

[7] J. Gama and P. Kosina, "Tracking Recurring Concepts with Meta-learners," in *Progress in Artificial Intelligence: 14th Portuguese Conference on Artificial Intelligence, Epia 2009, Aveiro, Portugal, October 12-15, 2009, Proceedings*. Springer, 2009, p. 423.

[8] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Tracking recurring contexts using ensemble classifiers: an application to email filtering," *Knowl. Inf. Syst.*, vol. 22, no. 3, pp. 371–391, Mar. 2010.

[9] Y. Yang, X. Wu, and X. Zhu, "Mining in anticipation for concept change: Proactive-reactive prediction in data streams," *Data mining and knowledge discovery*, vol. 13, no. 3, pp. 261–289, 2006.

[10] ——, "Combining proactive and reactive predictions for data streams," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, p. 715.

[11] W. Kosinski, P. Prokopowicz, and D. Slezak, "Calculus with fuzzy numbers," in *Proceedings of the Second international conference on Intelligent Media Technology for Communicative Intelligence*, ser. IMTCI'04. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 21–28.

[12] I. Žliobaitė, A. Bifet, M. M. Gaber, B. Gabrys, J. Gama, L. L. Minku, and K. Musial, "Next challenges for adaptive learning systems," *SIGKDD Explorations*, vol. 14, no. 1, pp. 48–55, 2012.

[13] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM New York, NY, USA, 2001, pp. 97–106.

[14] W. Street and Y. Kim, "A streaming ensemble algorithm (SEA) for large-scale classification," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM New York, NY, USA, 2001, pp. 377–382.

[15] S. Ramamurthy and R. Bhatnagar, "Tracking recurrent concept drift in streaming data using ensemble classifiers," in *Proc. of the Sixth International Conference on Machine Learning and Applications*, 2007, pp. 404–409.

[16] J. Bartolo Gomes, E. Menasalvas, and P. Sousa, "Tracking recurrent concepts using context," in *Rough Sets and Current Trends in Computing, Proceedings of the Seventh International Conference RSCTC2010*. Springer, 2010, pp. 168–177.

[17] D. Brzeziński and J. Stefanowski, "Accuracy updated ensemble for data streams with concept drift," in *Proceedings of the 6th international conference on Hybrid artificial intelligent systems - Volume Part II*, ser. HAIS'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 155–163.

[18] D. Brzezinski and J. Stefanowski, "Reacting to different types of concept drift: The accuracy updated ensemble algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 81–94, 2013.

[19] R. Elwell and R. Polikar, "Incremental learning of concept drift in nonstationary environments," *Neural Networks, IEEE Transactions on*, vol. 22, no. 10, pp. 1517–1531, 2011.

[20] M. Muhlbaier, A. Topalis, and R. Polikar, "Learn++. nc: Combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, 2009.

[21] M. J. Hosseini, Z. Ahmadi, and H. Beigy, "New management operations on classifiers pool to track recurring concepts," in *Data Warehousing and Knowledge Discovery*. Springer, 2012, pp. 327–339.

[22] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand, "Exponentially weighted moving average charts for detecting concept drift," *Pattern Recogn. Lett.*, vol. 33, no. 2, pp. 191–198, Jan. 2012.

[23] P. M. Gonçalves Jr and R. S. M. D. Barros, "RCD: A Recurring Concept Drift Framework," *Pattern Recogn. Lett.*, vol. 34, no. 9, pp. 1018–1025, Jul. 2013.

[24] P. Li, X. Wu, and X. Hu, "Mining recurring concept drifts with limited labeled streaming data," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 2, pp. 29:1–29:32, Feb. 2012.

[25] J. a. B. Gomes, E. Menasalvas, and P. A. C. Sousa, "Learning recurring concepts from data streams with a context-aware ensemble," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, ser. SAC '11. New York, NY, USA: ACM, 2011, pp. 994–999.

[26] R. Klinkenberg and I. Renz, "Adaptive information filtering: Learning in the presence of concept drifts," in *Learning for Text Categorization*. Menlo Park, California: AAAI Press, 1998, pp. 33–40.

[27] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting Change in Data Streams," in *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, ser. VLDB '04. VLDB Endowment, 2004, pp. 180–191.

[28] A. Dries and U. Rückert, "Adaptive Concept Drift Detection," *Stat. Anal. Data Min.*, vol. 2, no. 5–6, pp. 311–327, Dec. 2009.

[29] I. Adä and M. Berthold, "EVE: a framework for event detection," *Evolving Systems*, vol. 4, no. 1, pp. 61–70, 2013.

[30] K. Nishida and K. Yamauchi, "Detecting Concept Drift Using Statistical Testing," in *Proceedings of the 10th International Conference on Discovery Science*, ser. DS'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 264–269.

[31] M. Baena-Garcıa, J. del Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavalda, and R. Morales-Bueno, "Early drift detection method," in *Fourth International Workshop on Knowledge Discovery from Data Streams*. Citeseer, 2006, pp. 77–86.

[32] J. Mendel, "Fuzzy logic systems for engineering: a tutorial," *Proceedings of the IEEE*, vol. 83, no. 3, pp. 345 –377, mar 1995.

[33] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM New York, NY, USA, 2000, pp. 71–80.

[34] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, Jun. 1965.

[35] E. Cox, "Fuzzy fundamentals," *Spectrum, IEEE*, vol. 29, no. 10, pp. 58 –61, oct. 1992.

[36] M. Harries, "Splice-2 comparative evaluation: Electricity pricing. Technical report, The University of South Wales," 1999.

[37] X. Zhu, "Stream Data Mining Repository - http://www.cse.fau.edu/~xqzhu/stream.html," 2010.

[38] G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive Online Analysis, 2007 - http://sourceforge.net/projects/moa-datastream/," 2007.

[39] P. Cingolani and J. Alcala-Fdez, "jfuzzylogic: a robust and flexible fuzzy-logic inference system language implementation," in *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, june 2012, pp. 1 – 8.

[40] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.

[41] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[42] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*. Wiley–Interscience, 1999.

[43] T. Galili, "Post-hoc analysis for Friedman test. Code available in http://www.r-statistics.com/2010/02/post-hoc-analysis-for-friedmans-test-r-code," 2010.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 31–38

DOI: 10.15439/2015F231
ACSIS, Vol. 5

# Classification and Optimization of Decision Trees for Inconsistent Decision Tables Represented as MVD tables

Mohammad Azad and Mikhail Moshkov
Computer, Electrical & Mathematical Sciences & Engineering Division
King Abdullah University of Science and Technology
Thuwal 23955-6900, Saudi Arabia
{mohammad.azad, mikhail.moshkov}@kaust.edu.sa

***Abstract*—Decision tree is a widely used technique to discover patterns from consistent data set. But if the data set is inconsistent, where there are groups of examples (objects) with equal values of conditional attributes but different decisions (values of the decision attribute), then to discover the essential patterns or knowledge from the data set is challenging. We consider three approaches (generalized, most common and many-valued decision) to handle such inconsistency. We created different greedy algorithms using various types of impurity and uncertainty measures to construct decision trees. We compared the three approaches based on the decision tree properties of the depth, average depth and number of nodes. Based on the result of the comparison, we choose to work with the many-valued decision approach. Now to determine which greedy algorithms are efficient, we compared them based on the optimization and classification results. It was found that some greedy algorithms ($Mult\_ws\_entSort$, and $Mult\_ws\_entML$) are good for both optimization and classification.**

## I. Introduction

OFTEN in a decision table, we have different examples with the different values of decision and we call such table as a consistent decision table or single-valued decision table. But it is pretty common in real life problems to have inconsistent decision tables where there are groups of examples (objects) with equal values of conditional attributes and different decisions (values of the decision attribute).

In this paper, instead of the group of examples with equal values of conditional attribute, we consider only one example for this group and attach the set of decisions to it. We will call such tables as many-valued decision tables.

In the rough set theory [1], generalized decision ($GD$) has been used to handle inconsistency. In this case, an inconsistent decision table is transformed into the many-valued decision table and after that, each set of decisions has been encoded by a number (decision) such that equal sets are encoded by equal numbers and different sets by different numbers (see Figure 1). We have also used another approach named the most common decision ($MCD$) which is derived from the concept of using most common value in case of missing value [2]. Instead of a group of equal examples with (probably) different decisions, we consider one example given by values of conditional attributes and we attach to this example the most common decision for examples from the group (see Figure 1).

$T^0 =$

| $f_1$ | $f_2$ | $f_3$ | |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 3 |
| 1 | 1 | 0 | 2 |
| 0 | 0 | 1 | 2 |
| 0 | 0 | 1 | 3 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 2 |

$T^0_{MVD} =$

| $f_1$ | $f_2$ | $f_3$ | |
|---|---|---|---|
| 1 | 1 | 1 | $\{1\}$ |
| 0 | 1 | 0 | $\{1,3\}$ |
| 1 | 1 | 0 | $\{2\}$ |
| 0 | 0 | 1 | $\{2,3\}$ |
| 1 | 0 | 0 | $\{1,2\}$ |

$T^0_{GD} =$

| $f_1$ | $f_2$ | $f_3$ | |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 2 |
| 1 | 1 | 0 | 3 |
| 0 | 0 | 1 | 4 |
| 1 | 0 | 0 | 5 |

$T^0_{MCD} =$

| $f_1$ | $f_2$ | $f_3$ | |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 2 |
| 0 | 0 | 1 | 2 |
| 1 | 0 | 0 | 1 |

Fig. 1: Transformation of inconsistent decision table $T^0$ into decision tables $T^0_{MVD}$, $T^0_{GD}$ and $T^0_{MCD}$

In our approach, we can say that for a given example, we have multiple decisions that can be attached to the example but the goal is to find a single decision for each example. We refer this approach as many-valued decision ($MVD$) approach (see Figure 1). This approach is used for classical optimization problems (finding a Hamiltonian circuit with the minimum length or finding nearest post office [3]) where we have multiple optimal solutions but we have to give only one optimal output.

We studied a greedy algorithm for construction of decision trees for many-valued decision tables using the heuristic based on the number of boundary subtables in [4]. Besides, we have studied this algorithm in the cases of most common decision, and generalized decision approaches in [5]. In addition to this, we studied various greedy algorithms as well as dynamic programming algorithm to minimize the average depth in [6], minimize depth in [7], and as well as minimize size of the constructed decision tree in [8].

978-83-60810-66-8/$25.00©2015, IEEE

This paper is a continuation of the current research. We have compared three approaches $MVD$, $MCD$, and $GD$ by comparing the complexity of constructed decision trees. We choose $MVD$ approach based on the result of the comparison. After that, we have shown the average relative difference between greedy algorithm results and optimal results obtained by dynamic programming algorithms for the depth, average depth, and number of nodes of the constructed decision trees. Subsequently, we compare the performance of the classification error rates among classifiers constructed by the various greedy algorithms. We have presented results in the form of critical difference diagram [9] as well as average error rates using data sets from UCI ML Repository [10] and KEEL [11] repository. Finally, we found some of the greedy algorithms are in the top list for both optimization and classification tasks.

## II. Related Works

In literature, these types of tables are often referred as multi-label decision tables [12]. These tables are found in the problem of semantic annotation of images and videos, music categorization into emotions, functional genomics (gene and protein functions), and text classification (news article, email, bookmarks). There are two ways to solve the classification problem from these data sets: first one is algorithm adaptation method where usual classification methods are adapted or modified to handle multi-label data, and the second one is the problem transformation method where the multi-label data set is transformed into single label data set to work with usual classification methods without any modification of the algorithm. These papers solve the inconsistency of the decision table by dividing full set of decisions into relevant and irrelevant decision set for each example. The goal is to find the relevant set of decisions for unknown object.

There is another way to handle inconsistency which is mentioned in different names in literature: partial learning [13], ambiguous learning [14], and multiple label learning [15]. In this learning problem, each example is associated with multiple labels but only one label is correct, and all others are incorrect. The goal is to find out which label is correct. In [13], [15], the authors shows probabilistic methods to solve the learning problem whereas in [14], the author used standard heuristic approach to exploit inductive bias to disambiguate label information.

Our approach of $MVD$ is different from the above mentioned approaches in two ways:
- we assume that all our decisions are correct and there is no incorrect decisions attached with any of the examples,
- we assume that it is enough to find out one decision from the set of decisions rather than the relevant set of decisions.

Therefore, one can use our approach when it is enough to find one decision from the set of decisions.

## III. Preliminaries

### A. Many-valued Decision Table

A *many-valued decision table*, $T$ is a rectangular table whose rows are filled by nonnegative integers and columns are labeled with conditional attributes $f_1, \ldots, f_n$. If we have strings as values of attributes, we have to encode the values as nonnegative integers. There are no duplicate rows, and each row is labeled with a nonempty finite set of natural numbers (set of decisions). We denote the number of examples (rows) in the table $T$ by $N(T)$.

TABLE I: A many-valued decision table $T'$

$T' =$

| $f_1$ | $f_2$ | $f_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | {1} |
| 0 | 1 | 1 | {1,2} |
| 1 | 0 | 1 | {1,3} |
| 1 | 1 | 0 | {2,3} |
| 0 | 0 | 1 | {2} |

If there is a decision which belongs to all of the set of decisions attached to examples of $T$, then we call it a *common decision* for $T$. We will say that $T$ is a *degenerate* table if $T$ does not have examples or it has a common decision. We give an example of degenerate table in the Table II where 1 is the common decision.

TABLE II: A degenerate many-valued decision table

$T'' =$

| $f_1$ | $f_2$ | $f_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | {1} |
| 0 | 1 | 1 | {1,2} |
| 1 | 0 | 1 | {1,3} |

A table obtained from $T$ by removing some examples is called a subtable of $T$. We denote a *subtable* of $T$ which consists of examples that at the intersection with columns $f_{i_1}, \ldots, f_{i_m}$ have values $a_1, \ldots, a_m$ by $T(f_{i_1}, a_1), \ldots, (f_{i_m}, a_m)$. Such nonempty tables (including the table $T$) are called separable subtables of $T$. For example, if we consider subtable $T'(f_1, 0)$ for table $T'$, it will consist of examples $1, 2$, and $5$. Similarly, $T'(f_1, 0)(f_2, 0)$ subtable will consist of examples $1$, and $5$.

TABLE III: Example of subtables of many-valued decision table $T'$

$T'(f_1, 0) =$

| $f_1$ | $f_2$ | $f_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | {1} |
| 0 | 1 | 1 | {1,2} |
| 0 | 0 | 1 | {2} |

$T'(f_1, 0)(f_2, 0) =$

| $f_1$ | $f_2$ | $f_3$ | |
|---|---|---|---|
| 0 | 0 | 0 | {1} |
| 0 | 0 | 1 | {2} |

We denote the set of attributes (columns of table $T$), such that each of them has different values by $E(T)$. For example, if we consider table $T'$, $E(T') = \{f_1, f_2, f_3\}$. Similarly, $E(T'(f_1, 0)) = \{f_2, f_3\}$ for the subtable $T'(f_1, 0)$, because the value for the attribute $f_1$ is constant in subtable $T'(f_1, 0)$. For $f_i \in E(T)$, we denote the set of values from the attribute $f_i$ by $E(T, f_i)$. As an example, if we consider table $T'$ and attribute $f_1$, then $E(T', f_1) = \{0, 1\}$.

The minimum decision which belongs to the maximum number of sets of decisions attached to examples of the table

$T$ is called the *most common decision* for $T$. For example, the most common decision for table $T'$ is 1. Both 1 and 2 appears 3 times in the sets of decisions, but 1 is the most common decision as it is the minimum. We denote the number of examples for which the set of decisions contains the most common decision for $T$ by $N_{mcd}(T)$.

### *B. Decision tree*

A *decision tree* over $T$ is a finite tree with root in which each terminal node is labeled with a decision (a natural number), and each nonterminal node is labeled with an attribute from the set $\{f_1, \ldots, f_n\}$. A number of edges start from each nonterminal node which are labeled with the values of that attribute (e.g. two edges labeled with 0 and 1 for the binary attribute) .

Let $\Gamma$ be a decision tree over $T$ and $v$ be a node of $\Gamma$. We denote $T(v)$ as a subtable of $T$ that is mapped for a node $v$ of decision tree $\Gamma$. If the node $v$ is the root of $\Gamma$ then $T(v) = T$ i.e. the subtable $T(v)$ is the same as $T$. Otherwise, $T(v)$ is the subtable $T(f_{i_1}, \delta_1) \ldots (f_{i_m}, \delta_m)$ of the table $T$ where attributes $f_{i_1}, \ldots, f_{i_m}$ and numbers $\delta_1, \ldots, \delta_m$ are respectively nodes and edge labels in the path from the root to node $v$. We will say that $\Gamma$ is a decision tree for $T$ if $\Gamma$ satisfies the following conditions:
- if $T(v)$ is degenerate then $v$ is labeled with the common decision for $T(v)$,
- otherwise $v$ is labeled with an attribute $f_i \in E(T(v))$. In this case, $k$ outgoing edges from node $v$ are labeled with $a_1, \ldots, a_k$ where $E(T(v), f_i) = \{a_1, \ldots, a_k\}$.

An example of a decision tree for the table $T$ can be found in Fig. 2. If the node $v$ is labeled with the nonterminal attribute $f_3$, then subtable $T(v)$ corresponding to the node $v$ will be the subtable $T(f_1, 0)$ of table $T$. Similarly, the subtable corresponding to the node labeled with 2 will be $T(f_1, 0)(f_3, 1)$ and here 2 is the common decision.

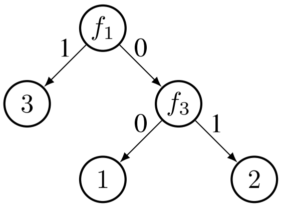


Fig. 2: Decision tree for the many-valued decision table $T'$

### *C. Impurity Functions and Uncertainty Measures*

In greedy algorithm, we need to choose attributes to partition the decision table into smaller subtables until we get degenerate table which then be used to label the terminal node. To choose which partition to consider for tree construction, we need to evaluate the quality of partition by impurity function. We assume that, the smaller the impurity function value, the better is the quality of partition. Impurity function can be calculated based on uncertainty measures for the considered subtables corresponding to the partitions. If we have a common decision, then there is no uncertainty in the data, and uncertainty measure is zero, otherwise uncertainty measure is positive.

**Algorithm 1** Greedy algorithm $A_I$

```
Input: A many-valued decision table $T$ with conditional attributes $f_1, \ldots, f_n$.
Output: Decision tree $A_I(T)$ for $T$.
  Construct the tree $G$ consisting of a single node labeled with the table $T$;
  while (true) do
    if No one node of the tree $G$ is labeled with a table then
      Denote the tree $G$ by $A_I(T)$;
    else
      Choose a node $v$ in $G$ which is labeled with a subtable $T'$ of the table $T$;
      if $U(T') = 0$ then
        Instead of $T'$ mark the node $v$ with the common decision for $T'$;
      else
        For each $f_i \in E(T')$, compute the value of the impurity function $I(T', f_i)$; Choose the attribute $f_{i_0} \in E(T')$, where $i_0$ is the minimum $i$ for which $I(T', f_i)$ has the minimum value; Instead of $T'$ mark the node $v$ with the attribute $f_{i_0}$; For each $\delta \in E(T', f_i)$, add to the tree $G$ the node $v_\delta$ and mark this node with the subtable $T'(f_{i_0}, \delta)$; Draw an edge from $v$ to $v_\delta$ and mark this edge with $\delta$.
      end if
    end if
  end while
```

*1) Uncertainty Measures:* Uncertainty measure $U$ is a function from the set of nonempty many-valued decision tables to the set of real numbers such that $U(T) \geq 0$, and $U(T) = 0$ if and only if $T$ is degenerate.

Let $T$ be a many-valued decision table having $n$ conditional attributes, $N = N(T)$ examples and its examples be labeled with sets containing $m$ different decisions $d_1, \ldots, d_m$. For $i = 1, \ldots, m$, let $N_i$ be the number of examples in $T$ that has been attached with sets of decisions containing the decision $d_i$, and $p_i = N_i/N$. Let $d_1, \ldots, d_m$ be ordered such that $p_1 \geq \cdots \geq p_m$, then for $i = 1, \ldots, m$, we denote by $N_i^{'}$ the number of examples in $T$ such that the set of decisions attached to example contains $d_i$, and if $i > 1$ then this set does not contain $d_1, \ldots, d_{i-1}$, and $p_i^{'} = N_i^{'}/N$. We have the following four uncertainty measures (we assume $0 \log_2 0 = 0$):

- Misclassification error: $me(T) = N(T) - N_{mcd}(T)$. It measures difference between total number of examples and number of examples with most common decision.
- Sorted entropy: $entSort(T) = -\sum_{i=1}^{m} p_i^{'} \log_2 p_i^{'}$ ([14]). First we sort the probabilities for all decisions. After that, for each example, keep the decision having maximum probability and discard others. Then we calculate entropy for this modified decision table.
- Multi-label entropy: $entML(T) = 0$, if and only if $T$ is

degenerate, otherwise, it is equal to $-\sum_{i=1}^{m}(p_i \log_2 p_i + q_i \log_2 q_i)$, where, $q_i = 1 - p_i$. ([16]).

- Absent: $abs(T) = \prod_{i=1}^{m} q_i$, where $q_i = 1 - p_i$. It measures the multiplication of all absent probability $q_i$'s.

*2) Impurity Functions:* Let $U$ be an uncertainty measure, $f_i \in E(T)$, and $E(T, f_i) = \{a_1, \ldots, a_t\}$. The attribute $f_i$ divides the table $T$ into $t$ subtables: $T_1 = T(f_i, a_1), \ldots, T_t = T(f_i, a_t)$. We now define three types of impurity function $I$ which gives us the impurity $I(T, f_i)$ of this partition.

- Weighted max ($wm$):
  $I(T, f_i) = \max_{1 \le j \le t} U(T_j)N(T_j)$.
- Weighted sum ($ws$):
  $I(T, f_i) = \sum_{j=1}^{t} U(T_j)N(T_j)$.
- Multiplied weighted sum ($Mult\_ws$):
  $I(T, f_i) = (\sum_{j=1}^{t} U(T_j)N(T_j)) \times \log_2 t$.

## IV. Greedy Algorithms for Decision Tree Construction

Let $I$ be an impurity function based on the uncertainty measure $U$. The greedy algorithm $A_I$, for a given many-valued decision table $T$, constructs a decision tree $A_I(T)$ for $T$ (see Algorithm 1).

It constructs decision tree sequentially in a top-down fashion. It greedily chooses one attribute at each step based on the considered impurity function. We have total 12 (= 4 × 3) algorithms. The complexities of these algorithms are polynomially bounded above by the size of the tables.

## V. Data Sets

We consider five decision tables from UCI Machine Learning Repository [10]. There were missing values for some attributes which were replaced with the most common values of the corresponding attributes. Some conditional attributes have been removed that take unique value for each example. For the sake of experiments, we removed from these tables more conditional attributes. As a result, we obtained inconsistent decision tables which contain equal examples with equal or different decisions. The information about obtained inconsistent (represented as many-valued decision) tables can be found in Table IV. These modified tables have been renamed as the name of initial table with an index equal to the number of removed conditional attributes.

We also consider five decision tables from KEEL [11] multi-label data set repository. Note that, these tables are already in many-valued decision format. The information about these table can be found in Table V. The decision table 'genbase' has one attribute with unique value for each example and therefore, it was removed, and renamed as 'genbase-1'.

Table IV and V also contain the number of examples (column "Row"), the number of attributes (column "Attr"), the total number of decisions (column "Label"), the cardinality of decision (column "$lc$"), the density of decision (column "$ld$"), and the spectrum of this table (column "Spectrum"). The decision cardinality, $lc$, is the average number of decisions for each example in the table. The decision density, $ld$, is the average number of decisions for each example divided by the total number of decisions. If $T$ is a many-valued decision table with $N$ examples $(x_i, D_i)$ where $i = 1, \ldots, N$, then $lc(T) = \frac{1}{N}\sum_{i=1}^{N}|D_i|$, where $|D_i|$ is the cardinality of decision set in $i$-th example, and $ld(T) = \frac{1}{|L|}lc(T)$, where $L$ is the total number of decisions in $T$. Spectrum of a many-valued decision table is a sequence $\#1$, $\#2, \ldots$, where $\#i$, $i = 1, 2, \ldots$, is the number of examples labeled with sets of decisions with the cardinality equal to $i$. For some tables (marked with * in Table V), the spectrum is too long to fit in the page width. Hence, we show what allows in the page width limit.

## VI. Comparison of Three Approaches

We compared the three approaches $MVD$, $MCD$, and $GD$ to handle inconsistency. We have published the results of the comparison using the decision tree complexity (depth, average depth and number of nodes) in [17]. For the sake of the discussion, we reproduced the result in Table VI for the above 10 decision tables using the algorithm $A_I$ (see Algorithm 1) which uses misclassification error uncertainty measure and weighted sum impurity type.

Data sets from KEEL are already in $MVD$ format. These tables are converted into formats $MCD$ (in this case, the first decision is selected from the set of decisions attached to a row) and $GD$ by the procedure described in Section I. Conversely, inconsistent tables from UCI ML Repository were converted into $MVD$, $MCD$ and $GD$. We then used such data sets to construct decision trees by the algorithm $A_I$, and further we listed the depth, average depth and number of nodes in the constructed decision trees. Note that, we interpreted single valued decision tables, i.e. $T_{GD}$, $T_{MCD}$, as many-valued decision tables where each row is labeled with a set of decisions that has one element. Hence, we can apply the same algorithm for all three cases.

Table VI shows the result of depth, average depth and number of nodes for decision trees $A_I(T_{MVD})$, $A_I(T_{GD})$ and $A_I(T_{MCD})$. Moreover, we took average among the 10 data sets. Since, the result varies in the range of the parameter, we took the normalized average. The normalization has been done by taking the value and dividing by the maximum of three approaches. For example, the maximum depth of the three approaches for the table 'bibtex' is 43. Then the normalized depth of $MVD$ approach will be 39/43 = 0.91. Similarly, the normalized depth of $MCD$ approach will be 42/43 = 0.98, and for $GD$ will be 1.

If we look at the result, the $MVD$ approach in many cases gives minimum depth, average depth and number of nodes. When we took the normalized average, this claim is pretty clear. If our goal is to represent knowledge from the given data using the decision tree, we should use $MVD$ approach as it produces simpler trees compared to other approaches. Therefore, we have used $MVD$ approach for the rest of the paper.

TABLE IV: Characteristics of modified UCI inconsistent data represented in $MVD$ format

| Decision | Row | Attr | Label | $lc$ | $ld$ | Spectrum | | |
|---|---|---|---|---|---|---|---|---|
| table $T$ | | | | | | $\#1$ | $\#2$ | $\#3$ |
| CARS-1 | 432 | 5 | 4 | 1.43 | 0.36 | 258 | 161 | 13 |
| FLAGS-5 | 171 | 21 | 6 | 1.07 | 0.18 | 159 | 12 | |
| LYMPHOGRAPHY-5 | 122 | 13 | 4 | 1.07 | 0.27 | 113 | 9 | |
| NURSERY-1 | 4320 | 7 | 5 | 1.34 | 0.27 | 2858 | 1460 | 2 |
| ZOO-DATA-5 | 42 | 11 | 7 | 1.14 | 0.16 | 36 | 6 | |

TABLE V: Characteristics of KEEL multi-label data

| Decision<br>table $T$ | Row | Attr | Label | $lc$ | $ld$ | Spectrum | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $\#1$ | $\#2$ | $\#3$ | $\#4$ | $\#5$ | $\#6$ | $\#7$ | $\#8$ | $\#9$ |
| $bibtex^*$ | 7355 | 1836 | 159 | 2.41 | 0.015 | 2791 | 1825 | 1302 | 669 | 399 | 179 | 87 | 46 | 18 |
| COREL5K | 4998 | 499 | 374 | 3.52 | 0.009 | 3 | 376 | 1559 | 3013 | 17 | 0 | 1 | 0 | 0 |
| $enron^*$ | 1561 | 1001 | 53 | 3.49 | 0.066 | 179 | 238 | 441 | 337 | 200 | 91 | 51 | 15 | 3 |
| GENBASE-1 | 662 | 1186 | 27 | 1.47 | 0.054 | 560 | 58 | 31 | 8 | 2 | 3 | 0 | 0 | 0 |
| MEDICAL | 967 | 1449 | 45 | 1 | 0.027 | 741 | 212 | 14 | 0 | 0 | 0 | 0 | 0 | 0 |

TABLE VI: Depth, average depth, and number of nodes for decision trees $A_I(T_{MVD})$, $A_I(T_{GD})$ and $A_I(T_{MCD})$ for UCI and KEEL data sets using misclassification error uncertainty measure and weighted sum impurity type

| Decision | Depth | | | Average Depth | | | Number of Nodes | | |
|---|---|---|---|---|---|---|---|---|---|
| table $T$ | $MVD$ | $MCD$ | $GD$ | $MVD$ | $MCD$ | $GD$ | $MVD$ | $MCD$ | $GD$ |
| BIBTEX | 39 | 42 | 43 | 11.52 | 12.24 | 12.97 | 9357 | 10583 | 13521 |
| CARS-1 | 5 | 5 | 5 | 1.958 | 2.583 | 3.813 | 43 | 101 | 280 |
| COREL5K | 156 | 156 | 157 | 36.1 | 36.41 | 36.29 | 6899 | 8235 | 9823 |
| ENRON | 28 | 26 | 41 | 9.18 | 9.62 | 11.18 | 743 | 1071 | 2667 |
| FLAGS-5 | 6 | 6 | 6 | 3.754 | 3.801 | 3.836 | 210 | 216 | 223 |
| GENBASE-1 | 12 | 12 | 11 | 4.718 | 4.937 | 5.762 | 43 | 49 | 81 |
| LYMPHOGRAPHY-5 | 7 | 7 | 7 | 3.787 | 4.115 | 4.311 | 77 | 94 | 112 |
| MEDICAL | 16 | 16 | 16 | 8.424 | 8.424 | 8.424 | 747 | 747 | 747 |
| NURSERY-1 | 7 | 7 | 7 | 2.169 | 3.469 | 4.127 | 198 | 832 | 1433 |
| ZOO-DATA-5 | 4 | 7 | 7 | 3.214 | 3.714 | 4.119 | 19 | 25 | 41 |
| AVERAGE | 28 | 28.4 | 30 | 8.48 | 8.93 | 9.48 | 1833.6 | 2195.3 | 2892.8 |
| NORMALIZED AVERAGE | 0.92 | 0.96 | 0.99 | 0.82 | 0.9 | 0.99 | 0.56 | 0.7 | 1 |

## VII. Decision Tree Optimization

We can optimize the depth, average depth and number of nodes of the constructed decision tree based on dynamic programming algorithms as shown in [6], [8], [7]. It builds all possible separable subtables from the root to the leaf. After that, it considers all possible decision trees by the bottom up approach based on the given criteria of minimizing depth or average depth, or number of nodes. We have compared the average relative difference (in %) $ARD = \frac{greedy-optimal}{optimal} \times 100$ between the sub-optimal results from the greedy algorithms and optimal result from the dynamic programming algorithm. ARD shows how close the greedy result compared to the optimal solution. We have produced the ARD results in Table VIIa, VIIb and VIIc for 3 top algorithms from 12 algorithms (see Section IV).

TABLE VII: ARD (in %) between results of greedy and dynamic algorithms

| Algorithm | $ARD$ |
|---|---|
| $ws\_entSort$ | 4.58 |
| $ws\_entML$ | 5.47 |
| $ws\_me$ | 6.03 |

(a) Average depth

| Algorithm | $ARD$ |
|---|---|
| $wm\_me$ | 12.08 |
| $wm\_entSort$ | 12.08 |
| $ws\_me$ | 12.08 |

(b) Depth

| Algorithm | $ARD$ |
|---|---|
| $Mult\_ws\_entML$ | 21.22 |
| $Mult\_ws\_entSort$ | 24.58 |
| $ws\_abs$ | 25.03 |

(c) Number of nodes

## VIII. Decision Tree Classifier

The examples in the many-valued decision table $T$ have been attached with sets of decisions $D \subset L$ where $L$ is the set of all possible decisions in the table $T$. We denote $D(x)$ as the set of decisions attached to the example $x$. If $X$ is the domain of the examples to be classified, the goal is to find a classifier $h : X \to L$ such that $h(x) = d$, where $d \in D(x)$, that means to find a decision from the ground truth set of decisions attached to the example. To solve the problem, we use decision tree as our model. We construct different kinds of decision trees using various impurity functions.

### A. Evaluation Measure

Here we use the common evaluation measure of classification error percentage. Let us assume, we have unknown instance $x'$ and corresponding decision set is $D(x')$. The classifier $h$ is applied on the new instance $x'$ and it gives the decision $d = h(x')$. If $d \in D(x)$ then $error(x') = 0$, otherwise $error(x') = 1$. Let us assume, we have total $M$ unknown instances to classify, then the error rate will be $\frac{1}{M}\sum_{i=1}^{M} error(x_i)$.

### B. Methodology

Let $T$ be a many-valued decision table with conditional attributes $f_1, \ldots, f_n$, and the decision attribute $D$. We have to divide the initial subtable into three subtables: training subtable $T_1$, validation subtable $T_2$, and test subtable $T_3$. The subtable $T_1$ is used for construction of initial classifier. The subtable $T_2$ is used for pruning of the initial tree. Let $\Gamma$ is a decision tree for $T_1$. For each node $v$ of $\Gamma$, we construct a subtable $T_1(v)$ of the table $T_1$. If $v$ is the root, then $T_1(v) = T_1$, otherwise $T_1(v) = T_1(f_{i_1}, a_1)\ldots(f_{i_m}, a_m)$ where $f_{i_1}, \ldots, f_{i_m}$ are the conditional attributes attached to nodes of the path from the root of $\Gamma$ to $v$, and $a_1, \ldots, a_m$ are numbers attached to edges of this path.

We denote $\alpha(v) = U(T_1(v))/U(T_1)$, where $U(T_1)$ is the misclassification error uncertainty of table $T_1$. Let $\Gamma$ contain $t$ nonterminal nodes, and $v_1, \ldots, v_t$ be all nonterminal nodes of $\Gamma$ in an order such that $\alpha(v_1) \leq \alpha(v_2) \ldots \leq \alpha(v_t)$. For any $i \in \{1, \ldots, t-1\}$, if $\alpha(v_i) = \alpha(v_{i+1})$ then the distance from the root of $\Gamma$ to $v_i$ is at least the distance from the root to $v_{i+1}$. We now construct a sequence of decision trees $\Gamma_0, \Gamma_1, \ldots, \Gamma_t$ where $\Gamma_0 = \Gamma$ (initial tree). The procedure of such decision tree construction is described below in an inductive way:

Let assume that, for some $i \in 0, \ldots, t-1$, the decision tree $\Gamma_i$ is already constructed. We now construct the decision tree $\Gamma_{i+1}$ from the decision tree $\Gamma_i$. Let $D$ be a subtree of $\Gamma_i$ with the root $v_{i+1}$. We remove all nodes and edges of $D$ from $\Gamma_i$ with the exception of $v_{i+1}$. After that, we transform the node $v_{i+1}$ into a terminal node which is labeled with the most common decision for $T_1(v_{i+1})$. As a result, we obtain the decision tree $\Gamma_{i+1}$.

For $i = 0, \ldots, t$, we used the decision tree $\Gamma_i$ to calculate the classification error rate for the table $T_2$. We choose the tree $\Gamma_i$ which has the minimum classification error rate (in case of tie, we choose the tree with smaller index). Now this tree can be used as the final classifier and we can evaluate the test error rate by using this tree to classify the examples in table $T_3$.

## IX. Statistical Comparison of Greedy Algorithms

To compare the algorithms statistically, we used Friedman test with the corresponding Nemenyi post-hoc test as suggested in [9]. Let we have $k$ greedy algorithms $A_1, \ldots, A_k$ for constructing trees and $M$ decision tables $T_1, \ldots, T_M$. For each decision table $T_i$, $i = 1, \ldots, M$, we rank the algorithms $A_1, \ldots, A_k$ on $T_i$ based on their performance scores of classification error rates, where we assign the best performing algorithm the rank of 1, the second best rank 2, and so on. We break ties by computing the average of ranks. Let $r_i^j$ be the rank of the $j$-th of $k$ algorithms on the decision table $T_i$. For $j = 1, \ldots, k$, we correspond to the algorithm $A_j$ the average rank $R_j = \frac{1}{M} \cdot \sum_{i=1}^{M} r_i^j$. For a fixed significance level $\alpha$, the performance of two algorithms is significantly different if the corresponding average ranks differ by at least the critical difference $CD = q_\alpha \sqrt{\frac{k(k+1)}{6M}}$ where $q_\alpha$ is a critical value for the two-tailed Nemenyi test depending on $\alpha$ and $k$.

## X. Classification Results

We used 3-fold cross validation to separate test and training data set for each decision table. The data set is divided into 3 folds, we run the experiment for 3 times. At $i$-th ($i = 1, 2, 3$) iteration, $i$-th fold is used as the test subset, and the rest of data is partitioned randomly into train (70%) and validation subset (30%). The validation subset is used to prune the trained tree. We successively prune the nodes of the trained decision tree model based on the accuracy of the classifier from validation data set unless its accuracy is maximum. After pruning, we used trained decision tree model to predict the decisions for test data sets. For each fold, we repeat the experiment 5 times and take the average of 5 error rates.

We have four uncertainty measures ($me$, $abs$, *entSort*, *entML*) and three types of impurity functions ($ws$, $wm$, *Mult_ws*). So, 12 greedy algorithms have been compared. We show the names of the algorithms as combined name of heuristic and impurity function types separated by '_' in CDD. For example, if the algorithm name is *wm_me*, this means it uses $wm$ as a type of impurity function and $me$ as uncertainty measure. Figure 3 shows the CDD containing average rank for each algorithm on the $x$-axis for significance level of $\alpha = 0.05$. The best ranked algorithm are shown in the leftmost side of the figure. When Nemenyi test cannot identify significant difference between some algorithms, then those are clustered (connected).

It is clear that, *Mult_ws_abs* is the best ranked algorithm to minimize the test error. We have shown classification error rate for each data sets for the three best ranked algorithm in Table VIII as well as the average error rate (AER) among all the data sets. We can see that for most of the data sets *Mult_ws_abs* gives minimum error rate than others. On average it gives the best result. We have also shown the overall execution time for the three best ranked algorithm in the Table IX and found that the *Mult_ws_entML* algorithm is faster than other algorithms.

Also note that, the *Mult_ws_entML* algorithm is the best for minimizing the number of nodes in the tree (see Section VII), and it is also one of the top algorithms for minimizing the classification error rates. Also there are two algorithms (*Mult_ws_entML*, and *Mult_ws_entSort*) for minimizing number of nodes in the tree intersects with the same two algorithms for minimizing the classification error rates. This result is interesting as we can relate the classification and optimization problem. It looks like the quality of classification is connected with the quality of minimizing the number of nodes.

## XI. Conclusion

We studied three different approaches to handle inconsistent decision tables and found $MVD$ approach performs better. We also have created different greedy algorithms based on various uncertainty measures and impurity types to construct decision trees, and compared the results with the optimal results. Finally, we compared these greedy algorithms statistically for classification task to get best ranked classifier and considered also the average error rate across all data sets. We found that $Mult\_ws\_abs$ gives lowest classification error rate than others for most of the data sets. We also found $Mult\_ws\_entML$ algorithm is faster than other top algorithms and also good for both classification and optimization of number of nodes.

In the future, our goal is to construct ensemble of decision trees to work with larger data sets efficiently. Also we are planning to consider more sophisticated pruning methods based on Pareto-optimal points using dynamic programming algorithms.

## Acknowledgement

Research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST).

## References

[1] K. DembczyÅĎski, S. Greco, W. KotÅĆowski, and R. SÅĆowiÅĎski, "Optimized generalized decision in dominance-based rough set approach," in *Rough Sets and Knowledge Technology*, ser. Lecture Notes in Computer Science, 2007, vol. 4481, pp. 118–125.

[2] J. Mingers, "An empirical comparison of selection measures for decision-tree induction," *Machine Learning*, vol. 3, no. 4, pp. 319–342, 1989. doi: 10.1007/BF00116837. [Online]. Available: http://dx.doi.org/10.1007/BF00116837

[3] M. Moshkov and B. Zielosko, *Combinatorial Machine Learning - A Rough Set Approach*, ser. Studies in Computational Intelligence. Springer, 2011, vol. 360. ISBN 978-3-642-20994-9

[4] M. Azad, I. Chikalov, M. Moshkov, and B. Zielosko, "Greedy algorithm for construction of decision trees for tables with many-valued decisions," in *Proceedings of the 21th International Workshop on Concurrency, Specification and Programming, Berlin, Germany, September 26-28, 2012*. CEUR-WS.org, 2012, vol. 928.

[5] M. Azad, I. Chikalov, and M. Moshkov, "Three approaches to deal with inconsistent decision tables - comparison of decision tree complexity," in *RSFDGrC*, 2013. doi: 10.1007/978-3-642-41218-9 pp. 46–54.

[6] M. Azad and M. Moshkov, "Minimization of decision tree average depth for decision tables with many-valued decisions," *Procedia Computer Science*, vol. 35, no. 0, pp. 368 – 377, 2014. doi: http://dx.doi.org/10.1016/j.procs.2014.08.117

[7] ——, "Minimization of decision tree depth for multi-label decision tables," in *Granular Computing (GrC), 2014 IEEE International Conference on*, vol. 0. IEEE, 2014. doi: 10.1109/GRC.2014.6982798

[8] ——, "Minimizing size of decision trees for multi-label decision tables," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*, vol. 0. IEEE, 2014. doi: 10.15439/2014F256

[9] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[10] A. Asuncion and D. J. Newman, "UCI Machine Learning Repository," http://www.ics.uci.edu/ mlearn/, 2007.

[11] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, and S. García, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2011.

[12] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *IJDWM*, vol. 3, no. 3, pp. 1–13, 2007.

[13] T. Cour, B. Sapp, C. Jordan, and B. Taskar, "Learning from ambiguously labeled images," in *CVPR*, 2009. doi: 10.1109/CVPRW.2009.5206667 pp. 919–926.

[14] E. Hüllermeier and J. Beringer, "Learning from ambiguously labeled examples," *Intell. Data Anal.*, vol. 10, no. 5, pp. 419–439, 2006.

[15] R. Jin and Z. Ghahramani, "Learning with multiple labels," in *NIPS*, 2002, pp. 897–904.

[16] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in *PKDD*, 2001. doi: 10.1007/3-540-44794-6 pp. 42–53.

[17] M. Azad and M. Moshkov, "'misclassification error' greedy heuristic to construct decision trees for inconsistent decision tables," in *International Conference on Knowledge Discovery and Information Retrieval*. SCITEPRESS, 2014, pp. 184–191.

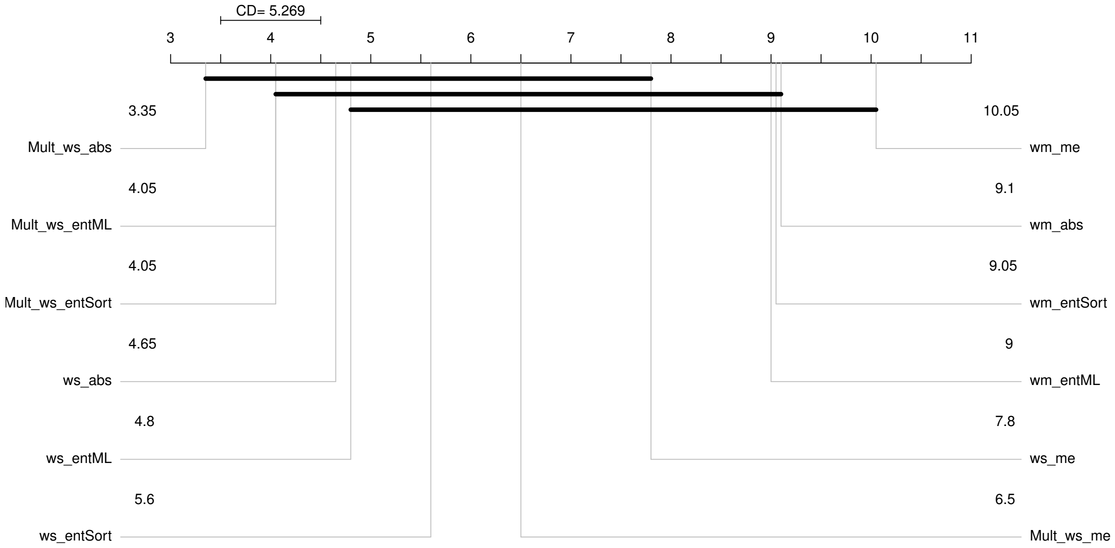


Fig. 3: Critical difference diagram for classification error rates

TABLE VIII: Classification error rate (in %)

| Filename | $Mult\_ws\_abs$ | $Mult\_ws\_entSort$ | $Mult\_ws\_entML$ |
|---|---|---|---|
| BIBTEX | 56.87 | 60.09 | 57.09 |
| CARS-1 | 3.33 | 4.49 | 5.56 |
| COREL5K | 74.3 | 76.57 | 77.72 |
| ENRON | 36.9 | 26.96 | 29.69 |
| FLAGS-5 | 58.83 | 62.11 | 63.27 |
| GENBASE-1 | 5.73 | 3.79 | 3.69 |
| LYMPHOGRAPHY-5 | 27.18 | 25.4 | 25.87 |
| MEDICAL | 24.05 | 26.66 | 26.6 |
| NURSERY-1 | 2.06 | 2.69 | 2.62 |
| ZOO-DATA-5 | 22.86 | 27.62 | 25.24 |
| AER (AVERAGE ERROR RATE) | 31.21 | 31.64 | 31.73 |

TABLE IX: Overall execution time (in sec)

| Filename | $Mult\_ws\_abs$ | $Mult\_ws\_entSort$ | $Mult\_ws\_entML$ |
|---|---|---|---|
| BIBTEX | 285.42 | 2012.34 | 117.55 |
| CARS-1 | 0.0046 | 0.006 | 0.0042 |
| COREL5K | 127.95 | 853.3 | 82.19 |
| ENRON | 6.82 | 14.35 | 5.45 |
| FLAGS-5 | 0.0098 | 0.0176 | 0.0102 |
| GENBASE-1 | 0.1174 | 0.17 | 0.111 |
| LYMPHOGRAPHY-5 | 0.0046 | 0.0056 | 0.0042 |
| MEDICAL | 1.2352 | 6.9238 | 1.1488 |
| NURSERY-1 | 0.0408 | 0.0622 | 0.0416 |
| ZOO-DATA-5 | 0.0024 | 0.0028 | 0.0028 |
| AVERAGE | 42.16 | 288.72 | 20.65 |

DOI: 10.15439/2015F121
ACSIS, Vol. 5

# Comparison of Decision Trees with Rényi and Tsallis Entropy Applied for Imbalanced Churn Dataset

Krzysztof Gajowniczek, Tomasz Ząbkowski, Arkadiusz Orłowski
Department of Informatics, Warsaw University of Life Sciences,
Nowoursynowska 159, 02-776Warsaw, Poland
Email: krzysztof_gajowniczek@sggw.pl, tomasz_zabkowski@sggw.pl, arkadiusz_orlowski@sggw.pl

*Abstract*—**Two algorithms for building classification trees, based on Tsallis and Rényi entropy, are proposed and applied to customer churn problem. The dataset for modeling represents highly unbalanced proportion of two classes, which is often found in real world applications, and may cause negative effects on classification performance of the algorithms. The quality measures for obtained trees are compared for different values of α parameter.**

## I. Introduction

DECISION trees are powerful and very popular tools for different classification tasks [1]-[3]. The attractiveness of this technique is due to the fact that they create rules that can be easily interpreted. Decision trees use some statistical property called information gain to measure the classification power of the input attributes on classification problem as the difference between the entropy before and after a decision. Entropy computation is used to generate simple decision trees, in terms of the structure, with effective classification, since tree size reduction depends on the attribute selection. For this purpose, usually Shannon entropy is used, but other entropy formulas, such as Rényi [4] and Tsallis [5] entropy, can also be applied. Here, a comparative study based on Rényi and Tsallis entropy is described taking into account the issue of imbalance in the class distribution. We used data from telecommunication industry to predict loss of customers to competitors what is known as customer churn. In this dynamical and liberal market customers can choose among cellular service providers and actively migrate from one service provider to another. This problem is especially interesting due to the fact that the portion of churning customers in business practice is low, between 1% and 5%, depending on the country and type of the telecommunication service.

The comparison of the trees is carried out by taking into account different values of α parameter and set of the following measures: classification accuracy, area under the ROC curve, lift, and number of leaves in a tree as complexity measure.

In the next section properties of Rényi and Tsallis entropies are described. The data used in this study are described in the third section. The empirical analysis and comparison of the entropies is shown in fourth section. This type of analysis is especially interesting for decision trees because of the high dimensionality of telecommunication data. Conclusions are given in the last section.

## II. Theoretical framework

In this paper we assume that observations may belong to two given classes and for the classification we use a modified algorithm similar to C4.5 [6] to construct a binary tree in R environment [7].

As a general measure of diversity of objects, a Shannon entropy is often used which is defined as [8]:

$$H_s = -\sum_{i=1}^{n} p_i \log p_i, \tag{1}$$

where $p_i$ is the probability of occurrence of an event $x_i$ being an element of the event $X$ that can take values $x_i,\ldots,x_n$. The value of the entropy depends on two parameters: (1) disorder (uncertainty) and is maximum when the probability $p_i$ for every $x_i$ is equal; (2) the value of n. Shannon entropy assumes a tradeoff between contributions from the main mass of the distribution and the tail. To control both parameters two generalizations were proposed by Rényi [4] and Tsallis [5].

The Rényi entropy is defined as:

$$H_R = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n} p_i^{\alpha}\right), \tag{2}$$

where parameter $\alpha$ is used to adjust the measure depending on the shape of probability distributions.

The Tsallis entropy is defined as:

$$H_R = \frac{1}{\alpha-1}\left(1-\sum_{i=1}^{n} p_i^{\alpha}\right). \tag{3}$$

With Shannon entropy, events with high or low probability have equal weights in the entropy computation. However, using Tsallis entropy, for $\alpha > 1$, events with high probability contribute more than low probabilities for the entropy value [9]. Therefore, the higher is the value of $\alpha$, the higher is the contribution of high probability events in the final result. Furthermore, increasing $\alpha$ parameter $(\alpha \to \infty)$ makes the Rényi entropy determined by events

978-83-60810-66-8/$25.00©2015, IEEE

with higher probabilities, and lower values of $\alpha$ coefficient $(\alpha \to 0)$ weigh the events more equally, no matter of their probabilities.

The Tsallis and Rényi entropies were successfully applied to many diverse practical problems, showing their high usefulness for accurate classification. For instance, in [10] the authors applied both entropies for variable selection in computer networks intrusion detection, analyzing models detection capabilities while providing a set of attributes coming from the network traffic. Their results showed that selecting attributes based on Rényi and Tsallis entropies can achieve better results as compared to Shannon entropy.

Modified C4.5 decision trees based on Tsallis and Rényi entropies have been tested on several high-dimensional microarray datasets in [11]. The results showed that use of non-standard entropies may be highly recommended for this kind of data.

In [12] the authors addressed the question whether the Rényi entropy is equally suit-able to describe systems with q-exponential behavior, where the use of the Tsallis entropy is relevant. The study confirmed that in this case Tsallis entropy is a more suitable choice than Rényi entropy.

Some other studies considered image segmentation based on Tsallis and Rényi-entropies [13]. Their conclusion was that entropic segmentation can give good results but is highly related to an appropriate choice of the entropic index $\alpha$.

## III. The churn dataset

Customer churn is a term used in the telecommunication industry to describe the customer movement from one provider to another, and the churn management strategy is a process aimed to retain profitable customers [14]. Every year telecommunication industry suffers from a substantial loss of valuable customers to competitors. In this liberal market customers can migrate between telecommunication operators freely. The motivation for churn research is based on the fact that it costs more to recruit new customers than to retain existing ones, especially those high profitable customers. The other motivation is the fact that the average churn at cellular providers is about 25% per year in Europe, according to [15], what means that one fourth of the customers' base is lost each year.

In order to check the performance of the proposed entropies, we conducted the simulations based on the data collection known as "Cell2Cell: The Churn Game" [16] derived from the Center of Customer Relationship Management at Duke University, USA. The data constitute a representative slice of the entire customer database, be-longing to an anonymous company operating in the sector of mobile telephony in the United States.

The data contains 71047 observations, wherein each observation corresponds to the individual customer. For each observation 78 variables are assigned, of which 75 potential explanatory variables are used for models construction. All explanatory variables are derived from the same time period, except the binary dependent variable (the values 0 and 1) labeled as "churn", which has been observed in the period from 31 to 60 days later than the other variables. In the collection there is an additional variable "calibrat" to identify the learning sample and test sample, comprising 40000 and 31047 observations, respectively. Learning sample contains 20000 cases classified as churners (leavers) and 20000 cases classified as non-churners. In the test sample, which is used to check the quality of the constructed model, there is only 1.96% of customers who quit. Such a small percentage of the modeled class can be often found in the business practice.

## IV. Analysis and results

### A. Accuracy measures

To compare the trees obtained for different values of $\alpha$ we define a set of three measures. These are: (1) AUC (area under the ROC curve), (2) Lift and, (3) Lv (number of leaves in a tree). The first two measures are related to efficiency and effectiveness of the tree and they have been often used for evaluation of classification models in the context of e.g. credit scoring [17], income and poverty determinants [18] or customer insolvency and churn [19]. The last measure Lv expresses a complexity of the tree as the number of its leaves. In this study we will favor small trees which usually lead to simple and general rules, thus having an advantage over other models. Therefore, a good tree will be characterized by the high accuracy of AUC and lift as well as the relatively small number of leaves. In other words we would like to obtain small but efficient structures for churn classification.

Since we deal with a problem of binary classification, the model yields two results: positive and negative. There are four possible outcomes, as shown in Table 1.

In order to construct AUC measure we need to define two indicators: $Tpr = TP / (\mathrm{TP+FP})$, $Fpr = FP / (\mathrm{FP}+T\,\mathrm{P})$ as well as a ROC curve. As mentioned earlier, each tree's node and leaf has a class assigned based on the share of churn classes. If the share exceeds the decision threshold, usually set to 0.5, a node or a leaf gets a class churn=1 assigned, otherwise class churn=0.

TABLE 1.
CONFUSION MATRIX FOR BINARY CLASSIFICATION

| **Predicted** | **Observed** | |
|---|---|---|
| | Positives | Negatives |
| Positives | True Positives (TP) | False positives (FP) |
| Negatives | False Negatives (FN) | True Negatives (TN) |

Defined indicators can be calculated for various values of the decision threshold. The increase of the threshold from 0 to 1 will yield to a series of points ( $Fpr$ , $Tpr$ ) forming the curve with $Tpr$ on horizontal axis and $Fpr$ on vertical axis. The curve is named receiver operating characteristics, ROC [20], [21]. The AUC measure is an area under the ROC curve which can be calculated using trapezoidal rule. Theoretically $AUC \in [0;1]$ and the larger the AUC the closer is the model to the ideal one and the better is its performance.

The lift measure is dictated by the economic considerations, because the telecom operator does not direct the retention campaign to a wide customer base, but focuses on a small percentage of approximately 1-2% of the customer database on a monthly basis, characterized by the highest probability of resignation. For instance, having the total number of customers of approximately 10 million, a group of 1% of customers is equal to 100 thousand customers per month, which would receive the retention offer.

The required input for lift calculation is a validation dataset that has been “scored" by assigning the estimated churn probability to each case. Next, the churn probabilities are sorted in descending order and for a given customers percentage, the measure is calculated in the following manner (for the first percentile) [22]:

$$Lift_{0.01} = \frac{TP_{0.01}}{TP} \tag{4}$$

The lift measure shows how much more likely we are to receive positive responses (detecting churn customers) in comparison to a random sample of customers.

### B. Experiments

Rényi and Tsallis entropy were compared to each other using the modified C4.5 algorithm for decision tree construction which has been applied to churn dataset. The modification of the algorithm concerned mainly the pruning part. The listing *Generate_decision_tree* presents the tree growing algorithm.

```
Algorithm: Generate_decision_tree
Input: training samples D, list of attributes L,
attribute_selection_method
Output: decision tree

Create a node N
if D has the same class C then
  return N as leaf node with class C label
if L is empty then
  return N as leaf node with class label that is the most class in D
Choose test-attribute a that has the most Gain-Ratio using attribute_selection_method
Give node N with test-attribute label
Find an optimal split that splits D into subsets D_i ($i = 1, ..., k$)
foreach $i = 1$ to $k$ do
  Add branch in node N to test-attribute = $a_i$
   Make partition for sample D_i from samples where test-attribute = $a_i$
  if D_i is empty then
    attach leaf node with the most class in D
       else attach node that generate by Generate_decision_tree(D_i, attribute-list, test-attribute)
endfor
return N
```

The algorithm is recursively called so that it works from the bottom of the tree upward, removing or replacing branches to minimize the predicted error on the validation dataset.

In order to obtain the optimal split while growing the tree (see part of the pseudo-code above) the gain ratio should be calculated. The listing *Prune* outlines the pruning process.

The algorithm is recursively called so that it works from the bottom of the tree up-ward, removing or replacing branches to minimize the predicted error on the validation dataset.

The decision trees were trained on training samples which reflected two designs: (1) learning on the balanced dataset (equal proportion of churn and non-churn classes); (2) learning on the imbalanced dataset with the churn rate of 1.96%. Both designs were then checked on the validation sample in which the churn rate was equal to 1.96%, as observed in real population. We considered α starting from 0.5 to 10 by 0.5.

The results obtained on the validation datasets are collected in Tables 2-3. The best results and corresponding values of α parameter differ in each case and can be summarized as follows:

i. Training the trees on the balanced dataset resulted in better classification performance;

ii. The Rényi entropy based trees trained on imbalanced dataset generated the splits only for α equal to 1 and 1.5; all the other α resulted in no split (Lv=1)

iii. In general, the Tsallis entropy based trees provided better generalization (smaller number of leaves) and the highest lift (3.612);

iv. The Rényi entropy based trees provided complex tree structures with questionable generalization abilities (although the high AUC observed).

v. The Shannon based tree trained on the balanced dataset resulted in high AUC and high lift; however the tree was very complex. The tree trained on the imbalanced dataset did not generate any splits.

The structure of the best tree, in terms of the lift, trained on the balanced dataset using Tsallis entropy is presented in Fig. 1. The tree has 27 leaves on 7 levels (including the root). Each node and each leave has indicated: decision rule, class (TRUE – if churn was observed and FALSE otherwise), and percentage of objects belonging to the majority class.

The first variable used for split was *EQPDYAS* (number of days of the current equipment). If the value of *EQPDYAS* was greater than 302 then the probability of churn increased, forming a group in which the percentage of churners amounted to 56.9%. On the other levels of the tree it was observed that the following variables were useful for detecting churners: *MONTHS* (months in service), *MOU* (mean monthly minutes of use), *RETCALLS* (number of calls previously made to retention team), *RECCHRGE* (mean total recurring charge), *RETACCPT* (number of

```
Algorithm: Prune
Input: node with an attached subtree, validation samples W
Output: pruned tree

    leafError = estimated leaf error on W
    if node is a leaf then
      return leaf error
    else
      subtreeError = $\sum_{N_i \in children(\text{node})} \Pr une(N_i)$
      branchError = error if replaced with most frequent branch
      if leafError is less than branchError and subtreeError then
        make this node a leaf
        error = leafError
      else if
        branchError is less than leafError and subtreeError then
        replace this node with the most frequent branch
        error = branchError
      else
        error = subtreeError
      return error
end
```

previous retention offers accepted), *PEAKVCE* (mean number of in and out peak voice calls), *DIRECTAS* (mean number of director assisted calls), *MOUREC* (mean unrounded MOU received voice calls), *CHANGEM* (% change in minutes of use), *CALLWAIT* (mean number of call waiting calls), *INCOME* (customer income), *REVENUE* (mean monthly revenue).

The final leaves contained the high proportion of churners ranging from 54.9% to 100%. Three rules lead to the leaves with 100% of churners. These were:

Rule 1 – *EQPDAYS* <= 302 & *MONTHS* <= 10 & *RECCHRGE* <= 37.8775 & *RETACCPT* > 0 & *DIRECTAS* <= 2.2275;

Rule 2 – *EQPDAYS* > 302 & *MONTHS* <= 12 & *RETCALLS* > 0 & *PEAKVCE* > 122.33 & *MOUREC* > 86.35 & *RETACCPT* <= 0;

Rule 3 – *EQPDAYS* > 302 & *MONTHS* > 12 & *MOU* <= 6 & *MOU* > 0 & *EQPDAYS* <= 375 & *CHANGEM* > -3.25.

The results presented in this paper are encouraging and provide high accuracy of classification, when compared to similar studies on this dataset. For example, the authors in [23] as an assessment of the quality of the model, chose lift in the first decile, which was equal to 2.61 for the best model. Finally, in our previous study [24] we obtained lift of 3.11 for the first percentile using C&RT tree on the same dataset, while current study delivers improved results.

TABLE 2.
RESULTS ON VALIDATION DATASET WHEN TRAINING THE TREE ON BALANCED DATASET. THE BEST RESULTS FOR EACH ACCURACY MEASURES ARE PRESENTED IN BOLD

| | **Tsallis** | | | **Rényi** | | |
|---|---|---|---|---|---|---|
| **Alpha** | ***AUC*** | ***Lift*** | ***Lv*** | ***AUC*** | ***Lift*** | ***Lv*** |
| 0,5 | 61,32 | 2,463 | 25 | 59,66 | 1,313 | 25 |
| 1 | 61,55 | 2,627 | 33 | 60,83 | 1,642 | 29 |
| 1,5 | 61,95 | 1,806 | 31 | 61,23 | 1,642 | 30 |
| 2 | 60,62 | 1,313 | 34 | 61,13 | **3,284** | 39 |
| 2,5 | 60,86 | 1,313 | 35 | 61,30 | 2,791 | 38 |
| 3 | 61,32 | 3,284 | 35 | 62,20 | 2,463 | 42 |
| 3,5 | 61,97 | 3,119 | 33 | 61,88 | 3,119 | 45 |
| 4 | 61,83 | 3,448 | 35 | 62,73 | 1,642 | 46 |
| 4,5 | 61,21 | 2,463 | 28 | 61,34 | 1,149 | 47 |
| 5 | **62,02** | 2,463 | 31 | 61,29 | 1,806 | 46 |
| 5,5 | 61,17 | 2,134 | 34 | **63,04** | 0,985 | 41 |
| 6 | 60,77 | 3,119 | 30 | 62,05 | 1,149 | 43 |
| 6,5 | 61,17 | **3,612** | 27 | 63,03 | 1,642 | 41 |
| 7 | 58,74 | 2,298 | 19 | 62,53 | 0,985 | 39 |
| 7,5 | 61,11 | 3,284 | 26 | 62,68 | 1,313 | 42 |
| 8 | 61,12 | 3,448 | 22 | 62,19 | 1,313 | 42 |
| 8,5 | 61,27 | 2,791 | 23 | 62,46 | 1,642 | 38 |
| 9 | 60,50 | 2,463 | 20 | 62,34 | 1,477 | 45 |
| 9,5 | 61,22 | 2,791 | 22 | 62,49 | 1,477 | 44 |
| 10 | 60,84 | 3,119 | 12 | 59,66 | 1,313 | 25 |
| | | **Shannon** | | | | |
| | 62,98 | 3,248 | 82 | | | |

TABLE 3.
RESULTS ON VALIDATION DATASET WHEN TRAINING THE TREE ON IMBALANCED DATASET. THE BEST RESULTS FOR EACH ACCURACY MEASURES ARE PRESENTED IN BOLD

| | **Tsallis** | | | **Renyi** | | |
|---|---|---|---|---|---|---|
| **Alpha** | ***Auc*** | ***Lift*** | ***Lv*** | ***Auc*** | ***Lift*** | ***Lv*** |
| 0,5 | 50,00 | 1,000 | 1 | 50,00 | 1,000 | 1 |
| 1 | 58,99 | 2,791 | 10 | **60,83** | 2,463 | 29 |
| 1,5 | 58,35 | 2,298 | 9 | 58,89 | **2,955** | 6 |
| 2 | **59,39** | 2,627 | 11 | 50,00 | 1,000 | 1 |
| 2,5 | 58,98 | 2,791 | 7 | 50,00 | 1,000 | 1 |
| 3 | 57,87 | 1,970 | 6 | 50,00 | 1,000 | 1 |
| 3,5 | 58,10 | 2,791 | 7 | 50,00 | 1,000 | 1 |
| 4 | 58,24 | **2,955** | 11 | 50,00 | 1,000 | 1 |
| 4,5 | 50,00 | 1,000 | 1 | 50,00 | 1,000 | 1 |
| 5 | 58,36 | 1,970 | 11 | 50,00 | 1,000 | 1 |
| 5,5 | 57,75 | 2,463 | 7 | 50,00 | 1,000 | 1 |
| 6 | 58,44 | 2,627 | 7 | 50,00 | 1,000 | 1 |
| 6,5 | 50,00 | 1,000 | 1 | 50,00 | 1,000 | 1 |
| 7 | 56,44 | 2,134 | 8 | 50,00 | 1,000 | 1 |
| 7,5 | 50,00 | 1,000 | 1 | 50,00 | 1,000 | 1 |
| 8 | 58,12 | 2,791 | 7 | 50,00 | 1,000 | 1 |
| 8,5 | 50,00 | 1,000 | 1 | 50,00 | 1,000 | 1 |
| 9 | 50,00 | 1,000 | 1 | 50,00 | 1,000 | 1 |
| 9,5 | 50,00 | 1,000 | 1 | 50,00 | 1,000 | 1 |
| 10 | 50,00 | 1,000 | 1 | 50,00 | 1,000 | 1 |
| | | **Shannon** | | | | |
| | 50,00 | 1,000 | 1 | | | |

## V. SUMMARY AND CONCLUDING REMARKS

In this paper, an evaluation of Rényi and Tsallis entropy applied to customer churn in telecommunication industry is performed. In particular, the modified C4.5 decision tree

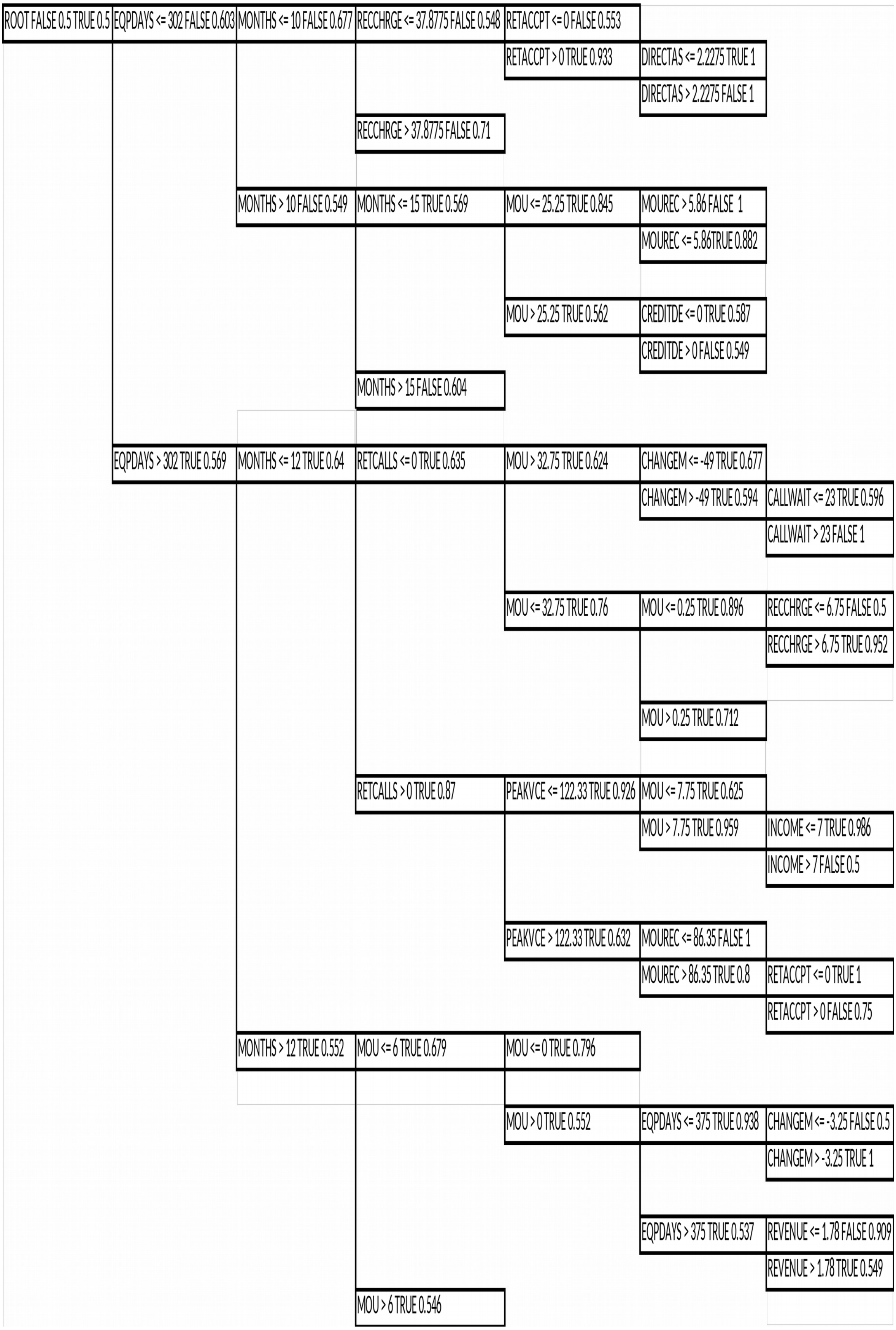


Fig. 1. Decision tree based on Tsallis entropy for α = 6.5 trained on the balanced data.

algorithm was used for classification since it can handle continuous and discrete input variables as observed in the churn dataset.

Additionally, we studied the performance of both entropies in the case of the learning dataset being balanced or imbalanced. The experimental results show that in general, Tsallis and Rényi entropies, with adequate $\alpha$ parameters, can lead to compact and efficient decision trees, with high accuracy measures. We observed that Tsallis entropy provided better generalization since the resulting trees were not as complex as for Rényi case. The study revealed that learning on the balanced learning dataset is beneficial for the final results. Finally, the use of Tsallis and Rényi entropies makes analysis more flexible than standard approach, e.g. Shannon entropy, since it allows for exploration of the tradeoff between the probability of different classes and the overall information gain.

## Acknowledgment

The computations were performed partially employing the computational resources of the Interdisciplinary Centre for Mathematical and Computational Modeling at the Warsaw University (Computational Grant No. G59-31).

## References

[1] Levashenko, V., Zaitseva, E., Pancerz, K., Gomuła, J.: Fuzzy decision tree based classification of psychometric data. In: Ganzha, M., Maciaszek, L., Paprzycki, M. (eds.) Position Papers of the 2014 Federated Conference on Computer Science and Information Systems. Annals of Computer Science and Information Systems, PTI, 3, 37–41, (2014)

[2] Popescu, A., Popescu, B., Brezovan, M., Ganea, E.: Image semantic annotation using fuzzy decision trees. In Computer Science and Information Systems (FedCSIS), IEEE, 597–601 (2013)

[3] Levashenko, V., Zaitseva, E.: Fuzzy decision trees in medical decision making support system. In Computer Science and Information Systems (FedCSIS), IEEE, 213–219 (2012)

[4] Rényi, A.: On measures of entropy and information. Proc. of the 4th Berkeley Symposium on Math. Statistics and Prob., University of California Press, Berkeley, 547–561 (1961)

[5] Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics. Journal of Statistical Physics 52(1-2), 479–487 (1988)

[6] Quinlan, J.: C 4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA (1993)

[7] Team, R. Core. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. (2014)

[8] Shannon, C.E.: A Mathematical Theory of Communication. The Bell System Technical Journal 27, 379–423, 623–656 (1948)

[9] Gajowniczek, K., Karpio, K., Łukasiewicz, P., Orłowski, A., Ząbkowski, T.: Q-entropy approach to selecting high income households, Acta Physica Polonica A, 127(3A), 38–44 (2015)

[10] Lima, C.F.L., Assis de, F. M., Souza de, C.P.: A Comparative Study of Use of Shannon, Rényi and Tsallis Entropy for Attribute Selecting in Network Intrusion Detection. H. Yin et al. (Eds.) IDEAL 2012, Lecture Notes in Computer Science 7435, 492–501 (2012)

[11] Maszczyk, T., Duch, W.: Comparison of Shannon, Renyi and Tsallis Entropy Used in Decision Trees. Rutkowski et al. (Eds.): ICAISC 2008, Lecture Notes in Computer Science 5097, 643–651 (2008)

[12] Johal, R.S., Tirnakli, U.: Tsallis versus Renyi entropic form for systems with q-exponential behaviour: the case of dissipative maps, Physica A 331, 487–496 (2004)

[13] Li, Y., Fan, X., Li, G.: Image segmentation based on Tsallis-entropy and Renyi-entropy and their comparison. IEEE International Conference on Industrial Informatics, 943–948 (2006)

[14] Berson, A., Smith, S., & Thearling, K.: Building data mining applications for CRM. New York, NY: McGraw-Hill (2002)

[15] Chang, Y.T.: Applying data mining to telecom churn management. International Journal of Reviews in Computing 1(10), 67–77 (2009)

[16] Neslin, S.: Cell2Cell: The churn game. Cell2Cell Case Notes. Hanover, NH: Tuck School of Business, Dartmoth College (2002)

[17] Chrzanowska, M., Alfaro. E., Witkowska, D.: The Individual Borrowers Recognition: Single and Ensemble Trees. Expert Systems with Applications 36(3), 6409- 6414 (2009)

[18] Madden, D.: Health and income poverty in Ireland 2003–2006. Journal of Economic Inequality 9, 23–33 (2011)

[19] Ząbkowski, T., Szczesny, W.: Insolvency modeling in the cellular telecommunication industry. Expert Systems with Applications 39, 879-6886 (2012)

[20] Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Machine Learning 31, 1–38 (2004)

[21] Gajowniczek, K., Ząbkowski, T., Szupiluk, R.: Estimating the ROC curve and its significance for classification models' assessment, Quantitative Methods in Economics, 15(2), 382–391 (2014)

[22] Larose, D.T.: Discovering knowledge in data: an introduction to data mining. John Wiley & Sons (2014)

[23] Bose, I., Chen, X.: Hybrid models using unsupervised clustering for prediction of customer churn. Proceedings of the International MultiConference of Engineers and Computer Scientists I, IMECS, Hong Kong (2009)

[24] Gajowniczek, K., Ząbkowski, T.: Problems of churn modeling at cellular telecommunication (in Polish). Quantitative Methods in Economics 13(3), 65–79 (2012)

DOI: 10.15439/2015F229
ACSIS, Vol. 5

# Equality in Computer Proof-Assistants

Adam Grabowski, Artur Korniłowicz
Institute of Informatics, University of Białystok,
ul. Ciołkowskiego 1 M, 15-245 Białystok, Poland
Email: {adam, arturk}@math.uwb.edu.pl

Christoph Schwarzweller
Department of Computer Science, University of Gdańsk,
Wita Stwosza 57, 80-952 Gdańsk, Poland
Email: schwarzw@inf.ug.edu.pl

***Abstract*—Equality is fundamental notion of logic and mathematics as a whole. If computer-supported formalization of knowledge is taken into account, sooner or later one should precisely declare the intended meaning/interpretation of the primitive predicate symbol of equality. In the paper we draw some issues how computerized proof-assistants can deal with this notion, and at the same time, we propose solutions, which are not contradictory with mathematical tradition and readability of source code. Our discussion is illustrated with examples taken from the implementation of the MIZAR system.**

## I. Introduction

THE ROLE of equality in mathematics is indispensable. Linear equations represented and accompanied by graphs seem to be one of the primary mathematical exercises for children, where also recognizing which objects are identical is important. Finding solutions for systems of equations [2], formulas for calculating integrals, values of various mathematical functions, etc., is the basic mathematical activity all engineers are familiar with. Hence it is not surprising that equational provers were one of the primary computerized tools developed for use by mathematicians, after offering simple numerical tools and methods – in fact also based on equality since they essentially handle sequences of equalities.

As a mathematical proof, a proof in a computerized system is just a sequence of proof steps, we could expect that it can be discovered automatically within a reasonable universe of discourse. Typically, a proof search explodes exponentially if the universe and/or the method is not properly chosen. Essentially the process of finding an equational proof is fairly simple. Given sets of equalities can be merged to define a substitution operation by iteration. E.g., in cases when the equalities are between terms and variables, or imply such equalities that can be derived, the substitution operation is the result of (1) (simultaneous) substitution of terms for corresponding variables, according to the given equalities, and (2) iterating this process on results. The iteration step (2) can lead to the target or fail if sooner or later the equalities became too complex to handle. The idea is to have a handy set of underlying techniques for machine learning and to restrict the field of expansions.

The results are always tempting, as after identifying two objects as equal – no matter how different they seem to have been – all knowledge about one of them applies to the other also. Moreover, from the moment of identifying objects, it does not matter with which of them we are dealing, properties discovered for one of them automatically carry over to the other one. Thus, identifying objects as equal can make certain work (e.g., in some non-procedural or extensional work) just easier. Depending on the mathematical context or computational environment, the use of the quality predicate may not be so simple as they look, and saying that two objects are equal we often mean certain level of *abstraction* which is the core of mathematics.

Informally the notion of equality is clear, but following [19] we cannot be sure about that:

> *... when it comes to a crisis of rigorous argument, the open secret is that, for the most part, mathematicians who are not focussed on the architecture of formal systems per se, mathematicians who are consumers rather than providers, somehow achieve a sense of utterly firm conviction in their mathematical doings, without actually going through the exercise of translating their particular argumentation into a brand-name formal system.*

Developers of computerized proof-assistants must be prepared for such exercises, furthermore – they have to provide tools for solving them. Our paper is a result of some thoughts presented in [7] and [3]. However, we are focusing not on computer algebra systems as in [7], and also not on the role of equality in mathematical education (although mathematical proof-assistants are definitely useful in this area). We started from the place where [7] posed some important questions: the area where automated reasoning and computer algebra can interact, and we went ahead of discussions, focusing on real-life implementations of the theoretical ideas: how the automatic proof-checker can cope with the predicate of mathematical equality and corresponding predicate symbol of equality. For our work in this paper, we have chosen MIZAR proof-checker described in [1]. The MIZAR system is much closer to automated theorem prover, although some basic elements of computer algebra are also implemented. In this sense this discussion is a kind of a counterpart of Davenport's research from the viewpoint of MIZAR proof-checker.

Essentially, all with the exception of the very basic MIZAR examples are of our authorship: the first author is mainly responsible for the formalization of lattice theory, rough and fuzzy sets [26], [37], the second author's work is on hard-coding of new MIZAR constructions (e.g. reductions), while the third author delivered examples from abstract algebra. We

978-83-60810-66-8/$25.00©2015, IEEE

hope then that our paper is much more than a theoretic discussion on selected issues on the usage of equality predicates.

The structure of the paper is as follows. At the beginning we draw some initial remarks on the properties of the equality and show how one can solve equational problems with the help of a computer, not necessarily in a fully readable way. Then we focus on the MIZAR system, explaining the implementation of the equality both from purely logical point of view, and extensionally, in set theory (including two extensions: rough and fuzzy sets). Section VI starts the discussion on specific implementation issues in MIZAR: analysis of terms, properties, and built-in computations. In Section XII, based on the concrete example, we show how the discussed techniques work to find the compromise between readability and writability, and then we present the statistics of the use of described constructions in MIZAR Mathematical Library (MML). The final part is devoted to the discussion of the structural understanding of the equality, where injections and isomorphisms are typically used by mathematicians.

## II. Equational Characterization

The importance of equational characterization is obvious – varieties are defined in this way. Among equationally defined classes of algebras, we can find all well-known problems solved with the help of powerful provers, with the Robbins problem [21] (the alternative set of axioms for Boolean algebras) and its solution by EQP/OTTER as the most prominent example. And even if Robbins problem's importance for Boolean algebras is not crucial, this specific set of axioms is quite interesting. E.g., we can point out many problems in lattice theory (presented as short equational bases) solved as a by-product of this computerized system achievement [9].

Absolute equality has the following properties defining it as equivalence relation between the objects in the considered universe of discourse (i.e., domain of objects):

- reflexivity,
- symmetry, and
- transitivity.

Furthermore, the absolute equality should satisfy the so-called Leibniz's Law (or identity of indiscernibles), which is a kind of closure with respect to properties or substitution property.

> *Two objects $x$ and $y$ are equal if for every predicate $P$ we have $P(x)$ if and only if $P(y)$, that is, $x$ and $y$ are equal if they cannot be distinguished using predicates.*

This is an intuitively clear definition, however hard to check, in particular in proof assistants – one can ask *which* universe of discourse should be taken. On the other hand, in some sense, the equality can be treated just like ordinary predicate.

## III. Automation – The Initial Approach

If we are not aware of all fears of the precise meaning and all shades of mathematical equality, we can see its use with computer knowledge systems very flawlessly. We can use equational provers e.g. in the area of lattice theory (as EQP/Prover9 was very successful in this domain), which gives representation of various mathematical areas – topology, algebra, logic, geometry, etc.

Remembering that lattices are structures

$$\langle L, \sqcup, \sqcap \rangle$$

where both binary operations $\sqcup$ and $\sqcap$ are commutative, associative, and satisfy the absorption laws, given as axioms, we can obtain

$$a \sqcup a = a$$

as a result of the six axioms. Essentially, the easy proof (sometimes credited to Dedekind) doesn't need them all, although caused some confusion (e.g., early versions of lattice axiomatics included both idempotences as additional axioms). We can easily obtain a proof using any equational prover. Let us stick to our favourite Prover9 – the direct successor of OTTER:

```
formulas(assumptions).
x ^ y = y ^ x.
x ^ (y ^ z) = (x ^ y) ^ z.
x ^ (x v y) = x.
x v y = y v x.
x v (y v z) = (x v y) v z.
x v (x ^ y) = x.
end_of_list.
```

pushing all the axioms into the assumptions and the desired equality into the goals:

```
formulas(goals).
x v x = x.
end_of_list.
```

and after a while one can get the answer, which is maybe not very readable for a human, but definitely assures us that the proven formula is true.

```
1   x v x = x.         [goal].
5   x ^ (x v y) = x.   [assumption].
9   x v (x ^ y) = x.   [assumption].
10  c1 v c1 != c1.     [deny(1)].
22  x v x = x.         [para(5(a,1),9(a,1,2))].
23  $F.                [resolve(22,a,10,a)].
```

Even for the reader unfamiliar with Prover9 syntax [20] it is clear that only two axioms are really needed. Bigger proofs are not that readable and additional transformations are useful to show the essence of the proof. In the column on the right hand side in square braces one can note the so-called *tactics* used in the proof search – among assumptions and goals, `deny` denotes the denial of the goal (Prover9 uses proofs by contradiction, then paramodulation and resolution is done). Essentially, the more readable counterpart of the proof given by Prover9 is the following *proof object* which contains concrete proof steps.

```
(5 (input) (= (meet v0 (v v0 v1)) v0) NIL)
(9 (input) (= (v v0 (meet v0 v1)) v0) NIL)
(10 (input) (not (= (v (c1) (c1)) (c1))) NIL)
(24 (instantiate 5 ((v0 . v100)))
     (= (meet v100 (v v100 v1)) v100) NIL)
(25 (instantiate 9 ((v0 . v100)
     (v1 . (v v100 v1))))
      (= (v v100 (meet v100 (v v100 v1))) v100)
      NIL)
(26 (paramod 24 (1) 25 (1 2))
     (= (v v100 v100) v100) NIL)
(22 (instantiate 26 ((v100 . v0)))
     (= (v v0 v0) v0) NIL)
(27 (instantiate 22 ((v0 . (c1))))
     (= (v (c1) (c1)) (c1)) NIL)
(23 (resolve 27 () 10 ()) false NIL)
```

## IV. The Mizar System

Formalization of mathematics is a practise of mathematics, or specific mathematical activity, by using a formal language suitable for computerized systems [36]. Here, we mean that a practise of mathematics means acts of proving theorems and correctness of definitions according to classical logic and Zermelo-Fraenkel set theory. Such activity, obviously without the use of computers, is dated back to Peano and Bourbaki, and typically, all areas of mathematics use specific, more-or-less formal languages. Computer certification of mathematics can be useful for many reasons – computers open new possibilities of information analysis and exchange, they can help to discover new proofs or to shed some light on approaches from various perspectives. With the help of such automated proof assistants one can observe deeper connections between various areas of mathematics.

Even if formalization of mathematics (even outside any computerized assistants) could potentially depend only on specific logical layer, set-theoretical counterpart is often, or typically, indispensable in practical applications (of course, one can imagine expressing e.g. Zermelo theorem in terms of pure predicates as Rasiowa and Sikorski mentioned in their *The Mathematics of Metamathematics* [30], but we aim at practically useful programs). Essentially then, alongside with logical connectives, e.g. equivalence, a mathematical system can include equality relation or predicate. Its characteristic properties are those commonly accepted in equivalential logics – reflexivity, symmetry, and transitivity, as we mentioned before.

As the testbed of our considerations we have chosen the MIZAR system, developed by the team we are members of. It was created in the early 1970s in order to assist mathematicians in their work. Now the system consist of three main parts: (1) a special formal language, in which mathematics can be expressed. This language is close to the vernacular used by human mathematician [14]. When used, the language expressions can be automatically checked for grammatical correctness; (2) the software, which verifies the correctness of formalized knowledge, in the classical logical framework; and last, but not least, (3) a huge collection of certified mathematical knowledge – the MML.

Here, we can quote the source written in MIZAR file HIDDEN containing basic built-in properties of primitives – one can see reflexivity and symmetry for the equality. Unfortunately, there is no *transitivity* property implemented for predicates in MIZAR, hence it was hard-coded by the developers of the system.

```
definition let x,y be object;
  pred x = y;
  reflexivity;
  symmetry;
end;
```

In fact, as HIDDEN is one of the two axiomatic files in the MML (the second one is TARSKI, which will be mentioned in the next section), there are no proofs for these properties. The formal language in our examples with MIZAR is pretty close to the every-day language in ordinary mathematics. For precise syntax details, we refer to [1].

## V. Set-theoretic Equality and Its Extensions

If we have primitives for chosen set theory, namely the notion of a set and a primitive set-theoretic membership predicate $\in$, we can express the equality of sets in terms of $\in$ and logical connectives (including quantifiers). Clasically, it is done via extensionality.

$$\forall_X \forall_Y (X = Y \Leftrightarrow (\forall_x (x \in X \Leftrightarrow x \in Y)))$$

```
theorem :: TARSKI:2 :: Extensionality
  (for x being object holds
    x in X iff x in Y) implies
      X = Y;
```

The implication in the opposite direction is provided by the implementation of the system. But obviously, this equality predicate is by no means *new* one. Of course, the above theorem can be read as the form of Leibniz's Law.

Although the notion of the equality of objects is quite concrete, its specific realizations are strongly dependent on how the object is mathematically defined. An illustrative example could be the notion of a rough set [26] treated formally in [8]. On the one hand, rough sets can be defined as the classes of equivalence with respect to a certain relation (it's really meaningless here that numerous generalizations are described in the literature, and partitions are hardly used nowadays in real-life applications of rough set theory). In such approach, two rough sets are equal if the families of subsets are equal (taken componentwise). Thus, we can define the alternative rough equality as below. It is coherent with the equality of ordered pairs, but not necessarily with families of subsets.

```
definition
  let A be Approximation_Space,
      X, Y be Subset of A;
  pred X _=^ Y means
:: ROUGHS_1:def 16
  LAp X = LAp Y & UAp X = UAp Y;
  reflexivity;
  symmetry;
end;
```

Note that the symbol of the rough equality predicate is not the ordinary =, but _=^ denoting the equality of lower and upper approximations. Earlier, we introduced similar predicates in cases of single approximations only [13], [10].

One can consider another approach, credited to Iwiński [16] – rough sets can be viewed as a pair of approximation operators – the lower and the upper one. Within a fixed approximation space, two sets are equal if both approximations are equal (from the very foundational part we can quote Kuratowski's definition of an ordered pair as

$$(a, b) = \{\{a, b\}, \{a\}\}),$$

but majority of mathematicians explore an usual equality of underlying elements.

Here we only mention that even if rough and fuzzy sets have much in common, the equality of fuzzy sets in MIZAR is not that harmless. Because in MML fuzzy sets are corresponding membership functions, their equality is just set-theoretic equality. For formal approach to fuzzy sets, see [11].

## VI. EQUATIONAL CALCULUS IN MIZAR

Equational calculus is mainly performed by a dedicated module EQUALIZER which computes the congruence closure over the collection of equalities accessible at a given inference, where the congruence closure of a relation $R$ defined on a set $A$ is the minimal *congruence relation* (a relation that is simultaneously reflexive, symmetric, transitive, and compatible) containing the original relation $R$. In other words, $R$ is congruence iff it is an equivalence relation satisfying

$$\forall_{x_1,x_2,y_1,y_2 \in A} (x_1, x_2) \in R \Leftrightarrow (y_1, y_2) \in R.$$

The computed congruence closure is used by the MIZAR CHECKER to detect a possible contradiction and to refute the inference (MIZAR is a disprover; by using a technique similar to adding the negation of the goal to the list of available premises can be observed in line 10 of the Prover9 proof in Section III), which can happen if one of the following cases holds

- there are two premises of the form $P[x]$ and $\neg P[y]$ and $x$, $y$ are congruent, or
- there is a premise of the form $x \neq y$ when $x$, $y$ are congruent;

where two elements $x$ and $y$ are congruent, if the pair $(x, y)$ belongs to the congruence closure.

There are many possible sources of equalities, which can be taken into account during the analysis of a given inference. They can be grouped into the following categories:

- occurring explicitly in a given inference,
- term expansions (`equals`),
- properties,
- term reductions,
- term identifications,
- arithmetic,
- type changing (`reconsider`),
- others, e.g. processing structures.

Some of them are of very basic character (e.g. arithmetic), two last ones are extremely dependent on the MIZAR language specification of objects (in this case, structures and type changing statements), but the remaining five are of very general character, and we can easily remap them with the ordinary human reasoning.

The following is an example of a very basic case of equalities, which are stated in the statement to be proved, e.g.:

```
theorem
  for a,b being Element of INT.Ring
    for c,d being Integer st a = c & b = d
      holds a + b = c + d;
```

In the following section(s), we give examples of equalities for the listed categories, at least for those that are not-so-intuitively clear.

## VII. TERM EXPANSIONS AND TERM REDUCTIONS

There are two MIZAR strategies based on substitutions – they act dually in a sense that one of them increases the length of the formula (measured by the number of characters), the other goes in the opposite direction preventing a little bit from uncontrolled growth of terms.

### A. Expansions

One of the methods for defining new functors can use the following syntax:

```
definition
  let $x_1$ be $\theta_1$, $x_2$ be $\theta_2$, ..., $x_n$ be $\theta_n$;
  func $\otimes(x_1, x_2, \ldots, x_n)$ -> $\theta$ equals :ident:
    $\tau(x_1, x_2, \ldots, x_n)$;
  coherence
  proof
    thus $\tau(x_1, x_2, \ldots, x_n)$ is $\theta$;
  end;
end;
```

which introduces a new functor $\otimes(x_1, x_2, \ldots, x_n)$ which is equal to $\tau(x_1, x_2, \ldots, x_n)$. Such definitions, whenever terms $\otimes(x_1, x_2, \ldots, x_n)$ occur in an inference, allow the VERIFIER to generate equalities

$$\otimes(x_1, x_2, \ldots, x_n) = \tau(x_1, x_2, \ldots, x_n).$$

For example,

```
definition
  let x,y be Complex;
  func x - y -> Complex equals
    x + (-y);
  coherence;
end;
```

causes that all instantiations of terms `x-y` are expanded to `x+(-y)`. As a gain of such expansions, for example, the equality `a-b-c = a+(-b-c)` is a direct consequence of associativity of addition. It holds because the term `a-b` is expanded to the term `a+(-b)`, and the term `a-b-c` is expanded to the term `a-b+(-c)`, and both give `a+(-b)+(-c)`. On the other hand, the term `-b-c` is expanded to the term `-b+(-c)`, which creates the term `a+(-b+(-c))`, that is, the associative form of `a+(-b)+(-c)`. An important feature of this kind of term expansions is that it is "a one-way" expansion, in the sense, that terms $\otimes(x_1, x_2, \ldots, x_n)$ are expanded to $\tau\,(x_1, x_2, \ldots, x_n)$ , but not vice-versa. The reason of such treatment is to avoid ambiguity of expansions and over-expanding terms.

### *B. Term Reductions*

Another method of imposing the EQUALIZER to generate extra equalities based on terms occurring in processed inferences are *term reductions*, presented in [17], with the following syntax:

```
registration
  let $x_1$ be $\theta_1$, $x_2$ be $\theta_2$, ..., $x_n$ be $\theta_n$;
  reduce $\tau_1(x_1, x_2, \ldots, x_n)$ to $\tau_2(x_1, x_2, \ldots, x_n)$;
  reducibility
  proof
    thus $\tau_1(x_1, x_2, \ldots, x_n)$ = $\tau_2(x_1, x_2, \ldots, x_n)$;
  end;
end;
```

Term reductions can be used to simplify terms to their proper parts of terms (sub-terms). This simplification relies on matching terms existing in the processed inference with left-side terms of all accessible reductions, and whenever the EQUALIZER meets an instantiation $\sigma$ of the term $\tau_1(x_1, x_2, \ldots, x_n)$, it makes $\sigma$ equal to its sub-term equivalent to $\tau_2(x_1, x_2, \ldots, x_n)$.

The restriction about simplifying terms to their proper subterms, not to any arbitrarily chosen terms, is to fulfill the general rule established for the system, that the EQUALIZER does not generate extra terms and does not expand the universe of discourse.

An example of a possible reduction could be reducing the first power of a complex number to the number (`c` is a subterm of `c|^1`).

example

```
registration
  let c be Complex;
  reduce c|^1 to c;
  reducibility;
end;
```

Reducing the zero power of a number to one is not allowed (`1` is not a sub-term of `c|^0`) – as in the source code below – as a result the following error will be output in a comment (“`::`” starts a comment).

prohibited use of a reduction

```
registration
  let c be Complex;
  reduce c|^0 to 1;
::>        *257
  reducibility;
end;

::> 257: Right term must be
::>      a sub-term of the left term
```

Reductions are recently implemented and their impact on the library is not very big yet. First of all, we can imagine even very complicated reductions, which can be useful in very exceptional cases. The main aim is to reflect human reasoning rather than to force unusual calculations, even if they are straightforward for the machine.

## VIII. PROPERTIES

*Properties* in MIZAR are special formulas, which can be registered while defining functors, see [1], [23]. MIZAR supports `involutiveness` and `projectivity`, for unary operations, and `commutativity` and `idempotence`, for binary operations. If a property is registered for some functor, the EQUALIZER processes appropriate equalities adequate to the property, where for `involutiveness` the equality is

$$f(f(x)) = x,$$

for `projectivity` it is

$$f(f(x)) = f(x),$$

for `commutativity` it is

$$f(x, y) = f(y, x),$$

and for `idempotence`

$$f(x, x) = x.$$

## IX. TERM IDENTIFICATIONS

In mathematics, there are different theories, which at some of their parts are about the same objects. For example, when one considers complex numbers and extended real numbers (reals augmented by $+\infty$ and $-\infty$) and discusses basic operations on such numbers (like addition, subtraction, etc.), it can be quickly recognized that if the numbers are reals, the results of these operations are equal to each other. That is, there is no difference, if one adds reals in the sense of complex numbers or in the sense of extended real numbers. Therefore, pairs of such operations could be identified on appropriate sets of arguments.

MIZAR provides a special construction for such identifications [4] with syntax:

```
registration
  let $x_1$ be $\theta_1$, $x_2$ be $\theta_2$, ..., $x_n$ be $\theta_n$;
  let $y_1$ be $\Xi_1$, $y_2$ be $\Xi_2$, ..., $y_n$ be $\Xi_n$;
  identify $\tau_1(x_1, x_2, \ldots, x_n)$ with $\tau_2(y_1, y_2, \ldots, y_n)$
   when $x_1$ = $y_1$, $x_2$ = $y_2$, ..., $x_n$ = $y_n$;
  compatibility
  proof
   thus $x_1$ = $y_1$ & $x_2$ = $y_2$ & ... & $x_n$ = $y_n$
     implies $\tau_1(x_1, x_2, \ldots, x_n)$ = $\tau_2(y_1, y_2, \ldots, y_n)$;
  end;
end;
```

and, whenever the EQUALIZER meets an instantiation $\sigma$ of the term $\tau_1(x_1, x_2, \ldots, x_n)$, it makes $\sigma$ equal to the appropriate instantiation of $\tau_2(y_1, y_2, \ldots, y_n)$. A gain of using such identifications is that all facts proven about $\tau_2(y_1, y_2, \ldots, y_n)$ are applicable for $\tau_1(x_1, x_2, \ldots, x_n)$, as well.

An example of identification taken from the MIZAR MATHEMATICAL LIBRARY could be the lattice of real numbers with operations `min`, `max` as the infimum and supremum, respectively, of two elements of the lattice, see [5].

```
registration
  let a,b be Element of Real_Lattice;
  identify a "\/" b with max(a,b);
  compatibility;
  identify a "/\" b with min(a,b);
  compatibility;
end;
```

By having such identifications declared, i.e., registered, as `registration`, reasonings about the lattice operations can automatically use facts about real numbers. For example, the associativity of the supremum is a direct consequence of the associativity of the maximum: `max(max(a,b),c) = max(a,max(b,c))`

A less obvious example of such term identifications is connection of lower and upper approximations of rough sets with the topological interior and topological closure, respectively, see [12]. The problem is that the topological closure coincides with the notion of the upper approximation (both possess Kuratowski closure's properties), but topological spaces and approximation spaces are formally distinct structures. We can lift both into some other common one.

```
registration
  let T be with_equivalence
    naturally_generated non empty TopRelStr;
  let A be Subset of T;
  identify LAp A with Int A;
  identify UAp A with Cl A;
end;
```

The latter registration would allow for mixed use of the lower approximation instead of interior operator and vice versa. Here the mathematician's understanding of the identification is as isomorphism (or an analogon) instead of equality.

## X. BUILT-IN COMPUTATIONS

Another source of equalities processed by the EQUALIZER are special built-in procedures for processing selected objects. Generating equalities by these routines is controlled by the environment directive **requirements**, see [23]. In our interest are two procedures dealing with boolean operations on sets (`BOOLE`) and basic arithmetic operations on complex numbers (`ARITHM`).

### *A. Requirements `BOOLE`*

```
X \/ {} = X;    X /\ {} = {};    X \ {} = X;
{} \ X = {};    X \+\ {} = X;
```

### *B. Requirements `ARITHM`*

```
x + 0 = x;    x * 0 = 0;    1 * x = x;
x - 0 = x;    0 / x = 0;    x / 1 = x;
```

Moreover, requirements `ARITHM` provides procedures for solving systems of linear equations over complex numbers.

## XI. TYPE CHANGING STATEMENTS

It is quite common situation, when one object can be treated as an element of different theories or different structures. For example, the empty set is the empty set in set theories, but it is also the zero number is some arithmetics. In computerized mathematics, to allow systems to distinguish and understand symbols clearly and to avoid ambiguity, it is often required to express types of objects explicitly.

MIZAR provides a special rule (**reconsider**) for forcing the system to treat a given term as if its type was the one stated.

For example, to consider the number 0 as an element of the field of real numbers (for example, to prove that it is the neutral element of its additive group), one can state

```
  reconsider z = 0 as Element of F_Real;
```

The equality `z = 0` is obviously processed by the EQUALIZER.

## XII. EXAMPLE WITH AUTOMATIZATION

In this section we present an example how all described above techniques can automatize reasoning and make proofs shorter or even make theorems obvious. The working example (about elements of the additive group of real numbers `G_Real`) with all automatizations switched-off and all basic proof steps written within the proof is as follows:

```
theorem
  for a being Element of G_Real holds
    a + 0.G_Real = a
  proof
    let a be Element of G_Real;
    reconsider x = a as Real;
B:  0 in REAL by XREAL_0:def 1;
A:  0.G_Real = the ZeroF of G_Real
                by STRUCT_0:def 6
      .= In(0,REAL) by VECTSP_1:def 1
      .= 0 by B,SUBSET_1:def 8;
    thus a + 0.G_Real
      = (the addF of G_Real).(a,0.G_Real)
                by ALGSTR_0:def 1
     .= addreal.(a,0.G_Real) by VECTSP_1:def 1
     .= x + 0 by A,BINOP_2:def 9
     .= x by ARITHM:1
     .= a;
  end;
```

while the theorem is obvious when all provided mechanism are utilized.

The equality

```
  a + 0.G_Real =
    (the addF of G_Real).(a,0.G_Real);
```

is a consequence of the "equals" expansion of the definition:

```
definition
  let M be addMagma;
  let a,b be Element of M;
  func a+b -> Element of M equals
:: ALGSTR_0:def 1
  (the addF of M).(a,b);
end;
```

The equality `a + 0.G_Real = x + 0` is a consequence of the equality `x = a` (reconsider), the equality `0.G_Real = 0` and the term identification:

```
registration
  let a,b be Element of G_Real, x,y be Real;
  identify a+b with x+y
  when a = x, b = y;
  compatibility by BINOP_2:def 9;
end;
```

The equality `0.G_Real = 0` is a consequence of the "equals" expansion of the definition:

```
definition
  let S be ZeroStr;
  func 0.S -> Element of S equals
:: STRUCT_0:def 6
  the ZeroF of S;
end;
```

and the "equals" expansion of the definition:

```
definition
  func G_Real -> strict addLoopStr equals
:: VECTSP_1:def 1
  addLoopStr (# REAL,addreal,In(0,REAL) #);
end;
```

and the term reduction:

```
registration
  let r be Real;
  reduce In(r,REAL) to r;
  reducibility;
end;
```

The equality `x + 0 = x` is a consequence of built-in calculations over complex numbers. Finally, the equality `x = a` is a trivial consequence of the "reconsider".

What is especially useful, the distribution of MIZAR contains programs which detect if the reference (or the proof step) is really necessary for the checker. Alternatively, we can test by brute force if the construction is useful in a specific case: adding an environment directive to the preamble of the article [25] and call the cleaning utilities; if no changes will be done, the directive is deleted. Of course, there is no need to add all directives to all articles (some of them can be just from another area of mathematics), note also that even useful constructions, say expansions, can significantly slown down the proof checking by expanding the universe of discourse.

An illustrating example was the splitting of the equality of sets into two inclusions: it is useful, but even if, according to the von Neumann construction, all natural numbers are sets, the expansion of the equality of numbers into two inclusions isn't especially feasible (even if mathematically reasonable, but who would prove inclusions like $2 \subseteq 3$, perhaps besides the person interested in the arithmetic of ordinals?).

Every MIZAR article contains a preamble with the list of files from MML which will be used. There are 10 keywords for that (and respectively, 10 environment directives). We focus only on those tightly connected with our paper, mainly of three items:

- registrations, responsible for various registrations of clusters, including reductions and identifications;
- equalities, which allow to expand definitions given by `equals`;
- requirements, which turn on (mainly arithmetical or set-theoretic) calculations.

With all the automatics switched on, one can obtain the formula from the third section

```
a "\/" a = a;
```

as a result of declared reduction (or alternatively, the `idempotence` property). This however cannot be enough for a human user who wants to see what is really going on here, i.e.:

```
a "\/" a = (a "/\" (a "\/" a)) "\/" a
                              by LATTICES:def 9
        .= a by LATTICES:def 8;
```

where both definitions are actually axioms – the absorption laws defined in the form of attributes.

Of course, the equality can be introduced in a non-explicit way as in the examples of problems of purely equational character. Remember that one can, apart from the above considerations, register the following cluster:

```
registration
  let X be set;
  cluster X \ X -> empty;
  coherence;
end;
```

which is effectively equivalent to

```
  X \ X = {};
```

adding it automatically as one of the premises to every inference where applicable. Remember that in order to assure the needed identification, one should not use the reduction ($\emptyset$ is not a sub-term of `X \ X`).

```
registration
  let T be TopSpace,
      A be 1st_class Subset of T;
  cluster Border A -> empty;
end;
```

To quote less straightforward example, in the area of rough sets (expressed in Isomichi style) it can be the identification of first class subsets with crisp sets, consequently their borders are the empty sets as the set-theoretic difference $A \setminus A$. The explanation and the details of the corresponding implementation of rough sets can be found in [13].

## XIII. Some Statistics

We illustrate our discussion by the number of concrete constructions used within the MML by Table I.

Table II shows how often selected properties are declared (and proved) in MML. The numbers however can be more or less accidental: authors can just omit a property when developing some theories, and can state it in a form of explicit formula, or just the proof of it can be too complex. In Table

Table I
OCCURRENCES OF BASIC CONSTRUCTIONS IN MML

| Construction | keyword | occurrences |
|---|---|---|
| type changing statements | `reconsider` | 66115 |
| equalities | `equals` | 3678 |
| identifications | `identify` | 172 |
| reductions | `reduce` | 139 |

Table II
OCCURRENCES OF PROPERTIES IN MML

| Property | occurrences |
|---|---|
| `commutativity` | 150 |
| `idempotence` | 18 |
| `involutiveness` | 36 |
| `projectivity` | 20 |

III, we listed how many times in MIZAR articles (among 1240 in total in the whole MML) selected properties were used. We have chosen among three areas: basic set theory, arithmetic of complex numbers, and some properties of pairs. Even if MML is based on ZFC set theory, the calculations are used very extensively.

Here we can point out the scale we refer to. There are about 11 thousand definitions and 55 thousand theorems proven in MML. This is contained among 2 million lines of code (about 90 MB of mathematical texts). One can ask a question why the commutativity of the set-theoretic union is used less often than that of set-theoretic intersection. Similar (although dual) issue with arithmetic is clear: the complex addition is just used more frequently. Symmetric difference $X \dot{-} Y$ is just of marginal interest and, which can be also a kind of explanation after deeper research, it is defined as equal to the combination of the other operations, namely $X \dot{-} Y = (X \setminus Y) \cup (Y \setminus X)$, where the commutativity of $\cup$ (understood automatically) can do some useful work.

## XIV. VARIOUS MATHEMATICAL EQUALITIES

Equality in mathematics has many facings; of course there are the usual identities $x = y$, such as for example $2+1 = 3$ or $(x+1)^2 = x^2+2x+1$. Note however, that even in these cases equality depends on the domain we are working in: $2+1 = 3$ is true in $\mathbb{Z}$ but not in $\mathbb{Z}_2$. The second equation can be seen as an equation between (real) numbers or between polynomials (not necessarily) over the real numbers.

Table III
THE NUMBER OF CONCRETE USES OF PROPERTIES

| Property | No. of Mizar articles | No. of implicit uses |
|---|---|---|
| $X \cup Y = Y \cup X$ | 483 | 4721 |
| $X \cap Y = Y \cap X$ | 547 | 6232 |
| $X \dot{-} Y = Y \dot{-} X$ | 7 | 57 |
| $a + b = b + a$ | 670 | 9548 |
| $ab = ba$ | 429 | 5734 |
| $\{x, y\} = \{y, x\}$ | 118 | 857 |

There are other kinds of equality. Among the most important ones is identifying objects of a given structure with respect to a certain property. Formally, this is described by an equivalence relation $r$, for example $r_p(x, y)$, if $x$ and $y$ are integers such that $x$ mod $p$ = $y$ mod $p$. This in particular is applied when, for example, constructing $\mathbb{Z}_p$, the integers modulo $p$: All integers leaving the same remainder when dividing them by $p$ are identified using the equivalence relation $r_p$. Formally, equivalence classes with respect to $r_p$ are built, and these are the elements of $\mathbb{Z}_p$ – on which the usual operations are defined. Using this construction $\mathbb{Z}_p$ does not contain any numbers, this is later achieved by "identifying an equivalence class with a number", so that $\mathbb{Z}_p = \{0, 1, \ldots, p-1\}$ is used. Note, however, that this formally also means "changing the operations of $\mathbb{Z}_p$ appropriately" to match with elements of $\{0, 1, \ldots, p-1\}$.

In some sense this last step – going from equivalence classes to non-negative integers smaller than $p$ – can be seen as changing the representation of the integers modulo $p$. In fact it does not matter whether we consider $\mathbb{Z}_p$ as a set of equivalence classes or as $\{0, 1, \ldots, p-1\}$. Aside from the names of the objects, the two structures are the same: The objects behave the same way, that is, the effect of the operations is identical.

This in fact is the most important kind of mathematical equality: The two structures are isomorphic, that is there exists a (bijective) function $i$ from $\mathbb{Z}_p$ to $\{0, 1, \ldots, p-1\}$ respecting the domain's operations, i.e. $i(x+y) = i(x) + i(y)$, and similarly for all other operations of $\mathbb{Z}_p$. From a mathematical point of view, this completely defines an equality in the sense of the first section: Properties discovered for one of the structures automatically carry over to the other one. In a proof-assistant, this however is not that obvious. First of all, both structures must have been constructed in the prover to describe the isomorphism. This is not always the case and is often technical and tedious, or impossible. Moreover, by having a property/theorem for one of the isomorphic structures, usually some work remains to be done in order to have the same property/theorem for the other one. For example $\mathbb{Z}_p$ is a field, if $p$ is prime. Proved in one of the isomorphic structures, carrying it over to the other one requires to translate the property with the help of the isomorphism $i$. Here of course the theorem that $i(x*y) = i(x)*i(y)$, where $y$ is the inverse of $x$ does the trick. The interesting discussion from the category theorist's point of view is given in [19].

Observe that our discussion here is not in contradiction with the Leibniz's Law mentioned at the beginning of our paper. Such an identity of indiscernibles does not coincide with equality via isomorphisms. For example, a field is prime, if it does not have any proper subfield. One could now expect that two isomorphic fields are either both prime or not. This, however, is not the case as the isomorphic fields $K(X)$ and $K(X^2)$ show. We will therefore mostly deal with equality via isomorphisms.

## XV. Data Structure Equality

From a computer scientist's view, the equality is not that easy as Leibniz's Law could suggest. Almost immediately one observes that, e.g. in a programming language, it is quite a difference whether the objects to be equal are just plain numbers or more structured objects such as arrays or trees. For short, equality, and consequently its handling, heavily depends on the object's types. In programming languages, of course, there exist a number of techniques allowing users to exactly define equality of objects of a given type, e.g. via type classes in Haskell [32]. In general, intelligent systems should be able to handle equalities of different kinds, even delivered by external tools [24].

Proof assistants dealing with the notion of a mathematical proof, therefore, should provide means for equalities occurring in mathematics. In mathematics, however, the notion of equality is much more elaborated. To give an easy example, two functions $f$ and $g$ are equal if their domain $D$ and codomain $C$ are equal, and if $f(x) = g(x)$, for all $x \in D$. Of course, this apparent incoherence can be easily explained considering natural hierarchy of notions many mathematicians don't use in their proofs. For example, in typical set-theoretic approaches, functions are just binary relations of a special type, and relations are subsets of the Cartesian product, which is a set of corresponding ordered pairs. Here equality takes into account not only two/four additional sets, but also equality of a – possibly infinite – number of objects of another type.

When it comes to structures, such as e.g. groups, fields or topologies, equality becomes even more sophisticated. This is not only due to the growing complexity of the objects; mathematicians have developed a special style of informal handling with various kinds of equality. The point is that, in proof assistants, each kind of equality between objects of a given type has to be formally defined, in order to prove that two objects of that type are equal.

## XVI. Isomorphisms and Inclusions

For a given field $F$, one can build the ring of polynomials $F[X]$. Actually, while the construction applies to arbitrary rings, our formalization work, however, concentrates on field theory, so we restrict ourselves to fields here. Elements of $F[X]$ – polynomials over $F$ – are basically functions from the natural numbers into $F$, hence obviously different from elements of $F$. Consequently, $F \subseteq F[X]$ cannot hold from a formal point of view. In particular one cannot prove this with the ordinary definition of $\subseteq$. Even if this would be possible, note that $F \subseteq F[X]$ here means not that $F$ is a subset of $F[X]$, but that $F$ is a subfield of $F[X]$, that is addition and multiplication of $F[X]$ are taken into account. Nevertheless, [35] states the following

> *We regard $F \subseteq F[X]$ by identifying the element $a \in F$ with the constant polynomial $a \in F[X]$, and observe that under this identification the units of $F[X]$ are precisely the nonzero elements of $F$.*

What is hidden in this statement, is the application of a mapping $i$ sending an element $a \in F$ to the constant polynomial $p(X) = a$. The mapping $i$ is an isomorphism between $F$ and its $i$-image in $F[X]$ (which is a subfield of $F[X]$). Thus, identifying $a$ with $a(x)$ actually means replacing the image of $i$ with $F$, which then formally gives an isomorphic copy $C$ of $F$ under the isomorphism $i$, such that $C \subseteq F[X]$:

```
theorem
  for F being Field
   ex R being Ring st
     R, Polynom-Ring F are_isomorphic &
       F is a Subfield of R;
```

and

```
theorem
  for F being Field,
      a being Element of F holds
    a|F is Unit of Polynom-Ring F
      iff a is non zero;
```

This is nice and handy, because it allows to work with the easier objects of $F$ when appropriate. In a proof assistant, however, it gives rise to quite a number of additional work and theorems: one has to define $i$.

Working with polynomials, however, is not the same in isomorphic polynomial rings, even if we restrict ourselves to polynomial rings over isomorphic fields. One has to apply the isomorphism $i$ to translate from one into the other. Note, that formally $i \circ p$ is an extension of $i$ transforming $p \in F[X]$ into a polynomial over $E$, that is $i$ is applied to all of $p$'s coefficients:

```
theorem
  for F, E being Field,
      p being Polynomial of F,
      x being Element of F,
      i being Isomorphism of F,E holds
    i(p(x)) = (i p)(i x);

theorem
  for F, E being Field,
      p being Polynomial of F,
      x being Element of F,
      i being Isomorphism of F,E holds
    p(x) = 0.F iff (i p)(i x) = 0.E;
```

Things can get even worse. The quotient field $F[X]/<p(X)>$, for a irreducible polynomial $p$, formally consists of equivalence classes $[q(X)]_{<p(X)>}$, where $q(X) \in F[X]$ – this is the field in which $p(X)$ has a root. However, in [35], after showing that $F[X]/<p(X)>$ is a field, the author states:

> *... to give an explicit description of the field $E = F[X]/<p(X)>$. Let $p(X) \in F[X]$ be an irreducible polynomial of degree $n$, $p(X) = a_nX^n + \cdots + a_0$. By taking representatives of equivalence classes we may regard*
>
> $$E = \{g(X) \in F[X] \mid deg\ g(X) < n\}$$
>
> *as a set. Addition in $E$ is the usual addition of polynomials, while multiplication in $E$ ...*

Actually, this means only that $E$ with the new defined operations is isomorphic to $F[X]/{<}p(X){>}$. Afterwards, the following consequence is drawn in [35]:

> *If $\pi : F[X] \longrightarrow F[X]/{<}p(X){>}$ is the canonical projection, then $\pi|_F$ is an injection, and using this we also regard $F \subseteq F[X]/{<}p(X){>}$.*

## XVII. Conclusion

Even if mathematicians have developed and used some human "mechanisms of obviousness", for hundreds of years, new technologies impose the need for new standards. Computerized systems should undoubtfully deliver the answers of some human questions and to construct models, which cannot be created by hand in a reasonable time – this is the testbed for computer algebra systems and model finders. Proof assistants, in order not to break human standards, should however support the traditional way of thinking, so there is a place for finding a reasonable balance between writability and readability of the source code, as we believe the Mizar system can deliver. We hope that we convinced the reader that the problem of the equality treatment is harder than it looks like at the very first sight, and if automated proof-assistants are taken into account, we should take care on something more than what we called absolute equality – Davenport's data structure equality.

## References

[1] G. Bancerek, C. Byliński, A. Grabowski, A. Korniłowicz, R. Matuszewski, A. Naumowicz, and K. Pąk, Mizar: state-of-the-art and beyond. In M. Kerber et al. (Eds) *Intelligent Computer Mathematics,* Lecture Notes in Artificial Intelligence, 9150, 261–279, Springer, 2015. doi:10.1007/978-3-319-20615-8_17

[2] M. Bartoszuk, Solving systems of polynomial equations: a novel end condition and root computation method. In *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, FedCSIS 2014,* M. Ganzha, L. Maciaszek, M. Paprzycki (Eds.), p. 543–552, IEEE, 2014. doi:10.15439/2014F183

[3] R. Bradford, J.H. Davenport, and C.J. Sangwin, A comparison of equality in computer algebra and correctness in mathematical pedagogy. In *Intelligent Computer Mathematics,* Lecture Notes in Computer Science, 5625, 75–89, Springer, 2009. doi:10.1007/978-3-642-02614-0_11

[4] M. Caminati, Custom automations in Mizar. *Journal of Automated Reasoning*, 50(2), 147–160, 2013. doi:10.1007/s10817-012-9266-1

[5] M. Chmur, The lattice of real numbers. The lattice of real functions. *Formalized Mathematics*, 1(**4**):681–684, 1990.

[6] Z.E. Csajbòk and T. Mihálydeák, Fuzziness in partial approximation framework. In *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, FedCSIS 2013,* M. Ganzha, L. Maciaszek, M. Paprzycki (Eds.), p. 35–41, IEEE, 2013.

[7] J.H. Davenport, Equality in computer algebra and beyond. *Journal of Symbolic Computation,* 34(4), 259–270, 2002. doi:10.1006/jsco.2002.0551

[8] A. Grabowski, Efficient rough set theory merging. *Fundamenta Informaticae,* 135(4), 371–385, 2014. doi:10.3233/FI-2014-1129

[9] A. Grabowski, Mechanizing complemented lattices within Mizar type system. *Journal of Automated Reasoning,* 55(3), 211-221, 2015, doi:10.1007/s10817-015-9333-5

[10] A. Grabowski, On the computer-assisted reasoning about rough sets. In *Monitoring, Security and Rescue Techniques in Multiagent Systems*, B. Dunin-Kęplicz, A. Jankowski et al. (Eds.), Advances in Soft Computing, 28, 215–226, Springer, 2005. doi:10.1007/3-540-32370-8_15

[11] A. Grabowski, On the computer certification of fuzzy numbers. In *Proceedings of 2013 Federated Conference on Computer Science and Information Systems, FedCSIS 2013,* M. Ganzha, L. Maciaszek, M. Paprzycki (Eds.), 51–54, IEEE, 2013.

[12] A. Grabowski, Topological interpretation of rough sets. *Formalized Mathematics*, 22(1):89–97, 2014. doi:10.2478/forma-2014-0010

[13] A. Grabowski and M. Jastrzębska, Rough set theory from a math-assistant perspective. In *Rough Sets and Intelligent Systems Paradigms,* M. Kryszkiewicz et al. (Eds.), Lecture Notes in Artificial Intelligence, 4585, 152–161, Springer, 2007. doi:10.1007/978-3-540-73451-2_17

[14] A. Grabowski and Ch. Schwarzweller, *Translating mathematical vernacular into knowledge repositories*. In: M. Kohlhase (ed.), Proceedings of the 4th International Conference on Mathematical Knowledge Management, Lecture Notes in Artificial Intelligence, 3863, 49–64, Springer Heidelberg, 2006. doi:10.1007/11618027_4

[15] G. Grätzer, *General Lattice Theory.* Birkhäuser, 1998.

[16] T.B. Iwiński, Algebraic approach to rough sets, *Bull. Polish Acad. Sci. Math.,* 35, 673–683, 1987.

[17] A. Korniłowicz, On rewriting rules in Mizar. *Journal of Automated Reasoning,* 50(2), 203–210, 2013. doi:10.1007/s10817-012-9261-6

[18] A. Korniłowicz, Definitional expansions in Mizar. *Journal of Automated Reasoning,* 55(3), 257–268, 2015. doi:10.1007/s10817-015-9331-7

[19] B. Mazur, When is one thing equal to some other thing? In *Proof and Other Dillemas: Mathematics and Philosophy,* B. Gold and R. Simons (Eds.), Spectrum, 2008.

[20] W. McCune, Prover9 and Mace4. Available at http://www.cs.unm.edu/~mccune/prover9/, 2005–2010.

[21] W. McCune, Solution of the Robbins problem. *Journal of Automated Reasoning,* 19(3), 263–276, 1997. doi:10.1023/A:1005843212881

[22] G. Moreno, J. Penabad, and C. Vázquez, Fuzzy logic rules modeling similarity-based strict equality. In *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki (Eds.), 119–128, IEEE, 2014. doi:10.15439/2014F387

[23] A. Naumowicz, Improving Mizar texts with properties and requirements. In *Mathematical Knowledge Management 2004*, A. Asperti, G. Bancerek, A. Trybulec (Eds.), Lecture Notes in Computer Science, 3119, 290–301, Springer Heidelberg, 2004. doi:10.1007/978-3-540-27818-4_21

[24] A. Naumowicz, Automating Boolean set operations in Mizar proof checking with the aid of an external SAT solver, *Journal of Automated Reasoning,* 55(3), 285–294, 2015, doi:10.1007/s10817-015-9332-6

[25] A. Naumowicz and A. Korniłowicz, A brief overview of Mizar. In *Theorem Proving in Higher Order Logics 2009,* S. Berghofer, T. Nipkow, Ch. Urban, M. Wenzel (Eds.), Lecture Notes in Computer Science, 5674, 67–72, Springer Heidelberg, 2009. doi:10.1007/978-3-642-03359-9_5

[26] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data.* Kluwer, Dordrecht, 1991.

[27] K. Pąk, Improving legibility of formal proofs based on the close reference principle is NP-hard, *Journal of Automated Reasoning,* 55(3), 295–306, 2015. doi:10.1007/s10817-015-9337-1

[28] K. Pąk, Improving legibility of natural deduction proofs is not trivial. *Logical Methods in Computer Science,* 10(3), 1–30, 2014. doi:10.2168/LMCS-10(3:23)2014

[29] K. Pąk, Methods of lemma extraction in natural deduction proofs. *Journal of Automated Reasoning,* 50(2), 217–228, 2013. doi:10.1007/s10817-012-9267-0

[30] H. Rasiowa and R. Sikorski, *The Mathematics of Metamathematics,* Polish Scientific Publishers, Warsaw, 1963.

[31] P. Rudnicki and A. Trybulec, On equivalents of well-foundedness. *Journal of Automated Reasoning*, 23, 197–234, 1999. doi:10.1023/A:1006218513245

[32] S. Thompson, *Haskell: The Craft of Functional Programming*. 3rd ed. Addison-Wesley, 2011.

[33] A. Trybulec, A. Korniłowicz, A. Naumowicz, and K. Kuperberg, Formal mathematics for mathematicians. *Journal of Automated Reasoning,* 50(2), 119–121, 2013. doi:10.1007/s10817-012-9268-z

[34] J. Urban and G. Sutcliffe, Automated reasoning and presentation support for formalizing mathematics in Mizar. In *Intelligent Computer Mathematics*, S. Autexier et al. (Eds.), Lecture Notes in Computer Science, 6167, 132–146, Springer Heidelberg, 2010. doi:10.1007/978-3-642-14128-7_12

[35] S. Weintraub, *Galois Theory*. 2nd edition, Springer-Verlag, 2009.

[36] F. Wiedijk, Formal proof – getting started. *Notices of the AMS,* 55(11), 1408–1414, 2008.

[37] L. Zadeh, Fuzzy sets. *Information and Control,* 8(3), 338–353, 1965.

[38] S. Żukowski, Introduction to lattice theory. *Formalized Mathematics,* 1(1), 215–222, 1990.

DOI: 10.15439/2015F14
ACSIS, Vol. 5

# Application of Artificial Neural Network and Support Vector Regression in Cognitive Radio Networks for RF Power Prediction Using Compact Differential Evolution Algorithm

Sunday Iliya, Eric Goodyer, John Gow, Jethro Shell and Mario Gongora
Centre for Computational Intelligence,
School of Computer Science and Informatics,
De Montfort University, The Gateway,
Leicester LE1 9BH, England, United Kingdom
Email: sundayiliyagoteng@yahoo.com, eg@dmu.ac.uk, jgow@dmu.ac.uk
jethros@dmu.ac.uk, mgongora@dmu.ac.uk

***Abstract*—Cognitive radio (CR) technology has emerged as a promising solution to many wireless communication problems including spectrum scarcity and underutilization. To enhance the selection of channel with less noise among the white spaces (idle channels), the a priory knowledge of Radio Frequency (RF) power is very important. Computational Intelligence (CI) techniques cans be applied to these scenarios to predict the required RF power in the available channels to achieve optimum Quality of Service (QoS). In this paper, we developed a time domain based optimized Artificial Neural Network (ANN) and Support Vector Regression (SVR) models for the prediction of real world RF power within the GSM 900, Very High Frequency (VHF) and Ultra High Frequency (UHF) FM and TV bands. Sensitivity analysis was used to reduce the input vector of the prediction models. The inputs of the ANN and SVR consist of only time domain data and past RF power without using any RF power related parameters, thus forming a nonlinear time series prediction model. The application of the models produced was found to increase the robustness of CR applications, specifically where the CR had no prior knowledge of the RF power related parameters such as signal to noise ratio, bandwidth and bit error rate. Since CR are embedded communication devices with memory constrain limitation, the models used, implemented a novel and innovative initial weight optimization of the ANN's through the use of compact differential evolutionary (cDE) algorithm variants which are memory efficient. This was found to enhance the accuracy and generalization of the ANN model.**

***Index Terms*—Cognitive Radio; Primary User; Artificial Neural Network; Support Vector Machine; Compact Differential Evolution; RF Power; Prediction.**

## I. Introduction

DUE TO the current static spectrum allocation policy, most of the licensed radio spectrum are not maximally utilized and often free (idle) while the unlicensed spectrum are overcrowded. Hence the current spectrum scarcity is the direct consequence of static spectrum allocation policy and not the fundamental lack of spectrum. The first bands to be approved for CR communication by the US Federal Communication Commission (FCC) because of their gross underutilization in time, frequency and spatial domain are the very high frequency and ultra-high frequency (VHF/UHF) TV bands [1] [2] [3]. In this paper, we focused on the study of real world RF power distribution in some selected channels (54MHz to 110MHz, 470MHz to 670MHz, 890MHz to 908.3MHz GSM up-link, 935MHz to 953.3MHz GSM down-link) within the VHF/UHF bands, FM band, and the GSM 900 band. The problem of spectrum scarcity and underutilization, can be minimized by adopting a new paradigm of wireless communication scheme. Advanced Cognitive Radio (CR) network or Adaptive Spectrum Sharing (ASS) is one of the ways to optimize our wireless communications technologies for high data rates in a dynamic environment while maintaining user desired quality of service (QoS) requirements. CR is a radio equipped with the capability of awareness, perception, adaptation and learning of its radio frequency (RF) environment [4]. CR is an intelligent radio where many of the digital signal processing that were traditionally done in static hardware are implemented via software. Irrespective of the definition of CR, it has the followings basic features: observation, adaptability and intelligence. CR is the key enabling tool for dynamic spectrum access and a promising solution for the present problem of spectrum scarcity and underutilization. Cognitive radio network is made up of two users i.e. the license owners called the primary users (PU) who are the incumbent legitimate owners of the spectrum and the cognitive radio commonly called the secondary users (SU) who intelligently and opportunistically access the unused licensed spectrum based on some agreed conditions. CR access to licensed spectrum is subject to two constrains i.e on no interference base, this implies that CR can use the licensed spectrum only when the licensed owners are not using the channel (the overlay CR scheme). The second constrain is on the transmitted power, in this case, SU can coexist with the PU as long as the interference to the PU is below a given threshold which will not be harmful to the PU nor degrade the QoS requirements of the PU (the underlay CR network scheme) [5] [1]. There are four major steps involved in cognitive radio network, these are: spectrum sensing, spectrum decision, spectrum sharing, and spectrum mobility [6] [7].

978-83-60810-66-8/$25.00©2015, IEEE

In spectrum sensing, the CR senses the PU spectrum using either energy detector, cyclostationary features detector, cooperative sensing, match filter detector, eigenvalue detector, etc to sense the occupancy status of the PU [8]. Based on the sensing results, the CR will take a decision using a binary classifier to classify the PU channels (spectrum) as either busy or idle there by identifying the white spaces (spectrum holes or idle channels). Spectrum sharing deals with efficient allocation of the available white spaces to the CR (SU) within a given geographical location at a given period of time while spectrum mobility is the ability of the CR to vacate the channels when the PU reclaimed ownership of the channel and search for another spectrum hole to communicate. During the withdrawal or search period, the CR should maintain seamless communication. Many wireless broadband devices ranging from simple communication to complex systems automation, are deployed daily with increasing demand for more, this calls for optimum utilization of the limited spectrum resources via CR paradigm. Future wireless communication device should be enhanced with cognitive capability for optimum spectrum utilization. CRs are embedded wireless communication devices with limited memory, thus in this paper, we utilized the power of compact differential evolutionary (cDE) algorithm which is memory efficient, to develop an optimized ANN and SVR model for the prediction of real world radio frequency (RF) power. RF power traffics is a function of time, geographical location (longitude and latitude), height above the sea level (altitude) and the frequency or channels properties. Since our experiment is conducted at a fixed geographical location and at constant height, the inputs of the ANN and SVR consist of only past RF power samples, current time domain information and frequency (channel) while the output is the predicted current RF power in decibel (dB) (i.e. the current RF power is modelled as a function of time, frequency and past RF power samples) hence forming a nonlinear time series prediction model. ANN and SVR models were adopted because of the dynamic nonlinearity often associated with RF traffic pattern, coupled with random interfering signals or noise resulting from both artificial and natural sources. The use of sensitivity analysis as detailed in Section VIII for the determination of the optimum number of past recent RF power samples to be used as part of the input of the ANN or SVR for prediction of current RF power, results into a more compact, robust, accurate, and well generalized models. The proposed algorithm used a priori data to enable the system to avoid noisy channels. The prior knowledge of the RF power allowed the cognitive radio to predictively select channels with the least noise among those that were unused or free. This would allow for a reduced utilization of radio resources including transmitted power, bandwidth, and in turn maximizing the usage of the limited spectrum resources. The data used in this study was obtained by capturing real world RF data for two months using Universal Software Radio Peripheral 1 (USRP 1). The digital signal processing and capturing of the data were done using gnuradio which is a combination of Python for scripting and C++ for signal processing blocks; while the models design and prediction were done in Matlab. The experiment was conducted at Centre for Computational Intelligence, De Montfort University, UK, located very close to Leicester city centre.

Many prediction models used in CR radio uses known RF related parameters as their inputs of which licensed owners will not be willing to dispose such information to CR users. Some of the models are based on explicit mathematical model which may be different from real world situation as highlighted in Section II. Some of the prediction models aim at prediction of spectrum holes, but the fact that spectrum holes (vacant channels) are known does not depict any information about the best channel to be used among the idle channels as the noise level is not flat for all the channels. Thus the major contribution of our model is that it can be used for Rf power prediction where the CR has no prior knowledge of any RF power related parameter. This will enable the CR to avoid noisy channels. The model is trained and tested using real world data. Also instead of training the ANN using back propagation algorithms (BPA) which often lack optimality due to premature convergent, the weights of the ANN are initially evolves using cDE and then fine tune using BPA, this was found to produce a more accurate and generalized model as compared with the one trained using only BPA. SVR was also examined using different kernels and we come up with the model that is more appropriate for our studied location.

The rest of this paper is consist of the following sections. Section II consist of previously presented related research in this field. This will be followed by Section III and Section V, that gives brief description of neural network and the optimization algorithms implemented. Experimental details are discussed in Section VII. The paper is concluded with Section IX, which discusses the results of the experiments, Section X gives the summary of the findings.

## II. Related Work

There are different types and variants of Computational Intelligence (CI) and machine learning algorithms that can be used in CR such as genetic algorithms for optimization of transmission parameters [9], swarm intelligence for optimization of radio resource allocation [10], fuzzy logic system (FLS) for decision making [11] [12], neural network and hidden Markov model for prediction of spectrum holes; game theory, linear regression and linear predictors for spectrum occupancy prediction [13] , Bayesian inference based predictors, etc. Some of the CI methods are used for learning and prediction, some for optimization of certain transmission parameters while others for decision making [14]. TV idle channels prediction using ANN was proposed in [15], however, data were collected only for two hours everyday day (5pm to 7pm) within a period of four weeks, this is not sufficient to capture all the various trends associated with TV broadcast. Also, identifying the idle channels does not depict any spatial or temporal information of the expected noise and/ or level of interference based on the channels history which is vital in selecting the channels to be used among the idle channels. Spectrum hole prediction using Elman recurrent artificial neural network (ERANN) was proposed in [16]. It uses the cyclostationary features of modulated signals to determine the presence or absence of primary signals while the input of the ERANN consists of time instances. The inputs and the target output used in the training of the ERANN and prediction were modelled using ideal multivariate time series equations, which are often different from real life RF traffics where PU signals can be embedded in noise and/ or interfering signals. Traffic pattern prediction using seasonal autoregressive integrated moving-average (SARIMA) was proposed for reduction of CRs hopping rate and interference effects on PU while maintaining a fare blocking rate [17]. The model (SARIMA) does not depict any information about the expected noise power.

Fuzzy logic (FL) is a CI method that can capture and represent uncertainty. As a result it has been used in CR research for decision making processes. In [11] an FL based decision-making system with a learning mechanism was developed for selection of optimum spectrum sensing techniques for a given band. Among these techniques are matched filtering, correlation detection, features detection, energy detection, and cooperative sensing. Adaptive neural fuzzy inference system (ANFIS) was used for prediction of transmission rate [18]. This model was designed to predict the data rate (6, 12, 24, 36, 48 and 54 Mbps) that can be achieved in wireless local area network (WLAN) using a 802.11a/g configuration as a function of time. The training data set was obtained by generating a random data rate with an assigned probability of occurrence at a given time instance, thus forming a time series. In this study, real world RF data wasn't used. More importantly, the research did not take into account the dynamic nature of noise or interference level which can affect the predicted data rates. Semi Markov model (SMM) and continuous-time Markov chain (CTMC) models have also been used for the prediction of packet transmission rates [19]. This avoids packet collisions through spectrum sensing and prediction of temporal WLAN activities combined with hoping to a temporary idle channel. However, SMM are not memory efficient, neither was there any reference made to the expected noise level among the inactive (idle) channels to be selected. An FL based decision system was modeled for spectrum hand-off decision-making in a context characterized by uncertain and heterogeneous information [12] and fuzzy logic transmit power control for cognitive radio. The proposed system was used for the minimization of interference to PU's while ensuring the transmission rate and quality of service requirements of secondary users [20]. The researcher did not, however, include any learning from past experience or historical data. An exponential moving average (EMA) spectrum sensing using energy prediction was implemented in [21]. The EMA achieved a prediction average mean square error (MSE) of 0.2436 with the assumption that the channel utilization follow exponential distribution with rate parameter $\lambda = 0.2$ and signal to noise (SNR) of 10dB; RF real world data was not used in their study. Within this paper we demonstrate the use of SVR and an ANN trained using cDE for prediction of real world RF power of selected channels within the GSM band, VHF and UHF bands. An optimized ANN model was produced by combining the global search capabilities of cDE algorithm variants and the local search advantages of back-propagation algorithms (BPA). The initial weights of the ANN were evolved using cDE after which the ANN was trained (fine tune) more accurately using back-propagation algorithms. This methodology demonstrates the application of previously acquired real world data to enhance the prediction of RF power to assist the implementation of CR applications. The meta parameters that govern the accuracy and generalization SVR model were evolves using cDE.

## III. Artificial Neural Network

Artificial Neural Networks (ANN) are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems [22], [23]. Due to the dynamic nonlinearity often associated with RF traffic pattern, coupled with random interfering signals or noise resulting from both artificial and natural sources, a fully connected multilayer perceptron (MLP) ANN with two hidden layers was used in this study. The input layer was cast into a high dimensional first hidden layer for proper features selection. The activation functions used in the two hidden layers are nonlinear hyperbolic tangent functions (1), and a linear symmetric straight line (2) is used for the output activation function. Implementation with other activation functions were also adopted, but this choice gave a better promising results. The nonlinear hyperbolic tangent functions introduced a nonlinear transformation into the network. The hidden layers serve as a feature detector i.e. during the training; they learn the salient attributes (features) that characterizes the training data. The ANN is trained using compact differential evolutionary algorithms variants after which the weights are further fine tuned using backpropagation algorithm (BPA). The training objective function is the minimization of the mean square error (MSE) i.e. the synaptic weights and biases were updated every epoch to minimize the MSE. A supervised batch training method was used with 60% of the data used for training the ANN, 20% for validation and 20% for testing the trained ANN. In this study, the back propagation algorithm is used as a local searcher, thus the learning rate was kept low at 0.01. The inputs of the ANN consist of seven past recent RF power, and time domain data of varying rates of change i.e. second, minute, hour, week day (1 to 7), date day (1 to at most 31), week in a month (1 to 5), and month while the output gives the power in Decibels (dB). Each input of the time domain, enables the ANN to keep track with the trend of RF power variation as a function of that particular input. The current RF power is modelled as a nonlinear function of recent past RF power samples and current time, thus forming a nonlinear time series model. The number of past samples to be used (in this study 7) for reliable prediction and efficient memory management was obtained experimentally as detailed in Section VIII. The actual past RF power (not the predicted RF power) samples fed at the input of the ANN, coupled with the long time training information captured via the time domain inputs, results in a robust ANN model that adapt well to the present trend of RF power variation. In this paper we designed three ANN models. The first model is shown in Fig 1; it consists of only one output neuron and is dedicated for RF power prediction of only one channel which implies that each channel will have its own dedicated ANN. To circumvent this problem, we designed two models for RF power prediction in multiple channels. The second model depicted in Fig 2 is used for prediction of RF power in many channels (for this study is 20 channels) but one at a time. It has only one output neuron, but in addition to the time and past RF power samples inputs, it has another inputs representing the channels. The output neurons of the third (parallel) model is equal to the number of channels to be considered Fig 3. The parallel model is used for simultaneous prediction of RF power in multiple channels given the current time instant and past RF power samples as inputs. For the parallel model, if 7 recent past samples of each of the channels were used as distinct feedback inputs, there will be a total of $7N$ feedback inputs; where $N$ is the number of channels Fig 3; and the training will be computationally expensive. These large feedback inputs ware reduced to 7 by using their average.The data used in this study were obtained by capturing real world RF signals within the GSM 900, VHF and UHV TV and FM bands for a period of two months. In all the models, no RF power related parameters such as signal to noise ratio (SNR), bandwidth, and modulation type, are used as the input of the ANN. Thus making the models robust for cognitive radio application where the CR has no prior knowledge of these RF power related parameters.

Artificial neural network architecture can be broadly classified as either feed forward or recurrent type [22]. Each of these two classes can be structured in different configurations. A feed forward network is one in which the output of one layer is connected to input of the next layer via a synaptic weight, while the recurrent type may have at least one feedback connection or connections between neurons within the same layer or other layers depending on the topology (architecture). The training time of the feed forward is less compared to that of the recurrent type but the recurrent type has better memory capability for recalling past events. Four ANN topologies were considered: feed forward (FF), cascaded feed forward (CFF), feed forward with output feedback (FFB), and layered recurrent (LR) ANN.

The accuracy and level of generalization of ANN depend largely on the initial weights and biases, learning rate, momentum constant,training data and also the network topology. In this paper, the learning rate and the momentum were kept constant at 0.01 and 0.008 respectively while the initial weights and biases were evolved using compact differential evolutionary algorithm variants. The first generation initial weights and biases were randomly generated and constrained within the decision space of -2 to 2. After 1000 generations, the ANN weights and biases were initialized using the elite i.e. the most fittest solution (candidate with the least MSE, obtain using test data) and then train further using backpropagation algorithm (BPA) to fine tune the weights as detailed in the training Section VI-A. Thus producing the final optimized ANN model.

$$F(x) = b \cdot tanh(ax) = b(\frac{e^{ax} - e^{-ax}}{e^{ax} + e^{-ax}}) \tag{1}$$

$$F(x) = mx + c \tag{2}$$

Where the intercept $c = 0$ and the gradient $m$ is left at Matlab default while the constants $a$ and $b$ are assigned the value 1.

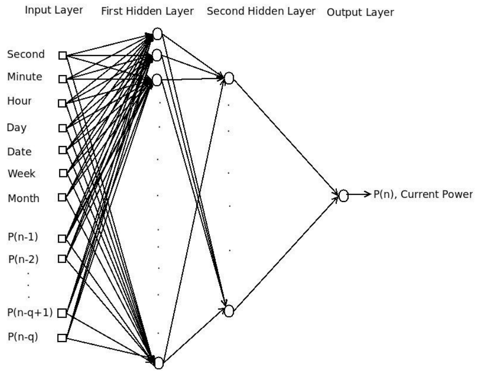


Fig. 1: Dedicated ANN model for one channel

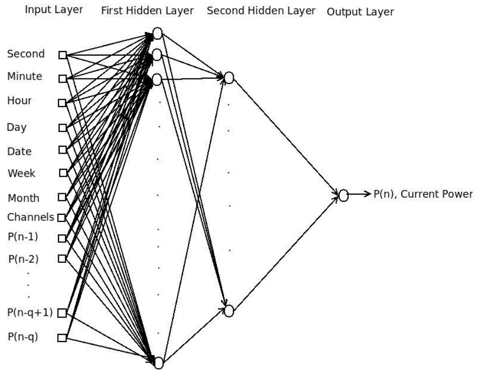


Fig. 2: Multiple channels, single output ANN model

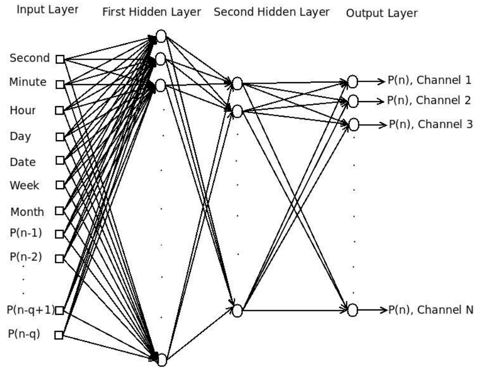


Fig. 3: Multiple channels, parallel outputs ANN model

Where $n$ is a time index, $P(n-1), P(n-2), \cdots, P(n-q)$ are the past $q$ RF power samples while $P(n)$ is the current predicted RF power.

## IV. Support Vector Machine

Support vector machine (SVM) used for regression is often known as support vector regression (SVR). In SVR, the input space $x$ is first mapped onto a high $m$ dimensional feature space by means of certain non-linear transformation (mapping), after which a linear model $f(x, w)$ is constructed in the feature space as shown in (4), [22]. Many time series regression prediction models uses certain lost functions during the training phase for minimization of the empirical risk, among these loss functions are mean square error, square error and absolute error. In SVM regression, a different loss function called $\varepsilon$-insensitive loss proposed in [24] [25], is used. When the error is within the threshold $\varepsilon$, it is considered as zero, beyond the threshold $\varepsilon$, the loss function (error) is computed as the difference between the actual error and the threshold as depicted in (5). The empirical risk function of support vector regression is as shown in (6). The gaol of SVR model is to approximate an unknown real-value function depicted by (3). Where $x$ is a multivariate input vector while $y$ is a scalar output, and $\delta$ is independent and identically distributed (i.i.d.) zero mean random noise or error. The model is estimated using a finite training samples $(x_i, y_i)$ for $i = 1, \cdots, n$ where $n$ is the number of training samples. For this study, the input vector $x$ of the SVR model consist of past recent RF power, current time and frequency while the scalar output $y$ is the current power in Decibels (dB).

$$y = r(x) + \delta \tag{3}$$

$$f(x, w) = \sum_{j=1}^{m} w_j g_j(x) + b \tag{4}$$

Where $g_j(x),\ j = 1, \cdots, m$ refer to set of non-linear transformations, $w_j$ are the weights and $b$ is the bias.

$$L_\varepsilon(y, f(x, w)) = \begin{cases} 0 & \text{if } |y - f(x, w)| \le \varepsilon \\ |y - f(x, w)| - \varepsilon & \text{otherwise} \end{cases} \tag{5}$$

$$R_{emp}(w) = \frac{1}{n} \sum_{i=1}^{n} L_\varepsilon(y_i, f(x_i, w)) \tag{6}$$

Support vector regression model is formulated as the minimization of of the following objective functions, [22]:

$$\text{minimise} \quad \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}(\xi_i + {\xi_i}^*) \tag{7}$$

$$\text{subject to} \quad \begin{cases} y_i - f(x_i, w) - b \le \varepsilon + {\xi_i}^* \\ f(x_i, w) + b - y_i \le \varepsilon + \xi_i \\ \xi_i, {\xi_i}^* \ge 0, \quad i = 1, \cdots, n \end{cases} \tag{8}$$

The non-negative constant $C$ is a regularization parameter that determined the trade off between model complexity (flatness) and the extend to which the deviations larger than $\varepsilon$ will be tolerated in the optimization formulation. It controls the trade-off between achieving a low training and validation error, and minimizing the norm of the weights. Thus the model generalization is partly dependent on C. The parameter C enforces an upper bound on the norm of the weights, as shown in (9). Very small value of C will lead to large training error while infinite or very large value of C will lead to over-fitting resulting from large number of support vectors, [26]. The slack variables $\xi_i$ and ${\xi_i}^*$ represent the upper and lower constrains on the output of the system. These slack variables are introduced to estimate the deviation of the training samples from the $\varepsilon$-sensitive zone thus reducing model complexity by minimizing the norms of the weights, and at the same time performing linear regression in the high dimensional feature space using $\varepsilon$-sensitive loss function. The parameter $\varepsilon$ controls the width of the $\varepsilon$-insensitive zone used to fit the training data. The number of support vectors used in constructing the support vector regression model (function) is partly dependent on the parameter $\varepsilon$. If $\varepsilon$-value is very large, few support vectors will be selected, on the contrary, bigger $\varepsilon$-value results in a more generalized model (flat estimate). Thus both the complexity and the generalization capability of the network depend on its value. One other parameter that can affect the generalization and accuracy of a support vector regression model is the kernel parameter and the type of kernel function used as shown in (11) to (14).

There are three meta-parameters or hyperparameters that determine the complexity, generalization capability and accuracy of support vector machine regression model, these are the $C$ Parameter, $\varepsilon$ and the kernel parameter $\gamma$, [27], [28], [29]. Optimal selection of these parameters is further complicated due to the fact that they are problem dependent and the performance of the SVR model depends on all the three parameters. There are many proposals how these parameters can be chosen. It has been suggested that these parameters should be selected by users based on the users experience, expertise and a priori knowledge of the problem, ( [24], [25], [30], [31]). This leads to many repeated trial and error attempts before getting the optimums if possible, and it limit the usage to only experts. In this study, we used cDE to evolves the three meta parameters of the SVR model. SVR optimization problem constitute a dual problem with a solution given by

$$f(x) = \sum_{i=1}^{s} (\alpha_i - \alpha_i^*) K(x, x_i) + b \tag{9}$$

The dual coefficients in (9) are subject to the constrains $0 \le \alpha_i \le C$ and $0 \le \alpha_i^* \le C$. Where $s$ is the number of support vectors, $K(x, x_i)$ is the kernel function, $b$ is the bias, while $\alpha$ and $\alpha^*$ are Lagrange multipliers. The training samples $x$ with non-zero coefficients in (9) are called the support vectors. The general expression for the kernel is depicted in (10). Any symmetric positive definite function, which satisfies Mercers Conditions [22] can be used as kernel function. In this study, four kernel were used, i.e. the Radial Basis Function (RBF), Gaussian Radial Basis Function, Exponential Radial Basis Function kernel and Linear kernel given by (11), (12), (13), and (14) respectively, [22]. In this study, we designed two SVR models for each kernel, one of the model shown in Fig. 4 is dedicated for prediction of RF power of only one channel or resource block which implies that each channel will have it own model; the second model shown in Fig. 5 has an additional channel input thus it can be used for prediction of RF power in many channels but one at a time.

$$K(x, x_i) = \sum_{j=1}^{m} g_j(x) g_j(x_i) \tag{10}$$

$$K(x, x_i) = e^{(-\gamma\|x - x_i\|^2)} \tag{11}$$

$$K(x, x_i) = e^{(-\frac{\|x - x_i\|^2}{2\sigma^2})} \tag{12}$$

$$K(x, x_i) = e^{(-\frac{\|x - x_i\|}{2\sigma^2})} \tag{13}$$

$$K(x, x_i) = x^T x_i + c \tag{14}$$

The adjustable constant parameter $c$ of the linear kernel was set to 0 while the kernel parameters $\sigma$ and $\gamma$ were evolved using cDE algorithm variants.

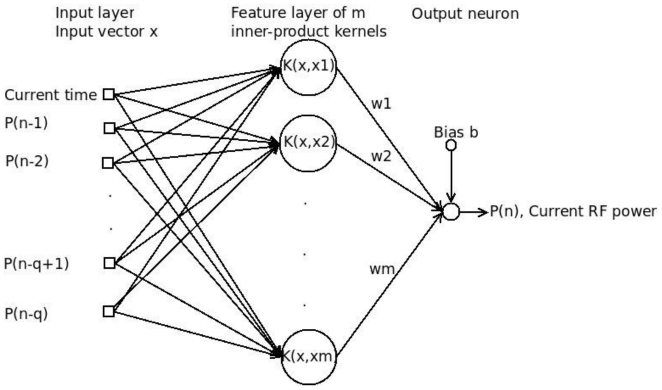


Fig. 4: Dedicated SVR model for one channel

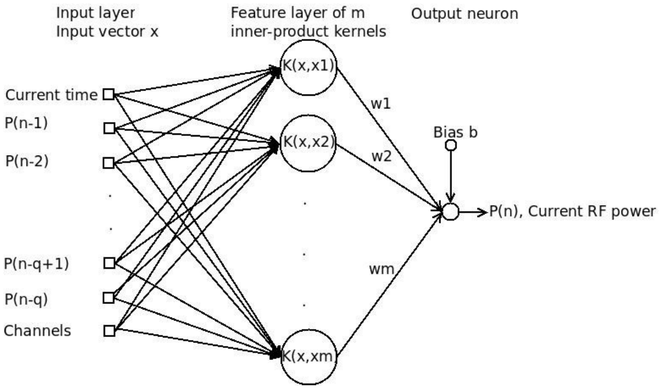


Fig. 5: Multiple channels, single output SVR model

Where $n$ is a time index, $P(n-1), P(n-2), \cdots, P(n-q)$ are the past $q$ RF power samples while $P(n)$ is the current predicted RF power.

## V. Optimization Algorithms

A brief description of the optimization algorithms implemented are presented in this section. We combine the global search capability of compact differential evolutionary algorithms with the single solution local search advantages of BPA to evolve the weights and biases of the optimized ANN model as described in the training, Section VI-A.

### A. Differential Evolution

Base on the original definition, DE are population based direct search algorithms used to solve continuous optimization problems [32] [33]. DE aims at evolving $NP$ population of $D$ dimensional vectors which encodes the $G$ generation candidate solutions $X_{i,G} = \{X_{i,G}^1, \cdots X_{i,G}^D\}$ towards the global optimum, where $i = 1, \cdots, NP$. The initial candidate solutions at $G = 0$ are evolves in such a way as to cover the decision space as much as possible by uniformly randomizing the candidates within the search domain using (15), [32].

$$X_{i,G} = X_{min} + rand(1,0) \cdot (X_{max} - X_{min}) \tag{15}$$

Where $i = 1, \cdots NP$. $X_{min} = \{X_{min}^1 \cdots X_{min}^D\}$, $X_{max} = \{X_{max}^1 \cdots X_{max}^D\}$ and $rand(1,0)$ is a uniformly distributed random number between 0 and 1.

### B. Mutation

For every candidates solution (individuals or target vectors) $X_{i,G}$ at generation $G$, a mutant vector $V_{i,G}$ called the provisional or trial offspring is generated via certain mutation schemes. The mutation strategies implemented in this study are as shown in (16) to (20), [32]:

- DE/rand/1:
$$V_{i,G} = X_{r_1,G} + F \cdot (X_{r_2,G} - X_{r_3,G}) \tag{16}$$
- DE/best/1:
$$V_{i,G} = X_{best,G} + F \cdot (X_{r_1,G} - X_{r_2,G}) \tag{17}$$
- DE/rand-to-best/1:
$$V_{i,G} = X_{i,G} + F \cdot (X_{best,G} - X_{i,G}) + F \cdot (X_{r_1,G} - X_{r_2,G}) \tag{18}$$
- DE/best/2
$$V_{i,G} = X_{best,G} + F \cdot (X_{r_1,G} - X_{r_2,G}) + F \cdot (X_{r_3,G} - X_{r_4,G}) \tag{19}$$
- DE/rand/2
$$V_{i,G} = X_{r_5,G} + F \cdot (X_{r_1,G} - X_{r_2,G}) + F \cdot (X_{r_3,G} - X_{r_4,G}) \tag{20}$$

Where the indexes $r_1$, $r_2$, $r_3$, $r_4$ and $r_5$are mutually exclusive positive integers and distinct from $i$. These indexes are generated at random within the range $[1\ PN]$. $X_{best,G}$ is the individual with the best fitness at generation $G$ while $F$ is the mutation constant.

### C. Cross Over

After the mutants are generated, the offspring $U_{i,G}$ are produced by performing a crossover operation between the target vector $X_{i,G}$ and its corresponding provisional offspring $V_{i,G}$. The two crossover schemes i.e. exponential and binomial crossover are used in this study for all the cDE algorithm variants implemented [34]. The binomial crossover copied the $jth$ gene of the mutant vector $V_{i,G}$ to the corresponding gene (element) in the offspring $U_{i,G}$ if $rand(0,1) \leq CR$ or $j = j_{rand}$. Otherwise it is copied from the target vector $X_{i,G}$ (parent). The crossover rate $Cr$ is the probability of selecting the offspring genes from the mutant while $j_{rand}$ is a random number in the range $[1\ D]$, this ensure that at least one of the offspring gene is copied from the mutant. The binomial crossover is represented by (21), [32]:

$$U_{i,G}^j = \begin{cases} V_{i,G}^j & \text{if} \quad (rand(0,1) \leq Cr \quad \text{or} \quad j = j_{rand}) \\ X_{i,G}^j & \text{otherwise} \end{cases} \tag{21}$$

For exponential crossover, the genes of the offspring are inherited from the mutant vector $V_{i,G}$ starting from a randomly selected index $j$ in the range $[1\ D]$ until the first time $rand(0,1) > Cr$ after which all the other genes are inherited from the parent $X_{i,G}$. The exponential crossover is as shown in Algorithm 1, [32].

*Algorithm 1:* Exponential Crossover

```
$U_{i,G} = X_{i,G}$
generate $j = randi(1, D)$
$U^j_{i,G} = V^j_{i,G}$
$k = 1$
**while** $rand\,(0, 1) \leq Cr$ AND $k < D$ **do**
    $U^j_{i,G} = V^j_{i,G}$
    $j = j + 1$
    **if** $j == n$ **then**
        $j = 1$
    **end if**
    $k = k + 1$
**end while**
```

**end**

### D. Selection Process

After every generation, the fitness function of each offspring $U_{i,G}$ and the corresponding parent $X_{i,G}$ are computed. A greedy selection schemes is used in which if the fitness function of the offspring is less than or equal to that of it parent, the offspring will replace the corresponding parent in the next generation otherwise the parent will be maintained among the next generation individuals. At the end of the generation, the most fittest individual (global best) among the final evolved solutions is selected. The DE algorithm pseudocode is depicted in Algorithm 2.

*Algorithm 2:* Differential Evolution

```
Generate an initial population $X_{G=0}$ of $Np$ individuals.
Evaluate fitness of each individuals (solutions).
**while** termination condition is not met (Generation) **do**
    **for** $i = 1$ to $Np$ **do**
        Evaluate parent ($X_{i,G}$) fitness .
        Generate trial offspring $V_{i,G}$ by mutation using (16).
        Generate offspring $U_{i,G}$ by either binomial crossover or exponential crossover.
        Evaluate offspring ($U_{i,G}$) fitness
    **end for**
    **for** $i = 1$ to $Np$ **do**
        Selection Process:
        Form the next generation solutions by selecting the best between parents and their offspring
    **end for**
**end while**
```

**end**

## VI. Compact Differential Evolution

Compact differential evolution (cDE) algorithm is achieved by incorporating the update logic of real values compact genetic algorithm (rcGA) within DE frame work [35] [36] [37]. The steps involves in cDE is as follows: A (2 x n) probability vector **PV** consisting of the mean $\mu$ and standard deviation $\sigma$ is generated. where $n$ is the dimensionality of the problem (in this case the number of weights and biases). At initialization, $\mu$ was set to 0 while $\sigma$ was set to a very large value 10, in order to simulate a uniform distribution. A solution called the elite is sampled from the **PV**. At each generation (step) other candidate solutions are sampled from the **PV** according to the mutation schemes adopted as described in Section V-B, e.g. for DE/rand/1 three candidate solutions $X_{r_1}$, $X_{r_2}$ and $X_{r_3}$ are sampled. Without lost of generality, each designed variable $X_{r_1}[i]$ belonging to a candidate solution $X_{r_1}$, is obtained from the **PV** as follows: For each dimension indexed by $i$, a truncated Gaussian probability density function (PDF) with mean $\mu[i]$ and standard deviation $\sigma[i]$ is assigned. The truncated PDF is defined by (22). The CDF of the truncated PDF is obtained. A random number rand(0,1) is sampled from a uniform distribution. $X_{r_1}[i]$ is obtained by applying the random number rand(0,1) generated to the inverse function of the CDF. Since both the PDF and CDF are truncated or normalized within the range [-1, 1]; the actual value of $X_{r_1}[i]$ within the true decision space of [a, b] is obtain as $(X_{r_1}[i] + 1)\frac{(b-a)}{2} + a$. The mutant (provisional offspring) is now generated using the mutation schemes. The offspring is evolved by performing a crossover operation between the elite and the provisional offspring as described in Section V-C. The fitness value of the offspring is computed and compare with that of the elite. If the offspring outperform the elite, it replaces the elite and declare the winner while the elite the loser; otherwise the elite is maintained and declare the winner while the offspring the loser. In this study, the fitness function is the MSE obtain using the test data. The weights and the biases of the ANN are initialized with the offspring and the MSE is obtain, this is repeated using the elite. The one with the least MSE is the winner. The **PV** is updated using (23) and (24). Hence in cDE, instead of having a population of individuals (candidates solutions) for every generation as in normal DE, the population are represented by their probability distribution function (i.e. their statistics), thus minimizing the computational complexity, amount of memory needed, and the optimization time. The psuedocode of cDE is as shown in Algorithm 3, [35].

$$PDF(\mu[i], \sigma[i]) = \frac{e^{\frac{-(x-\mu[i])^2}{2\sigma[i]^2}}\sqrt{\frac{2}{\pi}}}{\sigma[i](erf(\frac{\mu[i]+1}{\sqrt{2}\sigma[i]}) - erf(\frac{\mu[i]-1}{\sqrt{2}\sigma[i]}))} \tag{22}$$

$$\mu^{t+1}[i] = \mu^t[i] + \frac{1}{N_P}(Winner[i] - loser[i]) \tag{23}$$

$$\sigma^{t+1}[i] = \sqrt{(\sigma^t[i])^2 + \delta[i]^2 + \frac{1}{N_P}(Winner[i]^2 - loser[i]^2)} \tag{24}$$

where $\delta[i]^2 = (\mu^t[i])^2 - (\mu^{t+1}[i])^2$ , $t =$ steps or generations, $N_P$ is a vitual population and $erf$ is the error function.

*Algorithm 3:* Compact Differential Evolution Pseudocode

```
generation t=0
** PV Initialization **
**for** $i = 1$ to $n$ **do**
    Initialize $\mu[i] = 0$
    Initialize $\sigma[i] = 10$
**end for**
Generate the elite by means of **PV**
**while** buget condition **do**
    ** Mutation **
```

```
        Generate 3 or more individuals according to the mutation schemes e.g. $X_{r_1}$, $X_{r_2}$ and $X_{r_3}$ by means of **PV**
        Compute the mutant $V = X_{r_1} + F \cdot (X_{r_2} - X_{r_3})$
        ** Crossover **
        $U = V$, where $U =$ offspring
        for $i = 1 : N$ do
            Generate rand(0,1)
            if $rand(0, 1) > Cr$ then
                $U[i] = elite[i]$
            end if
        end for
        ** Elite Selcetion **
        [ Winner Loser] $= compete(U, elite)$
        if $U ==$ Winner then
            $elite = U$
        end if
        ** PV Update **
        for $i = 1 : n$ do
            $\mu^{t+1}[i] = \mu^t[i] + \frac{1}{N_P}(Winner[i] - loser[i])$
            $\sigma^{t+1}[i] = \sqrt{(\sigma^t[i])^2 + \delta[i]^2 + \gamma[i]^2}$
            Where: $\delta[i]^2 = (\mu^t[i])^2 - (\mu^{t+1}[i])^2$
            $\gamma[i]^2 = \frac{1}{N_P}(Winner[i]^2 - loser[i]^2)$
        end for
        $t = t + 1$
    end while
end
```

### A. Training of ANN and SVM

The objective function in this study is the MSE of the optimized ANN computed using the test data. After every generation, the offspring $U_G$ and the $elite$ are used to set the weights and biases of the ANN and the MSE of the ANN models are obtain using the test data. The use of the test data (data not known by the ANN nor used to train it) for computation of the fitness function (MSE) does not only result in a more accurate network but also a more robust and generalized ANN model. A greedy selection schemes is used in which if the MSE of the offspring is less than or equal to that of the elite, the offspring will replace the elite in the next generation otherwise the elite will be maintained. At the end of the generations, the most fittest candidate solution i.e. the final evolved elite; is used to initialize the weights and biases of the ANN which is further trained using back propagation algorithms (BPA) to fine tune the weights to produce the final optimized ANN model. The cDE is run for 1000 generations. The fine tuning of the ANN weights using BPA was constrained within a maximum of 200 epoch and 6 validation fails, i.e the training stop if any of these constrain thresholds is satisfied. One of the desirable feature of BPA is it simplicity but it often converges slowly and lack optimality as it can easily be trapped in a local optimum leading to premature convergent. Many approaches has been adopted to solve the problem of premature convergent associated with BPA such as the introduction of momentum constant, varying of the learning rate and retraining of the network with new initial weights. To circumvent the problem of premature convergent, and to have a robust ANN that is well generalized, we combine the global search advantages of cDE optimization algorithm and the local search capability of single solution BPA to evolve the weights and biases of the ANN. The combination of the global search capabilities of cDE and the local search advantages of BPA to evolve the weight and biases of ANN have proving to be superior to using only the famous BPA for this problem. The cDE algorithm pseudocode is depicted in Algorithm 3.

In constract to the training of ANN using BPA, the training of SVM is optimal with the optimality rooted in convex optimization. This desired feature of SVM is obtained at the cost of increased computational complexity. The fact that the training of SVM is optimal does not implies that the evolved machine will be well generalized or have a good performance. The optimality here is based on the chosen meta parameters ( i.e. $C$ parameter, $\varepsilon$ and the kernel parameter $\gamma$), the type of kernel function used and the training data. We used the same randomization cDE optimization algorithm variants to evolve the SVM meta parameters while the weights and bias of the SVM were evolves via convex optimization. At each generation, the meta parameters are set using each candidate solution, and the corresponding weights and bias are computed. In order to estimate how the SVM will generalize to an independent dataset (test data), we use two fold cross validation commonly known as holdout method. This has the advantage of having both large training and validation datasets, and each data point is used for both training and validation on each fold. The training data is randomly divided into two sets e.g. $A$ and $B$ of equal size. The SVM was trained on $A$ and test on $B$, after which it is trained on $B$ and test on $A$, the average of the MSE for the two test was used as the fitness function for the given sets of meta parameters. At the end of the generations, the SVM is reconstructed using the most fittest meta parameters and tested on the test datasets (data not known by the SVM nor used to train it).

## VII. Experiment and Simulation Data

The datasets used in this study were obtained by capturing real world RF signals using universal software radio peripheral 1 (USRP 1) for a period of two months. The USRP are computer hosted software-defined radios with one motherboard and interchangeable daughter board modules for various ranges of frequencies. The daughter board modules serve as the RF front end. Two daughter boards, SBX and Tuner 4937 DI5 3X7901, having a continuous frequency ranges of 4MHz to 4.4GHz and 50 MHz to 860 MHz respectively, were used in this research. The daughterboard perform analog operations such as up/down-conversion, filtering, and other signal conditioning while the motherboard perform the functions of clock generation and synchronization, analog to digital conversion (ADC),digital to analog conversion (DAC), host processor interface, and power control. It also decimate the signal to a lower sampling rate that can easily be transmitted to the host computer through a high-speed USB cable where the signal is processed by software. For TV channels with channels bandwidth of 8 MHz, we divided the channels into subchannels (resource block) each consisting of 500 KHz bandwidth. To ensure that no spectral information was lost, we used a sample frequency of 1MHz and obtained 1024 samples for each sample time. For GSM 900 and FM band with a bandwidth of 200 KHz, we used 1MHz sample frequency and 512 samples for each sample time. The power was obtained using both the time and frequency domain data. For the frequency domain, after passing the signal through the channel filter, the signal was windowed using a hamming window in order to reduce spectral leakage. The stream of the data was converted to a vector and decimated to a lower sampling rate that can easily be processed by the host computer at run time using the inbuilt decimation block in gnu-radio. This is then converted to the frequency domain and the magnitudes of the bins were passed to a probe sink. The choice of probe sink is essential because it can only hold the current data and does not increase thereby preventing stack overflow or a segmentation fault. This allows Python to grab the data at run time for further analysis. The interval of time between consecutive sample data was selected at a random value between 5 seconds and 30 seconds. The choice of this range is based on the assumption that for any TV programme, FM broadcast or GSM calls, will last for not less than 5 to 30 seconds. In order to capture all possible trends, the time between consecutive sample data is selected at random within the given range instead of using regular intervals. For the VHF and FM band we captured RF signals from 54MHz to 110MHz and 470 to 670MHz for the VHF TV bands. For the GSM band, 62 down-link channels (935MHz to 953.3MHz) and 62 uplink channels (890MHz to 908.3MHz) were captured. The real world RF data was divided into three subsets, randomly selected with 60% used for training the ANN, 20% for validation and 20% for testing the trained ANN model. The training or estimation data were the only known data sources used in training the ANN. The test data set was unknown to the network i.e. they are not used in training the network rather are used in testing the trained ANN as a measure of the generalization performance of the ANN model. The ANN design, optimization and the simulation were done in Matlab while the capturing of the data and the signal processing were implemented using gnu-radio which is a combination of Python and C++.

## VIII. Delayed Inputs Sensitivity Analysis

In order to examine how many numbers of recent past RF power samples are needed as feedback inputs for reliable prediction, and to have a model with reduced dimensionality of input vector, we carried out a sensitivity analysis. One way of evaluating the importance (significance) of an input in ANN is to measure the Change Of MSE (COM) when the input is deleted or added to the neural network [38]. In this study, the COM method is adopted with the time domain inputs unaltered, and the actual past RF power are added to the input one after the other starting from the most recent one. The ANN is trained with $i$ delay inputs (past RF power samples) and the MSE $MSE_i$ is evaluated. The network is retrained with $i+1$ delayed inputs, the MSE $MSE_{i+1}$ is obtained and the change in the MSE, $\delta_{mse} = MSE_{i+1} - MSE_i$ is computed as a means of evaluating the importance of the $i+1$ delay input, for $i = [0 \cdots q]$, where $q$ is the total number of past samples used; see Fig 1. Note, $\delta_{mse}$ is not computed relative to the MSE obtained when all the $q$ delayed inputs are used as in normal COM method, due to the fact that we don't know the required number of delay inputs $q$ at the start of the experiment; in this case $q$ is obtained by setting a constrain on $\delta_{mse}$. The importance of the inputs are ranked base on the one whose addition causes the largest decrease in MSE as the most important since they are most relevant to the construction of a network with the smallest MSE. In order to justify the importance or ranking of the inputs statistically, for every $i$ inputs delay, the ANN is trained 20 times, each time with a randomly generated initial weights and biases, the average of the 20 $MSE_i$ is used. The ranking using the normalized values of change in average MSE $\delta_{mse}$ as delayed inputs were added is as shown in Fig 7. The graph of the average MSE against number of delayed inputs is as depicted in Fig 6. From Fig 6 and Fig 7, it is obvious that when the number of delay inputs is $> 7$, the change in MSE $\delta_{mse}$, is very small. Thus in this study we decided to use 7 past recent RF power samples as part of the ANN inputs for current RF power prediction, taken into cognition the memory constrain of CR as an embedded device.

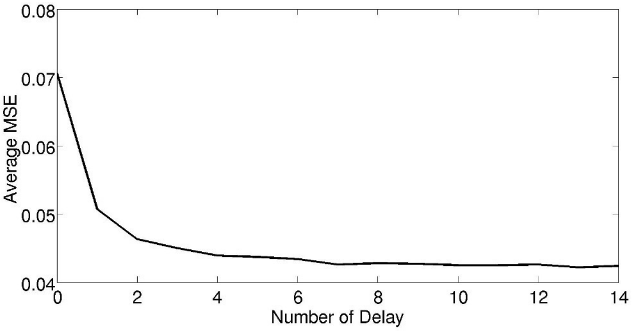


Fig. 6: Sensitive analysis curve

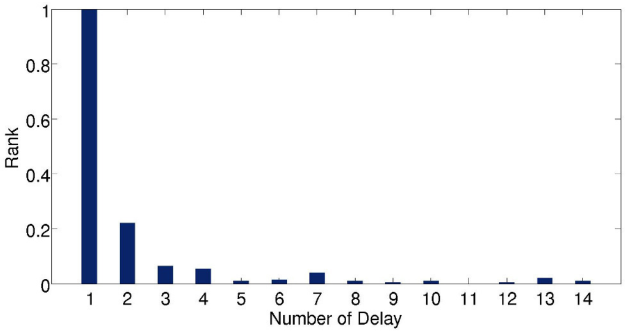


Fig. 7: Past RF power sensitivity ranking

## IX. Results

To minimize the MSE of the ANN when tested with the test data, the above listed algorithms were run for 30 independent runs. Each run has been continued with 30000 fitness evaluations for 1000 generations. After a manual tuning of the parameters, the following parameters are used in this study:

- cDE/rand/1/bin, cDE/rand/1/exp, cDE/rand/2/bin, cDE/rand/2/exp, cDE/best/1/bin, cDE/best/1/exp, cDE/best/2/bin, cDE/best/2/exp, cDE/rand-to-best/1/bin and cDE/rand-to-best/1/exp has been run with $F = 0.1$ and $Cr = 0.3$
- BPA has been run with $Epoch = 1200$, $learningrate = 0.01$ $mumentum = 0.008$ the other specifications are shown in Table I.

TABLE I: ANN Models Specification

| | ANN Models | | |
|---|---|---|---|
| | Dedicated one channel | Multiple channels, single output | Multiple channels, parallel output |
| First Hidden Neurons | 5 | 15 | 15 |
| Second Hidden Neurons | 3 | 10 | 10 |
| Output Neurons | 1 | 1 | 20 |
| Number of Channels | 1 | 20 | 20 |

Tables IV and III shows the numerical results in terms of the MSE obtained using the test data (data not known by the ANN nor used in training the ANN). The final results of each algorithm was obtained by taken the average of the MSE (AMSE) for the 30 independent runs and their corresponding standard deviation (STD). From the results, the combination of cDE/rand/1/exp and back propagation algorithm (BPA) outperform all the other algorithms with reference to the FFB ANN model while the combination of cDE/rand/1/bin and BPA is the best for the FF ANN model. These two bests are used as the reference for the Wilcoxon test [39]. A '+' indicate that the reference algorithm outperform the other algorithm while "−" mean that the other algorithm outperform the reference. For this problem, when the Wilcoxon test was perform by changing the reference algorithm, the second best algorithm for the FFB ANN model are cDE/rand-to-best/1/exp and cDE/rand/2/exp; both having the same AMSE and STD of 0.0290 and 0.0005 respectively while for FF ANN model, the second best algorithm is cDE/rand/1/exp. For this problem, the FF ANN model trained using cDE/rand/1/bin and BPA emerge as the best compared with other models (FFB, CFF, LR) and algorithms implemented. This implies that the feedback information may have been captured through the inputs assigned to the 7 recent past RF power samples. Comparing the results depicted in tables IV and III with our previous work detailed in [40]; the use of some of the most recent RF power samples as part of the input vector of the ANN, produces a more accurate and robust ANN model with reduced number of neurons. This form a non-linear time series predictive model. The neurons in the model adopted in this study is approximately half of the ones used in [40], thus it has less parameters (weights and biases) to be optimized. To validate the fact that the combination of these cDE variants of optimization algorithms with the famous BPA to evolves the weights and biases of ANN will produce a more accurate, robust and generalized model than using only BPA; we use the same topology but train with only BPA at constant learning rate of 0.01 and another one with varying learning rate starting from 0.8 and keep on changing with change in MSE using inbuilt Matlab training function traingda. For both models trained with only BPA, each was run 30 times, each run was constrain within a maximum of 1200 epoch and 6 validation fails, the average results is depicted in tables IVand III. For the hybridized training i.e combining cDE with BPA, the cDE is run for 1000 generations and the final best solution (elite) was used to reinitialized the ANN weights and further train using BPA constrained within 200 epoch and 6 validation fails. In almost all cases, the hybridize training outperform the training with only BPA. Fig 8, 9 and 10 shows the prediction graphs of some selected channels using test data. These results depict a good generalization of the three models. For this problem, the combination of the global search capabilities of cDE algorithm variants, and the local search advantages of BPA to evolve the weights of the ANN was found to yield an improved performance as compared to using only the famous BPA.

The prediction results of the dedicated SVR model shown in Fig. 4 is depicted in Table IV. From this result, the exponential kernel with meta parameters evolved using cDE/rand/2/exp seem to be more promising with an average MSE of 0.0226, the next best kernel is the linear kernel with AMSE of 0.0301 using cDE/rand/1/bin variant. For multiple channel, single output SVR model Fig. 5, trained for prediction of RF power of 20 resource blocks or channels, the linear kernel emerge as the best with AMSE of 0.0682 this is followed by the RBF kernel with AMSE of 0.0819. the best hyperparameters are evolved using cDE/rand/1/exp for linear kernel and cDE/rand/1/bin and cDE/rand/2/bin for RBF. The results for FF ANN, multiple channel, single output model is also shown in Table with best model having an AMSE of 0.0818 with weights and biases evolved using cDE/best/2/exp and BPA.

TABLE II: Test Results Using FFB ANN Model With DE/best/1/bin as Reference

| Algorithms | Algorithms | | Algorithms + BPA | |
|---|---|---|---|---|
| | AMSE | STD | AMSE | STD |
| cDE/rand/1/bin | 0.4055 | 0.0780 | 0.0403 | (0.0879+) |
| **cDE/rand/1/exp** | **0.1950** | **0.0964** | **0.0242** | **(0.0007)** |
| cDE/best/2/bin | 0.1496 | 0.0354 | 0.1999 | (0.2405+) |
| cDE/best/2/exp | 0.3243 | 0.0400 | 0.1635 | (0.2272+) |
| cDE/best/1/bin | 0.1436 | 0.0335 | 0.1644 | (0.2267+) |
| cDE/best/1/exp | 0.3376 | 0.0419 | 0.1808 | (0.2352+) |
| cDE/rand-to-best/1/bin | 0.1575 | 0.0212 | 0.0359 | (0.0275+) |
| cDE/rand-to-best/1/exp | 0.5233 | 0.1233 | 0.0290 | (0.0005+) |
| cDE/rand/2/bin | 0.0928 | 0.0119 | 0.0312 | (0.0112+) |
| cDE/rand/2/exp | 0.2087 | 0.0302 | 0.0290 | (0.0005+) |
| BPA (constant learning rate) | | | 0.1345 | (0.2029+) |
| BPA (varying learning rate) | | | 0.2508 | (0.0332+) |

TABLE III: Test Results Using FF ANN With DE/best/1/bin as Reference

| Algorithms | Algorithms | | Algorithms + BPA | |
|---|---|---|---|---|
| | AMSE | STD | AMSE | STD |
| **cDE/rand/1/bin** | **0.3294** | **0.1664** | **0.0203** | **(0.0004)** |
| cDE/rand/1/exp | 0.1720 | 0.1490 | 0.0204 | (0.0006+) |
| cDE/best/2/bin | 0.1181 | 0.0447 | 0.0240 | (0.0134+) |
| cDE/best/2/exp | 0.1656 | 0.0833 | 0.0205 | (0.0005+) |
| cDE/best/1/bin | 0.1145 | 0.0314 | 0.0206 | (0.0015+) |
| cDE/best/1/exp | 0.2121 | 0.1189 | 0.0205 | (0.0005+) |
| cDE/rand-to-best/1/bin | 0.3102 | 0.1136 | 0.0205 | (0.0006+) |
| cDE/rand-to-best/1/exp | 0.1646 | 0.1147 | 0.0205 | (0.0008+) |
| cDE/rand/2/bin | 0.3039 | 0.1448 | 0.0214 | (0.0042+) |
| cDE/rand/2/exp | 0.1627 | 0.1050 | 0.0207 | (0.0009+) |
| BPA (constant learning rate) | | | 0.0267 | (0.0004+) |
| BPA (varying learning rate) | | | 0.0476 | (0.0180+) |

## X. Conclusion

This paper demonstrates the power of ANN to produce a robust time series prediction models for RF power traffics in some selected channels. The combination of the global search capabilities of memory efficient cDE and the local search advantages of single solution BPA to evolves the weights and biases of the ANN prediction models, proved to produce a more robust, accurate and well generalise ANN models than

TABLE IV: SVR results using one channel dedicated model

| Algorithms | RBF | | Gaussian RBF | | Exponential | | Linear | |
|---|---|---|---|---|---|---|---|---|
| | AMSE | STD | AMSE | STD | AMSE | STD | AMSE | STD |
| **cDE/rand/1/bin** | 0.4055 | 0.0780 | **0.0614** | **0.0198** | 0.0243 | 0.0006 | **0.0301** | **0.0005** |
| cDE/rand/1/exp | 0.1950 | 0.0964 | 0.1333 | 0.0916 | 0.0298 | 0.0111 | 0.0313 | 0.0012 |
| cDE/best/2/bin | 0.0508 | 0.0174 | 0.2043 | 0.1437 | 0.0350 | 0.0181 | 0.0303 | 0.0003 |
| **cDE/best/2/exp** | **0.0416** | **0.0078** | 0.2114 | 0.0788 | 0.0437 | 0.0190 | 0.0313 | 0.0005 |
| cDE/best/1/bin | 0.0474 | 0.0015 | 0.2670 | 0.0385 | 0.0266 | 0.0041 | 0.0310 | 0.0007 |
| cDE/best/1/exp | 0.0511 | 0.0035 | 0.2115 | 0.1355 | 0.0323 | 0.0015 | 0.0323 | 0.0015 |
| cDE/rand-to-best/1/bin | 0.0477 | 0.0045 | 0.2062 | 0.0856 | 0.0243 | 0.0001 | 0.0314 | 0.0007 |
| cDE/rand-to-best/1/exp | 0.0507 | 0.0037 | 0.1346 | 0.0806 | 0.0239 | 0.0014 | 0.0306 | 0.0008 |
| cDE/rand/2/bin | 1.6718 | 0.0000 | 0.1937 | 0.0946 | 0.0248 | 0.0008 | 0.0306 | 0.0005 |
| **cDE/rand/2/exp** | 0.2087 | 0.0302 | 0.1912 | 0.0625 | **0.0226** | **0.0009** | 0.0313 | 0.0015 |

TABLE V: SVR and ANN results using one output, multiple channel model for 20 channels

| Algorithms | Linear | | Gaussian RBF | | Exponential RBF | | RBF | | ANN FF | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AMSE | STD | AMSE | STD | AMSE | STD | AMSE | STD | AMSE | STD |
| **cDE/rand/1/bin** | 0.0700 | 0.0033 | 0.1274 | 0.0035 | 0.2422 | 0.0296 | **0.0819** | **0.0093** | 0.0842 | 0.0029 |
| **cDE/rand/1/exp** | **0.0682** | **0.0016** | **0.1238** | **0.0025** | 0.2609 | 0.0651 | 0.0848 | 0.0016 | 0.0835 | 0.0028 |
| cDE/best/2/bin | 0.1104 | 0.0012 | 0.3565 | 0.3051 | 0.1472 | 0.0225 | 0.1056 | 0.0084 | 0.0875 | 0.0033 |
| **cDE/best/2/exp** | 0.1078 | 0.0020 | 0.1609 | 0.0155 | 0.1434 | 0.0136 | 0.0855 | 0.0119 | **0.0818** | **0.0017** |
| cDE/best/1/bin | 0.1112 | 0.0017 | 0.6258 | 0.4953 | 0.1466 | 0.012 | 0.1057 | 0.0087 | 0.0869 | 0.0018 |
| cDE/best/1/exp | 0.1112 | 0.0016 | 0.3395 | 0.2010 | 0.3020 | 0.2732 | 0.0947 | 0.0087 | 0.0832 | 0.0032 |
| cDE/rand/best/1/bin | 0.1096 | 0.0019 | 0.2854 | 0.0425 | 0.1361 | 0.0060 | 0.0910 | 0.0040 | 0.0839 | 0.0028 |
| **cDE/rand/best/1/exp** | 0.1086 | 0.0016 | 0.4062 | 0.0433 | **0.1357** | **0.0035** | 0.0868 | 0.0036 | 0.0851 | 0.0012 |
| **cDE/rand/2/bin** | 0.1094 | 0.0019 | 0.1281 | 0.0095 | 0.2345 | 0.0162 | **0.0819** | **0.0074** | 0.0844 | 0.0035 |
| cDE/rand/2/exp | 0.0687 | 0.0011 | 0.1255 | 0.0036 | 0.2257 | 0.0104 | 0.0902 | 0.0156 | 0.0859 | 0.0042 |

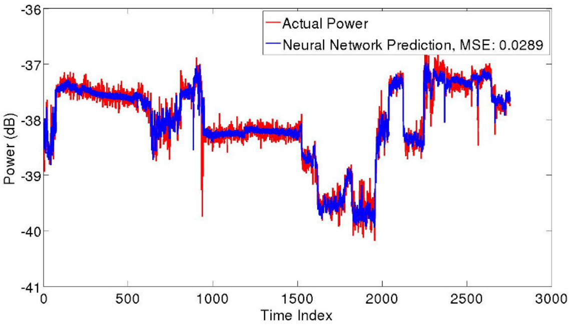


Fig. 9: Cascaded feed forward, parallel output model prediction

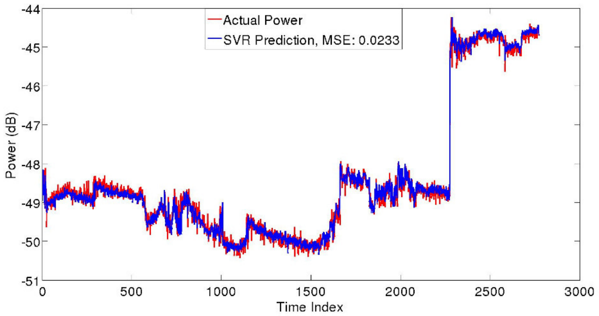


Fig. 10: Multiple channels, single output SVR model prediction

TABLE VI: Best SVR model parameters for Table IV

| | Kernel | | | |
|---|---|---|---|---|
| | RBF | Gaussian RBF | Exponential RBF | Linear |
| MSE | 0.0416 | 0.0614 | 0.0226 | 0.0301 |
| $C$ | 7.72 | 7.62 | 22.28 | 3.76 |
| $\varepsilon$ | 0.000653 | 0.0003331 | 0.000228 | 0.000472 |
| $\gamma$ or $\sigma$ | 7.38 | 1.49 | 7.57 | - |
| Algorithms cDE/ | best/2/exp | rand/1/bin | rand/2/exp | rand/1/bin |

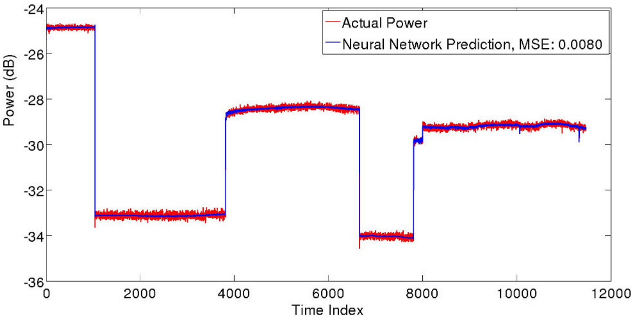


Fig. 8: Feed forward dedicated model prediction

using only BPA for this problem. For the dedicated one channel model, the ANN outperform the SVR model for all the kernels implemented while for the multiple channels, single output model, only the linear kernel SVR model outperform the ANN. The a priori knowledge of the RF power resulting from either communication signals, noise and/or interferences, is not only applicable to cognitive radio network, but in any wireless communication system for noisy channels avoidance.

## Acknowledgment

This work is supported by Petroleum Technology Development Fun (PTDF) Scholarship, Nigeria and De Montfort University, United Kingdom.

## References

[1] FCC, "Federal comminucation commission notice of inquiry and notice of proposed rule making, in the matter of establishment of an interference temperature metric to quantify and manage interference and to expand available unlicensed operation in certain fixed, mobile and satellite frequency bands," no. 03-237, November 13, 2003.

[2] V. Valenta, R. Marsalek, G. Baudoin, M. Villegas, M. Suarez, and F. Robert, "Survey on spectrum utilization in europe: Measurements, analyses and observations," in *Cognitive Radio Oriented Wireless Networks Communications (CROWNCOM), 2010 Proceedings of the Fifth International Conference on*, June 2010, pp. 1–5.

[3] S. Haykin, D. J. Thomson, and J. H. Reed, "Spectrum sensing for cognitive radio," in *IEEE Transactions on Cognitive Radio*, May 2009.

[4] J. Oh and W. Choi, "A hybrid cognitive radio system: A combination of underlay and overlay approaches," in *IEEE Transactions on Cognitive Radio*, 2009.

[5] C. Stevenson, G. Chouinard, Z. Lei, W. Hu, J. Stephen, and W. Caldwell, "The first cognitive radio wireless regional area network standard," in *IEEE 802.22*, 2009.

[6] X. Xing, T. Jing, W. Cheng, Y. Huo, and X. Cheng, "Spectrum prediction in cognitive radio networks," in *1536-1284/13/$25.00 ©2013 IEEE Transactions on Wireless Communications*, April 2013.

[7] A. M. Wyglinski, M. Nekovee, and Y. T. Hou, *Cognitive Radio Communications and Networks*, 2009.

[8] M. Subhedar and G. Birajdar, "Spectrum sensing techniques in cognitive radio networks: A survey," *International Journal of Next-Generation Networks*, vol. 3, no. 2, pp. 37–51, 2011.

[9] T. W. Rondeau, B. Le, C. J. Rieser, and C. W. Bostian, "Cognitive radios with genetic algorithms: Intelligent control of software defined radios," in *©2004 SDR Forum, Proceeding of the SDR 2004 Technical Conference and Product Exposition*, 2004.

[10] S. K. Udgata, K. P. Kumar, and S. L. Sabat, "Swarm intelligence based resource allocation algorithm for cognitive radio network," in *Parallel Distributed and Grid Computing (PDGC), 2010 1st International Conference on*, Oct 2010, pp. 324–329.

[11] M. Matinmikko, J. Del Ser, T. Rauma, and M. Mustonen, "Fuzzy-logic based framework for spectrum availability assessment in cognitive radio systems," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 11, pp. 2173–2184, November 2013.

[12] L. Giupponi and A. Perez, "Fuzzy-based spectrum handoff in cognitive radio networks," 2008.

[13] Y. Chen and H.-S. Oh, "A survey of measurement-based spectrum occupancy modelling for cognitive radios," in *1553-877X ©2013 IEEE IEEE Communications Surveys and Tutorials*, 2013.

[14] R. Azmi, "Support vector machine based white space predictors for cognitive radio," Master's thesis, 2011.

[15] O. Winston, A. Thomas, and W. OkelloOdongo, "Optimizing neural network for tv idle channel prediction in cognitive radio using particle swarm optimization," in *Computational Intelligence, Communication Systems and Networks (CICSyN), 2013 Fifth International Conference on*, June 2013, pp. 25–29.

[16] M. I. Taj and M. Akil, "Cognitive radio spectrum evolution prediction using artificial neural networks based mutivariate time series modelling," in *European Wireless, Vienna Austria*, April 2011.

[17] X. Li and S. A. Zekavat, "Traffic pattern prediction and performance investigation for cognitive radio systems," in *IEEE Communication Society, WCNC Proceedings*, 2008.

[18] S. Hiremath and S. K. Patra, "Transmission rate prediction for cognitive radio using adaptive neural fuzzy inference system," in *IEEE 5th International Conference on Industrial and Information Systems (ICIIS), India*, Aug 2010.

[19] S. Geirhofer, J. Z. Sun, L. Tong, and B. M. Sadler, "Cognitive frequency hopping based on interference prediction: Theory and experimental results," vol. 13, no. 2, march 17, 2009.

[20] Z. Tabakovic, S. Grgic, and M. Grgic, "Fuzzy logic power control in cognitive radio," in *IEEE transactions*, 2009.

[21] Z. Lin, X. Jian, L. Huang, and Y. Yao, "Energy prediction based spectrum sensing approach for cognitive radio network," in *978-1-4244-3693-4/09/$25.00 ©2009 IEEE*, 2009.

[22] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed., 2008.

[23] Z. Jianli, "Based on neural network spectrum prediction of cognitive radio," in *978-1-4577-0321-8/11/$26.00 ©2011 IEEE*, 2011.

[24] V. Vapnik, *The nature of statistical learning theory*. Springer-Verlag New York Inc, 1999.

[25] V. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.

[26] E. Alpaydin, *Introduction to Machine Learning*, ser. Adaptive computation and machine learning. MIT Press, 2004.

[27] V. Kecman, *Learning and soft computing*. MIT Press Cambridge, Mass, 2001.

[28] C. Vladimir and Y. MA, "Selection of meta-parameters for support vector regression," pp. 687–693, August 2002.

[29] W. Wenjian, Z. Xu, W. Lu, and X. Zhang, "Determination of the spread parameter in the gaussian kernel for classification and regression," vol. 55, no. 3, pp. 643–663, October 2003.

[30] V. S. Cherkassky and F. Mulier, *Learning from Data: Concepts, Theory, and Methods*, 1st ed. New York, NY, USA: John Wiley and Sons, Inc., 1998.

[31] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[32] K. V. Price, R. Storn, and J. Lampinen, *Differential Evolution: A Practical Approach to Global Optimization*. Springer, 2005.

[33] A. K. Qin, V. L. Huang, and P. N. Suganthan, "Differential evolution algorithm with strategy adaptation for global numerical optimization," in *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 2, April 2009.

[34] D. Zaharie, "A comparative analysis of crossover variants in differential evolution," in *Proceedings of the International Multiconference on Computer Science and Information Technology*, 2007, pp. 171–181.

[35] E. Mininno, F. Neri, F. Cupertino, and D. Naso, "Compact differential evolution," *Evolutionary Computation, IEEE Transactions on*, vol. 15, no. 1, pp. 32–54, Feb 2011.

[36] G. Harik, F. Lobo, and D. Goldberg, "The compact genetic algorithm," *Evolutionary Computation, IEEE Transactions on*, vol. 3, no. 4, pp. 287–297, Nov 1999.

[37] C. W. Ahn and R. Ramakrishna, "Elitism-based compact genetic algorithms," *Evolutionary Computation, IEEE Transactions on*, vol. 7, no. 4, pp. 367–385, Aug 2003.

[38] A. H. Sung, "Ranking importance of input parameters of neural networks," *Expert Systems with Applications*, vol. 15, no. 3, pp. 405–411, November 1998.

[39] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[40] S. Iliya, E. Goodyer, J. Shell, J. Gow, and M. Gongora, "Optimized neural network using differential evolutionary and swarm intelligence optimization algorithms for rf power prediction in cognitive radio network: A comparative study," in *978-1-4799-4998-4/14/$31.00 ©2014 IEEE International Conference on Adaptive Science and Information Technology*, 2014.

DOI: 10.15439/2015F168
ACSIS, Vol. 5

# DISESOR - decision support system for mining industry

Michał Kozielski
Institute of Electronics,
Silesian University of Technology
ul. Akademicka 16,
44-100 Gliwice, Poland
Email: michal.kozielski@polsl.pl

Marek Sikora
Institute of Informatics,
Silesian University of Technology
ul. Akademicka 16,
44-100 Gliwice, Poland
Email: marek.sikora@polsl.pl

Łukasz Wróbel
Institute of Innovative Technologies EMAG
Leopolda 31,
40-189 Katowice, Poland
Institute of Informatics,
Silesian University of Technology
ul. Akademicka 16,
44-100 Gliwice, Poland
Email: lukasz.wrobel@ibemag.pl

***Abstract*—This paper presents the DISESOR integrated decision support system. The system integrates data from different monitoring and dispatching systems and contains such modules as data preparation and cleaning, analytical, prediction and expert system. Architecture of the system is presented in the paper and a special focus is put on the presentation of two issues: data integration and cleaning, and creation of prediction model. The work contains also a case study presenting an example of the system application.**

## I. Introduction

COAL mining is a heavy industry that plays an important role on an energy market and employs hundreds of thousands of people. Coal mining is also an industry, where large amount of data is produced but little is done to utilise them in further analysis. There is also a justified need to introduce a decision support system (DSS) integrating different aspects of coal mine operation in order to maintain continuity of mining.

Currently coal mines are well equipped with the monitoring, supervising and dispatching systems connected with machines, devices and transport facilities. There are also the systems for monitoring natural hazards (methane-, seismic- and fire hazards). All these systems are provided by many different companies, what causes problems with quality, integration and proper interpretation of the collected data. The collected data are used chiefly for current (temporary) visualisation on boards which display certain places in the mine. Whereas, application of domain knowledge and the results of historical data analysis can improve the operator's and supervisor's work significantly. For example, thanks to short-term prognoses about methane concentration, linked with the information about the location and work intensity of the cutter loader, it is possible to prevent emergency energy shutdowns and maintain continuity of mining (the research on this methodology was discussed in [1]). This will enable to increase the production volume and to reduce the wear of electrical elements whose exploitation time depends on the number of switch-ons and switch-offs.

It is possible to see the rising awareness of monitoring systems suppliers who begin to understand the necessity to make the next step in these systems development. Therefore, the companies providing monitoring systems seek their competitive advantage in equipping their systems with knowledge engineering, modelling and data analysis methods. This is a strong motivation to consider a DSS presented in this paper.

The goal of this paper is to present an architecture of the integrated decision support system DISESOR. The system integrates data from different monitoring systems and contains an expert system module, that can utilise domain expert knowledge, and analytical module, that can be applied to diagnosis of the processes and devices and to prediction of natural hazards. The special focus of the paper is put on the data integration and data cleaning issues realised by means of the data warehouse and ETL process. The work also contains a more detailed presentation of the prediction module, which is complemented by a presentation of a simple case of methane concentration prediction in a coal mine.

The contribution of the paper consists of:

- the architecture of the integrated decision support system DISESOR,
- presentation of the approaches to the preparation and cleaning of the data collected by monitoring systems,
- presentation of the prediction module architecture and principles of the module operation,
- case study presenting application of the presented system to methane concentration prediction in a coal mine.

The structure of the paper is as follows. Section II presents the works related to the presented topic. The architecture of the DISESOR system and its data repository are presented in section III. The more detailed descriptions of the data preparation and cleaning and prediction modules are presented in sections IV and V respectively. The case study of methane concentration prediction task is presented in section VI and section VII presents the final conclusions.

978-83-60810-66-8/$25.00©2015, IEEE

## II. Related work

The typical environments deployed in a coal mine are monitoring and dispatching systems. These systems collect a large number of data which can be utilised in further analysis, e.g., on-line prediction of the sensor measurements, which area was surveyed in [2]. The analysis can address different aspects of coal mine operation such as, e.g., equipment failure or natural hazards.

The examples of the research in the field of natural hazards in an underground coal mine cover, e.g., methane concentration prediction and seismic hazard analysis. The research on the prediction of the methane concentrations was presented in [3, 1]. Application of data clustering techniques to seismic hazard assessment was presented in [4]. There are also approaches to prediction of seismic tremors by means of artificial neural networks [5] and rule-based systems [6]. Each research listed above is a stand alone approach not incorporated into any integrated system.

Analytical methods that were mentioned require the data which are extracted, cleaned, transformed and integrated. Decision support systems utilise a data repository of some kind, e.g., a data warehouse [7]. The critical dependence of the decision support system on a data warehouse implementation and an impact of the data quality on decision support is discussed in [8].

There are applications of machine learning methods to diagnostics of mining equipment and machinery presented in literature [9, 10]. Also some initial concepts of the system that processes data streams delivered by the monitoring systems were presented in [11]. However, to the best of the authors knowledge there is no example of the integrated decision support system for monitoring processes, devices and hazards in a coal mine (except the work dealing with DSS for coal transportation [12] which loosely corresponds to the given topic).

## III. System architecture

The general architecture of the DISESOR integrated decision support system is presented in Fig. 1. The architecture of the system consists of data repository and data preparation and cleaning, that are presented in more detail in the following sections, and analytical, prediction and expert system modules shortly presented below, as they are not the main focus of the paper.

### A. Decision support system

The core of analytical, prediction and expert system modules is based on the RapidMiner [13] platform. The RapidMiner environment was customised to the requirements of the non-advanced user by disabling unnecessary options and views. Therefore, an advanced user can use the whole functionality of RapidMiner, whereas the non-advanced user can use such thematic operators as e.g., "Solve a methane concentration prediction issue" or "Solve a seismic hazard issue". Also due to the target application of the system in Polish coal mines the RapidMiner environment was translated into Polish. Finally, RapidMiner was extended in the created application by additional operators wrapping R [14] and MOA (Massive On-line Analysis) [15] environments.

The goal of the Data preparation and cleaning module, which is referred further as ETL2, is to integrate the data stored in data warehouse and process them to the form acceptable by the methods creating prediction and classification models. In other words the ETL2 module prepares the training sets.

Prediction module is aimed to perform incremental (on-line) learning of predictive models or apply classification and prediction models created in analytical module for a given time horizon and frequency of the values measured by the chosen sensors. This module also tracks the trends in the incoming measurements. The created predictive models are adapted to the analysed process on the basis of the incoming data stream and the models learnt on historical data (within the analytical module). The module provides the interfaces that enable the choice of quality indices and their thresholds that ensure the minimal prediction quality. If the quality of predictions meets the conditions set by a user, the predictions will be treated as the values provided by a soft sensor. They can be further utilised by e.g., expert system but also they can be presented to a dispatcher of a monitoring system.

Analytical module is aimed to perform analysis of historical data (off-line) and to report the identified significant dependencies and trends. The results generated by this module are stored in the repository only when accepted by a user. Therefore, this module supports a user in decision-making of what is interesting from monitoring and prediction point of view. It also provides additional information that can be utilised to enrich the knowledge of expert system or that can be utilised to comparative analysis. The module supports identification of changes and trends in the monitored processes and tools and it also enables to compare the operator's and dispatcher's work.

Expert system module is aimed to perform on-line and off-line diagnosis of machines and other technical equipment. It is also aimed to supervise the processes and to support the dispatcher or expert decision-making with respect to both technical condition of the equipment and improper execution of the process. The inference process is performed by means of classical inference based on stringent rules and facts or probabilistic inference based on belief networks. The system contains also a knowledge base editor that allows a user to define such rules and network. The expert system module is currently being developed.

### B. Data repository

Data repository was designed as a data warehouse of a snowflake structure (as some dimensions have multiple levels), that is presented in Fig. 2 in a reduced, general form. The structure of a data warehouse results from the analysis of databases of the existing monitoring systems and the characteristics of the known sensors. The full list of tables with their description is presented in Table I.

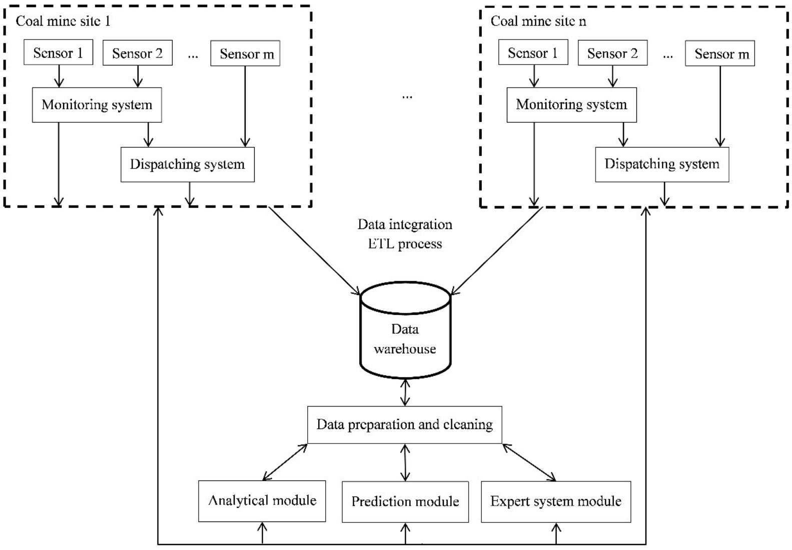


Fig. 1. Architecture of the DISESOR integrated decision support system

| Table | Description |
|---|---|
| Measurement | Value of a measurement |
| State | State of a measurement, e.g., *alarm*, *calibration*, *breakdown* |
| Discretisation | The measured values can be of discrete type |
| Time | Time of a measurement, range $[00:00:00, 23:59:59]$, 1 second resolution |
| Time_category | Category, e.g., *mining* or *no mining* |
| Date | Date of a measurement |
| Location | Location of the measurement source |
| Location_attribute | Characteristics of the given location |
| Location_hierarchy | Hierarchical structure of location |
| Source | Measurement source, e.g., sensor or device |
| Source_attribute | Characteristics of the given source |

TABLE I
TABLES CREATING A DATA WAREHOUSE STRUCTURE.

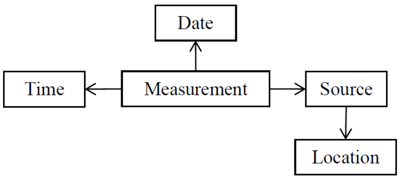


Fig. 2. Simplified schema of data repository

The central table of the data repository is *Measurement* where all the measurements are stored. The dimensions related to the *Measurement* table are *Date*, *Time* and *Source*. *Date* and *Time* describe when the measurement was registered, whereas *Source* describes what registered the given measurement. The *Source* table contains among others such information about sensors/devices as:

- name (e.g., MM256),
- description (e.g., methane meter number 256),
- type name (e.g., methane meter),
- measured quantity (e.g., methane concentration),
- measurement unit (e.g., %CH4),
- name of a system that collects the data (e.g., THOR),
- range of measurements.

The *Source* table is described by means of *Location* dimension, that describes where in a coal mine it is located. The location has hierarchical structure, some sample hierarchy is presented in Fig. 3. The top-most level of hierarchy are formed by coal mine divisions. Divisions consist of seams, which are divided into mining areas. At the bottom of hierarchy there are mining workings.

The data warehouse is loaded with data by means of the ETL process designed for the main monitoring and dispatching systems for coal mining, which are deployed in Poland, Ukraine and China, e.g., THOR dispatching system [16] or Hestia natural hazards assessment system [6]. The ETL process was designed by means of Open Talend Studio [17].

During the tests of the created solution the data warehouse was loaded with 800 million records what resulted in 200 GB of data. Therefore, it enabled the performance tests and optimisation of both the logical data warehouse structure and database management system (PostgreSQL [18]). As a result the *Measurement* data table was partitioned according

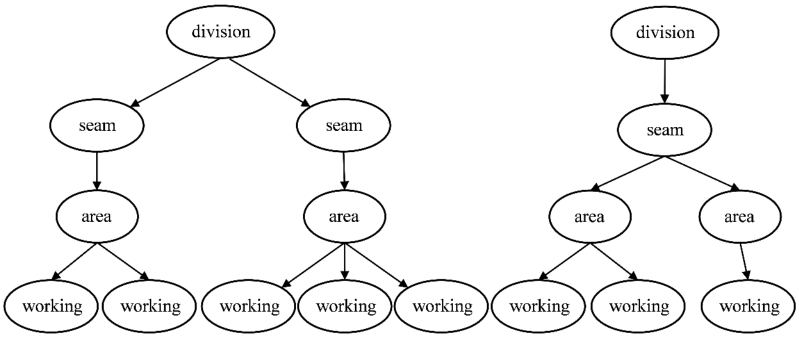


Fig. 3. Location hierarchy in a coal mine

to the months of measurements and the indices for foreign keys in this table were created. On the DBMS side several configuration parameters were adjusted, e.g., shared_buffers, work_mem, maintenance_work_mem, checkpoint_segments, checkpoint_completion_target, effective_cache_size.

## IV. Data preparation and cleaning

The goal of ETL2 module is to deliver integrated data (in a form of a uniform data set) coming from chosen sources (especially sensors) in a chosen time range.

The measurements can be collected with different frequencies. Additionally, some systems collect a new measurement only after significant (defined in a monitoring system) change of the measured value. Table II presents how the measurements of two methanometers can look like when collected directly from the data warehouse. The ETL2 process uniforms the data to the form where each recorded measurement represents the time period defined by a user, e.g., 1 second (Table III).

| MN234 | MN345 | T[s] |
|---|---|---|
| 0.1 | 0.1 | 0 |
| 0.2 | - | 1 |
| - | 0.2 | 4 |
| 0.5 | ? | 7 |
| 0.3 | 0.3 | 9 |

TABLE II
Data collected directly from data warehouse (- means that the measurement value does not change, ? means a missing value)

| MN234 | MN345 | T[s] |
|---|---|---|
| 0.1 | 0.1 | 0 |
| 0.2 | 0.1 | 1 |
| 0.2 | 0.1 | 2 |
| 0.2 | 0.1 | 3 |
| 0.2 | 0.2 | 4 |
| 0.2 | 0.2 | 5 |
| 0.2 | 0.2 | 6 |
| 0.5 | ? | 7 |
| 0.5 | ? | 8 |
| 0.3 | 0.3 | 9 |

TABLE III
Data prepared to the further transformation, cleaning, etc.

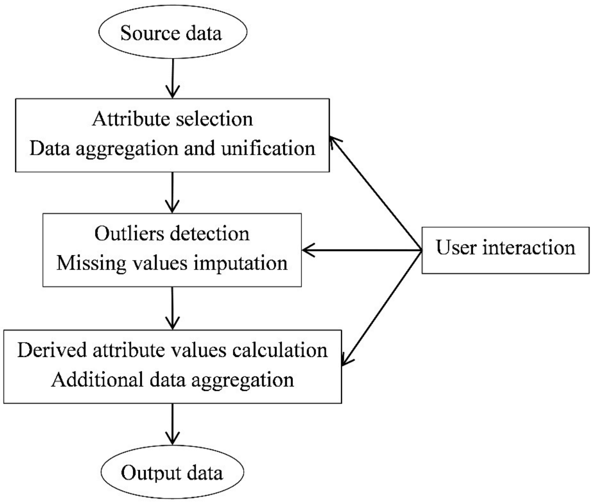

Source data
Attribute selection
Data aggregation and unification
Outliers detection
Missing values imputation
User interaction
Derived attribute values calculation
Additional data aggregation
Output data

Fig. 4. General characteristics of the data processing in ETL2 module

Within the ETL2 module there are also executed procedures of data cleaning, that identify outlier values and impute the missing values. This task is realised both by means of the simple functions presented below and by means of operators available in RapidMiner environment. Also data aggregation (e.g., 10 measurements are replaced with 1 measurement) and manual definition of derived variables (e.g., a new variable can be calculated as a sum of the values of two other variables) are performed by means of the methods included in ETL2 module.

The general scheme of data processing within ETL2 module is presented in Fig. 4.

As a result of the processing performed by means of the ETL2 module we receive a data set that can be either analysed (by means of analytical module), or utilised to prediction model creation (by means of prediction module), or utilised within diagnosis process (by means of expert system). All the phases of processing are performed as separate RapidMiner operators.

In order to select the variables that should be analysed a user can utilise THOR dispatching system 5, where each sensor (and attributes) are presented on a map of the region of interest. The system that is being created enables in turn, data (time-series) visualisation in order to select the time periods, that are the most interesting from the analyst point of view. Fig. 6 presents the visualisation of time-series consisting of several thousands of records. The developed operator creating such visualisation utilises R environment.

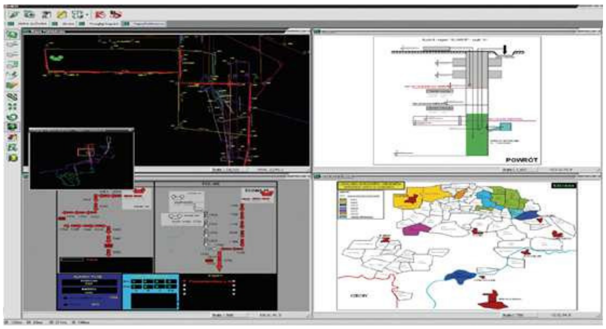


Fig. 5. Visualisation available in THOR dispatching system

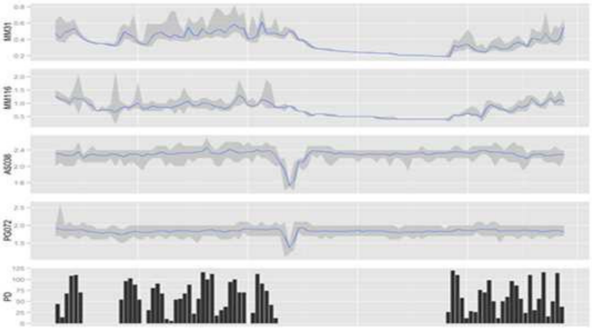


Fig. 6. Visualisation of exemplary time-series: methane concentration, air flow and mining cycle on a chosen longwall

Aggregation of the measurements replaces several values with a single one. The period of aggregation is chosen by a user, who sets a number of measurements that should be aggregated or a time unit defining the windows containing measurements to be aggregated. The following aggregation operators are available for each attribute: average, minimum, maximum, median, dominant, the number of occurrences. For each record being the result of the aggregation there is calculated a weight, that is inversely proportional to the number of missing values existing in the aggregated data. The weight calculation is also based on a weighted average for all the attributes. This approach enables us to reduce the number of missing values in data and introduce weights that can be utilised by the chosen methods (e.g., rule induction).

The operator that imputes missing values performs the analysis of each attribute separately. The following methods of changing the value or imputing the missing value can be utilised:

- a logical expression defining the replacing values (e.g., replace each value <1 with "low state"),
- the way how to receive the replacing values:
  - the value set by a user,
  - the last valid measurement,
  - average of the neighbouring measurements (with the parameter defining the number of neighbours),
  - linear regression of the two points (the last one before missing values section and the first one after this section),
  - linear regression of the data preceding missing values (with the parameter defining the window size).

The maximal number of consecutive missing values that can be imputed is defined as a separate parameter, as imputing the values for the long breaks in the measurements has no practical meaning. Therefore, the resulting data set can still contain missing values. In such case, the analyst can use a number of methods that are able to analyse data with missing values.

Introduction of a new derived variable can cover, among others, introduction of delays (the values of the previous measurements) or calculation of increments and trends (e.g., as an ordinal - increases, decreases). Another operator enables data smoothing by means of different filters (e.g., average, median). Finally, the last operator enables creation of dependent variable (decision variable). Typically, this variable contains the moved forward values of the chosen attribute, what enables to receive a proper prediction horizon. The operator defining the dependent variable has expanded functionality what enables e.g., to define the dependent variable as a maximal value of a given attribute in a defined time interval (e.g., 3 to 6 minutes in advance).

It is also important that within the developed framework the operators can be applied multiple times and in unrestricted order. Moreover, it is possible to pre-process data by means of the operators delivered by RapidMiner, that are dedicated to multidimensional analysis/identification of outliers and missing values (e.g., the operator applying local k-NN to missing values imputation).

When data pre-processing is finished, the whole process is saved according to RapidMiner-XML standard, that was created for the needs of the system. Thereby, the prediction module is able to transform the incoming data to the form that is acceptable by prediction models (see section V).

## V. Prediction module

Prediction module is based on, so called, prediction services. Prediction service is a webservice that predicts values of a variable (discreet or continuous) on the basis of input vector. Prediction service is inseparably connected with a model (regression or classification one) that is the basis of the prediction. The basic scenario of prediction service application is as follows:

1) Client sends a prediction execution request accompanied by a vector of conditional attributes and a timestamp.
2) Service calculates the prediction delivering the vector of conditional attributes as a model input. The attribute values come directly from the monitoring system, because

the data warehouse is not loaded online. The values of the attributes are transformed according to the dedicated ETL2 process to the form acceptable by the prediction model.

3) Service loads the results to a database.

The architecture of the prediction module is presented in Fig. 7.

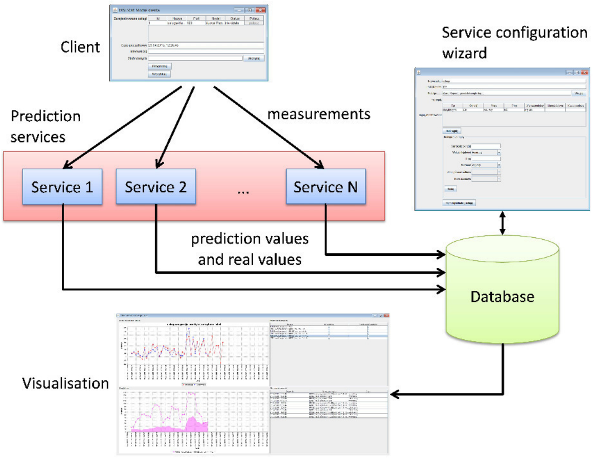


Fig. 7. Architecture and operation of prediction module

Database, which is an internal RapidMiner repository, stores the description of a model and the transformations of the attributes. It also stores the information about training data, the parameters of the minimal model quality and both predicted and real values of dependent variable. Each model adaptation results in a new database entry what makes the history of the changes available to the users.

The predictions can be visualised and compared on a single plot with the real values that are measured. Such visualisation can be performed by a monitoring or dispatching system (e.g., THOR dispatching system), where predicted values are delivered as measurements of a virtual sensor and the values of both sensors (virtual and real) can be easly compared.

It is assumed for the current module version, that if the quality of the predictions decreases below a given threshold, then a new training set is automatically collected. The size of this new data set is the same as size of the original data. The model adaptation is performed by modifying only the parameters of the existing model (the method and algorithm is not changed). Next, the quality of the model is verified on the same data that triggered the model adaptation (these data are not the part of the new training data set). If the quality of the adapted model is satisfactory, then this new model is applied to prediction. Otherwise a message is generated stating that prediction cannot be continued and it is needed to come back to analytical module in order to create a new prediction model.

The configuration wizard enables to define the so-called quality monitoring rules. From the practical point of view there is no point in presenting the minimum model quality by means of the well known measures, such as overall classification accuracy, g-mean, specificity, sensitivity, RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), etc. Therefore, quality monitoring rules are based on: a sliding time-window (e.g., 1 hour) in which the quality is verified, frequency of the prediction calculation (e.g., 1 minute) and the indicators which are typically called *FalsePositive* and *FalseNegative*. The values of these indicators are explicitly defined by a user for each decision class or only for a target class, e.g., corresponding to "danger". Therefore, knowing the values of *FalsePositive* and *FalseNegative*, and a number of predictions that are calculated in a given time-window it is possible to calculate the values of almost all the possible quality measures of prediction model. In case of regression task the module allows so-called insensitivity, what means that the predictions that differ less than the given threshold from the real values are not treated as an error. Additionally, it is possible to define that the values within the given range (e.g., corresponding to the "normal" state) are not counted as errors.

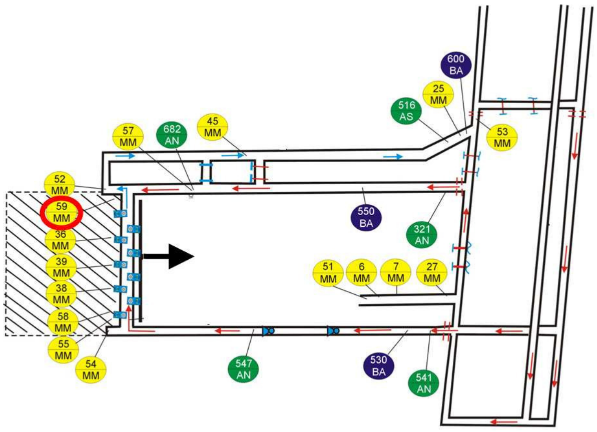


Fig. 8. Topology of the mining area and location of the sensors - MM59 sensor chosen as dependent variable is outlined a thick line

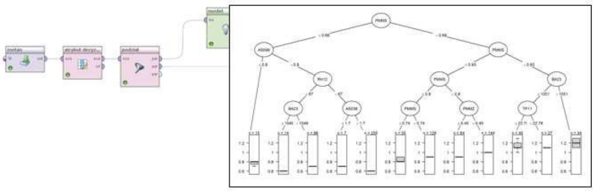


Fig. 9. The process of data preparation and prediction model creation together with the initial regression tree that was created

## VI. Example of the system application to the task of methane concentration prediction in mining excavation

The DISESOR system can be applied to several different tasks solving. This section presents an example, how the system can be utilised to methane concentration prediction.

Methane concentration monitoring is one of the main tasks of the natural hazard monitoring systems in mining industry.

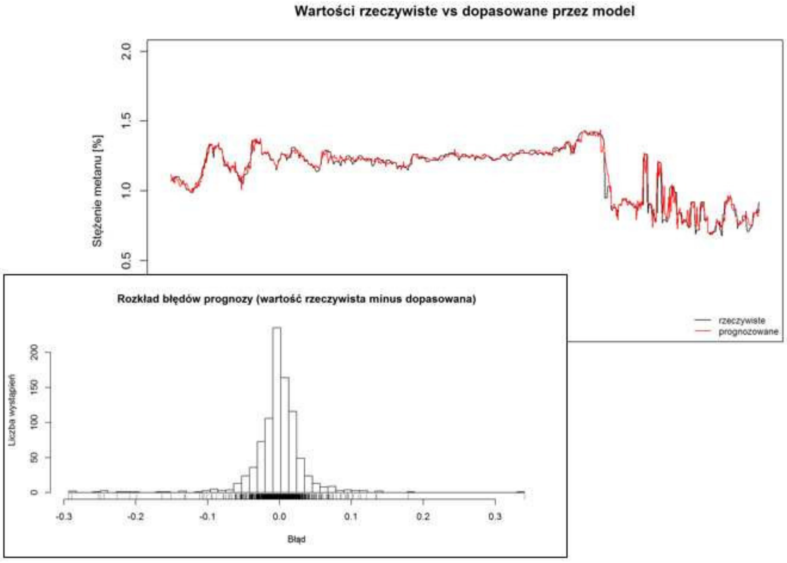


Fig. 10. The plot of the real methane concentration and the predicted maximum concentration together with the histogram of errors that are reported to a user

Such system is in charge of automatic and immediate shutdown of electricity within a given area, if a methane concentration exceeds a given alarm threshold. The power turn-on is possible after a certain time (from 15 minutes to even several hours), when the methane concentration decreases to the acceptable level. This results in large losses associated with downtime of production. Information from a soft (virtual) sensor presenting to a dispatcher the prediction of the methane concentration with a few minute horizon can allow the prevention electricity shut-down or can allow to lower the mining activity and increase the air flow if possible. Therefore, these actions allow to avoid undesirable situations and unnecessary downtimes.

The task of maximal methane concentration prediction with the horizon from 3 to 6 minutes was realised within the DISESOR system. By means of ETL2 module a set of the following sensors was selected: AN321, AN541, AN547, AN682, BA1000, BA603, BA613, BA623, MM11, MM21, MM25, MM31, MM36, MM38, MM39, MM41, MM45, MM52, MM53, MM54, MM55, MM57, MM58, MM59, MM61, MM81. The data were aggregated applying minimum operation to anemometer (AN) measurements, average operation to barometer (BA) measurements and maximum operation to methanometer (MM) measurements. The missing values were imputed applying linear regression method. As a dependent variable MM59 sensor was chosen. A map presenting the topology of the mining area and location of the sensors is presented in Fig. 8.

Analytical module is currently being developed and the analysis presented below is an example of the possible system apllication. Therefore, to create the examplary prediction model the method of regression tree induction was chosen arbitrary. The initial tree was created on the basis of data coming from 1 shift. The model and the list of sensors (variables) together with the defined transformations were forwarded to prediction model running a proper service. The time-window defined for prediction quality monitoring was set to 1 hour and the model adaptation was executed each hour regardless the minimum quality requirements. The adaptation could be executed more often if the minimum quality requirements were not met but there was no such situation. The data that were predicted were delivered on-line by the simulator of THOR system in order to simulate the real stream of measurements.

Fig. 9 presents the process of data preparation and the prediction model creation together with the initial regression tree that was created. Whereas, Fig. 10 presents the plot of the real methane concentration and the predicted maximum concentration together with the histogram of errors that are reported to a user. Currently, the user interface is in Polish as the deployment in Poland was planned in the project. However, the English and Chinese versions are also planned.

## VII. Conclusions

The system that is being developed delivers the solutions for decision support of a dispatcher and process operator. This system is complete as it delivers the tools that can be applied to data storage, processing and preparation, and also to definition of the models based on expert knowledge (expert system) and the models based on the results of both historical and on-line data analysis. Due to the application and proper customisation of existing tools (RapidMiner, R) and development of the proprietary solutions (e.g., ETL2, rule induction and optimisation [19, 20], rough set operators [21] and semantic analysis of data mining processes [22] that are not available in RapidMiner) a user receives a broad set of tools that can be applied to different tasks. Finally, the case study that was presented shows that the system can be practically utilised in a coal mine industry.

## Acknowledgement

This research was supported by Polish National Centre for Research and Development (NCBiR) grant PBS2/B9/20/2013 in frame of Applied Research Programme. The work was also financially supported by POIG.02.03.01-24-099/13 grant: GeCONiI Upper Silesian Center for Computational Science and Engineering.

## References

[1] M. Sikora and B. Sikora, "Rough natural hazards monitoring," in *Rough Sets: Selected Methods and Applications in Management and Engineering*. Springer, 2012, pp. 163–179. [Online]. Available: http://dx.doi.org/10.1007/978-1-4471-2760-4_10

[2] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Computers & Chemical Engineering*, vol. 33, no. 4, pp. 795–814, 2009. doi: 10.1016/j.compchemeng.2008.12.012. [Online]. Available: http://dx.doi.org/10.1016/j.compchemeng.2008.12.012

[3] M. Sikora and B. Sikora, "Improving prediction models applied in systems monitoring natural hazards and machinery," *International Journal of Applied Mathematics and Computer Science*, vol. 22, no. 2, pp. 477–491, 2012. doi: 10.2478/v10006-012-0036-3. [Online]. Available: http://dx.doi.org/10.2478/v10006-012-0036-3

[4] A. Leśniak and Z. Isakow, "Space-time clustering of seismic events and hazard assessment in the zabrze-bielszowice coal mine, poland," *International Journal of Rock Mechanics and Mining Sciences*, vol. 46, no. 5, pp. 918–928, 2009. doi: 10.1016/j.ijrmms.2008.12.003. [Online]. Available: http://dx.doi.org/10.1016/j.ijrmms.2008.12.003

[5] J. Kabiesz, "Effect of the form of data on the quality of mine tremors hazard forecasting using neural networks," *Geotechnical & Geological Engineering*, vol. 24, no. 5, pp. 1131–1147, 2006. doi: 10.1007/s10706-005-1136-8. [Online]. Available: http://dx.doi.org/10.1007/s10706-005-1136-8

[6] J. Kabiesz, B. Sikora, M. Sikora, and Ł. Wróbel, "Application of rule-based models for seismic hazard prediction in coal mines," *ACTA MONTANISTICA SLOVACA*, vol. 18, no. 4, pp. 262–277, 2013.

[7] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[8] S. T. March and A. R. Hevner, "Integrated decision support systems: A data warehousing perspective," *Decision Support Systems*, vol. 43, no. 3, pp. 1031–1043, 2007. doi: 10.1016/j.dss.2005.05.029. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2005.05.029

[9] M. Michalak, M. Sikora, and J. Sobczyk, "Analysis of the longwall conveyor chain based on a harmonic analysis," *Eksploatacja i Niezawodność - Maintenance and Reliability*, vol. 15, no. 4, pp. 332–333, 2013.

[10] M. Kalisch, P. Przystalka, and A. Timofiejczuk, "Application of selected classification schemes for fault diagnosis of actuator systems," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*. IEEE, 2014. doi: 10.15439/2014F158 pp. 1381–1390. [Online]. Available: http://dx.doi.org/10.15439/2014F158

[11] M. Grzegorowski, "Scaling of complex calculations over big data-sets," in *Active Media Technology*. Springer, 2014, pp. 73–84. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-09912-5_7

[12] E. Kozan and S. Q. Liu, "A demand-responsive decision support system for coal transportation," *Decision Support Systems*, vol. 54, no. 1, pp. 665–680, 2012. doi: 10.1016/j.dss.2012.08.012. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2012.08.012

[13] RapidMiner. (2015) Rapidminer. [Online]. Available: http://rapidminer.com

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org

[15] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: Massive online analysis," *The Journal of Machine Learning Research*, vol. 11, pp. 1601–1604, 2010.

[16] Sevitel. (2015) Thor. [Online]. Available: http://www.sevitel.pl/product,25,THOR.html

[17] Talend. (2015) Talend open studio. [Online]. Available: https://www.talend.com/products/talend-open-studio

[18] PostgreSQL. (2015) Postgresql. [Online]. Available: http://www.postgresql.org/

[19] T. Amin, I. Chikalov, M. Moshkov, and B. Zielosko, "Relationships between length and coverage of decision rules," *Fundam. Inform.*, vol. 129, no. 1-2, pp. 1–13, 2014. doi: 10.3233/FI-2014-956. [Online]. Available: http://dx.doi.org/10.3233/FI-2014-956

[20] U. Stanczyk, "Decision rule length as a basis for evaluation of attribute relevance," *Journal of Intelligent and Fuzzy Systems*, vol. 24, no. 3, pp. 429–445, 2013. doi: 10.3233/IFS-2012-0564. [Online]. Available: http://dx.doi.org/10.3233/IFS-2012-0564

[21] L. S. Riza, A. Janusz, C. Bergmeir, C. Cornelis, F. Herrera, D. Ślęzak, J. M. Benítez *et al.*, "Implementing algorithms of rough set theory and fuzzy rough set theory in the r package "roughsets"," *Information Sciences*, vol. 287, pp. 68–89, 2014. doi: 10.1016/j.ins.2014.07.029. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2014.07.029

[22] A. Lawrynowicz and J. Potoniec, "Pattern based feature construction in semantic data mining," *Int. J. Semantic Web Inf. Syst.*, vol. 10, no. 1, pp. 27–65, 2014. doi: 10.4018/ijswis.2014010102. [Online]. Available: http://dx.doi.org/10.4018/ijswis.2014010102

DOI: 10.15439/2015F264
ACSIS, Vol. 5

# Emotion Monitor - Concept, Construction and Lessons Learned

Agnieszka Landowska
Gdansk University of Technology, Narutowicza St. 11/12, 80-233, Gdansk, Poland
E-mail: nailie@eti.pg.gda.pl

*Abstract*—**This paper concerns the design and physical construction of an emotion monitor stand for tracking human emotions in Human-Computer Interaction using multi-modal approach. The concept of the stand using cameras, behavioral analysis tools and a set of physiological sensors such as galvanic skin response, blood-volume pulse, temperature, breath and electromyography is presented and followed by details of Emotion Monitor construction at Gdansk University of Technology. Some experiments are reported that were already held at the stand, providing observations on reliability, accuracy and value the stand might provide in human-systems interaction evaluation. The lessons learned at this particular stand might be interesting for the other researchers aiming at emotion monitoring in human-systems interaction.**

## I. Introduction

THIS paper concerns challenges in automatic multimodal affect recognition. Although it seems that the domain is well established and there are many off-the-shelf solutions, the reliability, accuracy and granularity of emotion recognition is still a challenge. In 2013 a project was started at Gdansk University of Technology (GUT) to build an emotion monitor stand that uses existing technologies in order to extend human-systems interaction with emotion recognition and affective intervention. The concept of the stand assumed combining multiple modalities used in emotion recognition in order to improve the accuracy of affect classification. The considered input channels included the ones that are most frequently used in the emotion recognition: physiological signals (skin conductance, respiration, electromyography, EEG, heart rate, peripheral temperature) [1], video input for facial expression analysis [2], keyboard and mouse usage patterns [3] as well as textual inputs for sentiment analysis [4]. The hardware layer of the stand was constructed in 2013, and the software layer in 2014 and 2015, however, the latter still requires extension, including improvement of classification algorithms. The paper describes the concept and construction details of the Emotion Monitor stand at GUT. Selected experiments held at the stand are described that provide an insight into practical aspects of automatic emotion recognition. The experiments are diverse, from the ones that aimed at establishment of reliable measurement procedures of physiological signals, ones that aimed at classification algorithms training and others that were practical applications of the stand in systems design. Although not every experiment was successful, all of them contributed to the knowledge on practical aspects of multimodal emotion monitoring. Lessons learned on the way are the main theme of this paper and the research questions of the article might be formulated as follows: *how to monitor emotional states in Human-Computer Interaction with acceptable reliability, accuracy and granularity and what are the main challenges in automatic multimodal affect recognition?* The main purpose of the paper is to evaluate the usability of the stand as well as to express the main limitations of emotion recognition in human-computer interaction.

## II. Related work

Works that are mostly related to this research fall into two categories: research on applicability of emotion monitoring in the context of human-system interaction and studies on emotion recognition based on different input channels.

First group of related papers provides rationale for emotion recognition application in human-systems interaction. Software usability testing can be extended with observation of human emotions [5][6] and it is also possible to measure and optimize software development processes [7][8][9]. Based on the methods for usability evaluation, it would be possible to evaluate educational software and resources designed for e-learning [10][11][12]. Physiological parameters can be also used for optimization of other emotion recognition algorithms and in affect-aware games and other intelligent personalized systems [13].

There are numerous emotion recognition algorithms that differ on input information channels, output labels, affect models and classification methods. As literature on affective computing tools is broad and has already been summarized several times (eg. [14]), only example papers are referenced. The most frequently used emotion recognition methods that might be considered for emotion monitor stand include:

- facial expression analysis (requires video as an input channel, however expressions might be partially controlled by people, especially when they know they are being observed or recorded) [2][14][16];
- audio (voice) signal analysis in terms of modulation (this method is seldom used in human-computer interaction as the voice communication channel is rarely used) [2][16];

This work was supported by Polish-Norwegian Financial Mechanism Small Grant Scheme under the contract no Pol-Nor/209260/108/2015 and by DS Funds of ETI Faculty, Gdansk University of Technology.

- textual input analysis (sentiment analysis requires conversational system interface) [17];
- physiological signals - although very precise and cannot be controlled by most of the people, require specialized equipment [1][15];
- behavioral patterns analysis (keystroke dynamics and mouse usage patterns) combined with other modalities can improve the accuracy of affect recognition; moreover those are the most natural input channels in HCI [3].

The best recognition results are obtained when fusing information from diverse input channels and early and late fusion can be distinguished [18]. Early fusion methods combine features derived from separate input channels to create a common feature vector for classification [19]. Late fusion combines the classification results provided by separate classifiers for every input channel; however, this requires some mapping between emotion representation models used as classifier outputs [20]. The highest accuracies are obtained mainly for two-class classifiers and multimodal input channels (including physiological measurements).

## III. Emotion monitor concept

The emotion monitor stand objective is to conduct experiments on computer users affective states retrieval and analysis. The stand is equipped with computers, cameras and a set of biosensors, which allow to monitor user activities and record multiple user observation channels at the same time. The data are then processed further to extract features and classify emotional states from single or multimodal input channels. Schema of the emotion monitor stand is provided in Figure 1.

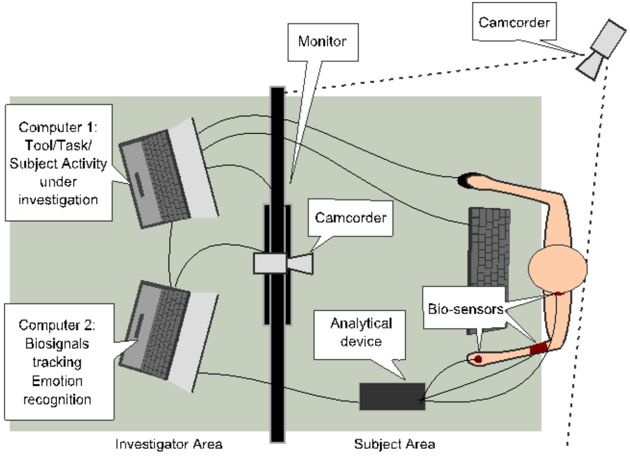

Camcorder
Monitor
Computer 1: Tool/Task/ Subject Activity under investigation
Camcorder
Computer 2: Biosignals tracking Emotion recognition
Analytical device
Bio-sensors
Investigator Area
Subject Area

Fig. 1 The concept of emotion monitor stand hardware configuration

The emotion monitor stand is divided into two separated areas: a subject area, where a user performs tasks that are under observation and an investigator area, where a researcher is able to monitor user's activities and track biometric parameters. During investigation, when different emotional states would be evoked on purpose, it would disturb experiments, when a subject would have a possibility to turn to the investigator personally, therefore the investigator and subject areas are separated.

The emotion monitor stand is equipped with the following devices: (1)biometric sensors set, including skin conductance, blood-volume pulse, respiration, temperature, electromyography and EEG sensors, (2) analytical device that allows to simultaneously sample multiple channels with high frequency, (3)front camera that records face and upper part of the subject's body, (4)side camera that records experiment execution, (5)computer 1, which allows the subject to perform tasks under investigation, (6)computer 2, which allows the investigator to monitor user activities and parameters.

Software layer of emotion monitor includes an application to store and track biometric data, tools for observation and recording of video images, keyboard and mouse usage tracker and user activity logger. Apart from the applications recording input channels, the main emotion monitor's application is the one that combines input channels and multiple classifiers in order to provide an affective state estimate. The result might be displayed with diverse visualization tools (general or dedicated for emotion representations).

## IV. Emotion monitor hardware layer construction

In 2013 an emotion monitor stand was constructed at Gdansk University of Technology as a dedicated stand in research laboratory room. The investigator's area and the subject's area are separated with a part-wall made out of furniture, which allows for visual, and partly acoustic separation. Photos of the subject's and the investigator's area are provided in Fig. 2 (left and right respectively).

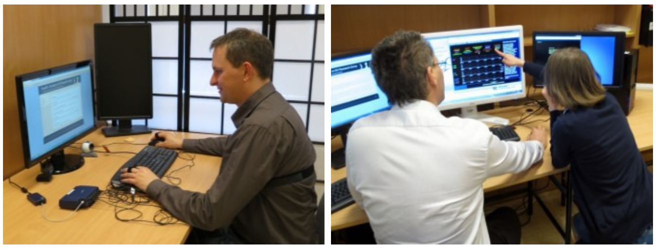


Fig. 2 Emotion monitor stand at GUT (left - the participant's area, right - the investigator's area)

The equipment of the stand was chosen based on capabilities and availability. Coder FlexComp Infiniti by Thought Technology, Canada was chosen as biosensors analytical device. The coder is a ten-channel multimodal device dedicated for real-time biofeedback, psychophysiology training and monitoring. It is connected with computer database via TT-USB device and allows to simultaneously record up to ten channels with sampling rate 2048 samples per second. The coder removes noise from all input channels and performs signal amplification and preliminary filtration that is adjusted to sensor type. EEG sensors, which are compatible with the coder, allow to perform impedance measurement, which allows to provide EEG signal of high quality. Other available devices had less input channels, lower sampling rates or did not allow to perform impedance measurement.

The biometric sensor set for the emotion monitor stand was designed to measure physiological parameters that are commonly used in affect recognition. and the choice was justified by literature review. Sensors are compatible with the FlexComp Infiniti coder and other coders produced by Thought Technology. Some of the sensors in the predefined set were doubled in order to try out multiple locations at the same time. Detailed list of sensor types includes: skin conductance, electromyography, respiration, temperature, electroencephalography and blood-volume pulse sensors.

Additionally three standard computer sets with two monitors each were provided for the stand. One computer is dedicated for biometric recording and emotion recognition and two monitors allow to display more parameters at the same time. The second computer is provided for a subject to perform tasks under investigation, additional monitor allows an investigator to track user's progress with the tasks. The third computer was added for the investigator-subject communication (investigator displays commands for a subject) – see a black monitor on the left photo in Figure 2.

The stand is normally equipped with three cameras: two in front of the subject and one side camera. The front cameras include one standard RBG camera of medium quality and additionally RGB-D camera with infrared depth sensor that allows for posture analysis independent of illumination for special applications.

## V. Emotion monitor software layer structure

There is a number of applications installed at the stand, however not all of them are used simultaneously. The applications might be divided into the following categories: (1) tools for input channels recoding and pre-processing; (2) applications for the data processing into feature vectors and classifiers' training; (3) software for classification and validation of the results (might be the same tools as above); (4) tools for emotional state visualization and interpretation.

The conceptual and actual diagram regarding software layer of the emotion monitor stand is provided in Fig. 3.

### *A. Data acquisition tools*

Thought Technology BioGraph Infiniti application was installed for gathering biometric data from the coder. The system was chosen due to compatibility with the coder, but also due to signal quality optimization features including verifying signal quality and adjusting sensor placement, integrated electrode impedance measurement as well as artifact rejection feature, both automatic and manual. Additionally, an optional Physiology Suite, which is specifically designed for monitoring and assessing physiological functions: recording biomeasurement sessions, reviewing recorded data for the purpose of artifact rejection, generating session reports and demonstrating the results, was installed.

Emotion monitor is also equipped with a set of applications for computer user behavioral observations: Mobii eye tracker, keystroke tracker and mouse tracker, Morae Recorder and Observer, and Logitech Video Capture.

As in some experiments there is a need for simultaneous recording of up to 6 camera images, the change of the software recording video is considered, as it allows to capture one camera image only (at one computer) and in the experiments with multiple cameras is not sufficient.

### *B. Training, classification and validation tools*

This part of the emotion monitor stand is still under

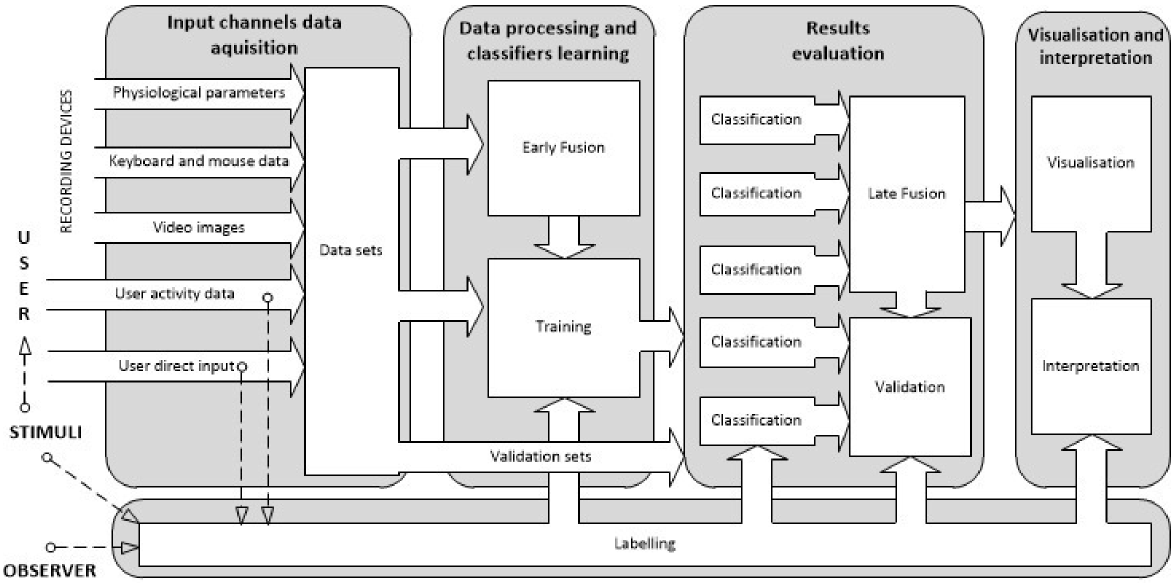


Fig. 3 Emotion monitor stand at GUT - the software layer

development, as there are several challenges in the automatic classification of computer users emotional states. Although the concept was to train classifiers off-line and then use them in real time, this would require more research, than was assumed. Therefore currently, during experiments only data acquisition is performed and the analysis and interpretation is performed afterwards. The analytical and training tools we use include the following: Morae Manager, Knime, Statistica, MatLab, SAS, Origin.

### *C. Visualization tools*

There is an emotional state visualizer that was prepared for the stand at GUT, however now it is not used, as the integration tool for real-time analysis and visualization is under development. Therefore apart from the dedicated tools we use a number of external tools for the data visualization, eg. Knime, Origin.

### *D. Early and late fusion*

The main concept of the stand was to perform the integration of emotional activation information from multiple input channels. One might consider early fusion, in which the data is combined to create common feature vectors. This approach has at least two important drawbacks: timing synchronization and temporary unavailability of input channels.

The first challenge we have encountered is timing synchronization. It seems simple with the timestamps provided by computers, however in practical acquisition of the affective activations the issue lies in the delay of the emotion expression for different channels, eg. heart rate change and skin conductance change last up to one/two seconds, facial expressions are delayed in comparison with the physiological signals. This is the result of the sympathetic and parasympathetic system activation and this is how it works. As a result it is hard to assign the same label to exactly the same period of time for different channels.

Another challenge is temporary unavailability of the input channels. All of the channels used are subject to temporal unavailability: for biosignals movement artifacts must be removed, as they interfere with informative peaks, face recognition is dependent on: face position and illumination conditions eg. if a user moves head a little bit, face recognition tool must follow the face (find it again), moreover keyboard and mouse are usually used interchangeably, with pauses. The common feature vectors are full of blank values, asynchronously for the input channels.

Late fusion that is based on the integration of the recognized emotional states from different classifiers suffers from diverse emotion representation models and lack of mapping between them. Facial expression analysis algorithms for emotion recognition use Ekman's Facial Coding System and provide six basic emotions as a result. The biosignals are best in recognition of the arousal dimension of emotional state, and not the valence (positive and negative experience might cause the same activation of the nervous system). There is a constant challenge of labeling, which will be depicted with the experiment in section VI.

## VI. Experiments in emotion monitoring

There were 5 experiments already held at the stand: (1) a study on reliability of physiological signals measurements in the HCI context, (2) experiment with picture stimuli, two experiments (3) and (4) with sound that provided the knowledge required for the optoelectronic system for autistic children, and finally (5) game experience monitoring.

As some of the experiments were already reported [21] [22][23][24], this section would summarize results of experiments (1), (3), (4) and (5) as well as provide more detailed description of experiment (2), which was not reported before. The descriptions would focus on revealing observations on usefulness of the emotion monitor stand.

### *A. Experiment 1. The challenge of biosignals acquisition in human-computer interaction*

After the stand was constructed at GUT in 2013, the first challenge was encountered in the sensitiveness of biometric sensors readings to movements. The typical locations of the sensors are finger tips or finger bases, which is inconvenient while using mouse and/or keyboard in human-computer interaction. Therefore, an experiment aiming at eliminating sensors from hands and finding alternative locations that are as good as typical finger placement was held. As the experiment was already reported in detail [21], only the major findings are summarized. The experiment allowed to draw some conclusions on human-computer interaction monitoring based on bio-measurements of muscle electric activity, respiration, temperature, pulse or skin conductance:

1. Emotion recognition in human-computer interaction should not use EMG measurements placed on trapesius muscle nor sensors located at finger tips (temperature, BVP sensor). Alternative locations of temperature and BVP sensor on earlobes could be accepted as a solution for human-computer interaction monitoring.

2. Skin conductance sensor location on forearm is perceived as less disturbing than location on fingers. Location on forearm is also less sensitive to mouse movements.

3. Respiration sensor are perceived as low disturbing and insensitive to movements, except from body movements.

4. For the signal recorded by all sensors artifacts connected with large movements (body movements) should be removed independently of their location on body.

In emotion recognition algorithms, which are based on biomeasurements, relative values should be provided rather than absolute values, as there are significant differences between individuals. Normalization or standardization procedure should be performed with personal average, instead of overall average calculated for all subjects.

Therefore baseline recording is important for experimental settings, as well as natural environments [21]. The experiment allowed to find a method for reliable acquisition of physiological signals in human-computer interaction.

### *B. Experiment 2. The challenge of labelling*

The concept of the experiment assumed registration of biometric parameters when evoking emotions on the basis of pictures. GAPED (Geneva Affective Picture Database) set was chosen, as the pictures in the set are labeled with the emotional activations in PAD (Pleasure Arousal Dominance) model. The experiment aimed at: acquisition of data for learning algorithms that cause emotions as well as determination whether the readings in alternative locations of sensors vary with emotions.

The pictures were grouped into 6 groups of similar emotional activations for: fear, sadness, anger, disgust and joy. The sixth group represented photos considered neutral for the reference. In each group 5 pictures were chosen and displayed one after another. Picture groups were separated with rest, when no picture was shown. Moreover, before the slide show of photos, a baseline was recorded. Necessity of the baseline recording results from the diversity of individual biometric readings.

The experiment succeeded in determining whether the readings in alternative locations of sensors vary with emotions, however failed in training classifiers for emotion recognition. The resulting recognition rate for the six emotional states never reached more than 30% independently of the classifier, training method and its parameters. Detailed analysis of the results for specific people revealed that pictures were not efficient stimuli to evoke emotions for part of them. There were a number of computer science students involved in the experiment, for whom even the drastic pictures from wars and hospitals were not enough to evoke reactions, as they are used to such views by intensively playing shooter games. For this group of people, the slide show of pictures that lasted for 20 minutes was simply boring and it was visible in physiological signals that headed towards relaxation state, independently of how drastic picture was shown.

Another interesting group of participants reacted to each and every picture independently of what was presented, even a photo of a blue mug or a box caused the reaction (visible with skin conductance fluctuation). This group of participants, which might be described as highly-reactive, also sometimes exhibited skin conductance raise before the picture was actually shown.

The minority of the participants exhibited the expected reactions: high for the drastic photos and low for more neutral ones.

The experiment revealed that choosing a stimuli is an important matter (obviously), but also that stimuli's label is not enough for labeling the data for emotion recognition. In the picture set there was one picture repeated twice in different contexts (in different picture groups) and the reaction to it depended mostly on the context, and not the actual photo (if any reaction was there). As a result we have tried alternative labeling methods, however they were found insufficient in this experiment.

However, the experiment brought to deeper understanding of the challenge of labeling with emotional states and those are the lessons learned:

1) There is a difference between the expression of emotion and the actual emotion eg. one might smile, although feeling embarrassed by the picture.

2) People exhibit more facial expressions when talking with others than in front of the computer screen (eg. typical surprise reaction of “jaw dropping” was never encountered). Micro-expressions must be recognized and interpreted instead.

3) Expectation of a stimuli results in pre-condition reaction (difficult to synchronize it with label).

4) There is no way of actually determine, what the emotional state of a person is (emotion has external expressions, however it is an internal phenomena, there is a two-factor theory of emotional reaction: both stimuli and the interpretation is required for the emotion to appear).

5) Some people do not exhibit facial expressions, for the others, the actual expression varies significantly.

6) When labeling pictures, sounds, videos or other stimuli with questionnaires one might obtain anticipated emotional state instead of the actual one.

7) People differ significantly in the ability to recognize and express their own emotional states.

8) Emotional reactions might be caused by some internal thoughts (some participants exhibited skin conductance fluctuations during baseline and rest recordings).

Although finished with no success, it was a valuable experiment, which revealed lots of research issues to work.

### *C. Experiments 3 and 4. Practical applications*

The other three experiments were practical applications of the stand. Experiments (3) and (4) were conducted to construct an optoelectronic system supporting behavioral therapy of autistic children [22][23]. The first experiment aimed at selection of physiological parameters which are closely correlated with person's emotional state. Measuring changes of those parameters and adequate data processing can show the emotion of investigated person. The experiment used a stimuli of 1 minute sound that started with 1kHz constant sound that was gradually silenced and finished with a shot sound. Only few parameters gave very strong change in measured signal after the shot sound: skin conductance, respiration and electromyography, however the individual reaction varied for subjects. The respiration rate and skin conductance changes were chosen to be monitored in the elaborated system. Another experiment (4) was conducted in order to evaluate the prototype of the optoelectronic system supporting behavioral therapy of autistic children. Recently, this system has been tested in one kindergarten for children with disabilities. Such support will be very useful and can significantly improve psychological treatment of those children [23].

## VII. Results and discussion

Apart from typical challenges in classification: feature selection, choosing classifier and its structure, proper training methods and validation, there are more challenges in automatic emotion recognition. The first one ins reliable data acquisition, as all input channels are subject to some noise and temporal unavailability. The main challenge seem still the labeling with emotional states. One might consider labeling with user reports, expert observations, user activity or stimuli labels, however all of the techniques could be questioned. Perhaps a combination of the two or three different labeling methods is a way, however the problem of blending in the case of constrictions remains. Another challenge is the fusion (early or late) of the emotional expressions estimate from different input channels and this field has gathered some researcher attention so far [19], however still much is to be revealed.

The author is aware of the fact that this study is not free of some limitations. First of all, the report on the experiment is subjective one, as the paper aimed at sharing lessons learned rather than reporting the actual experiments.

The question formulated at the beginning of the paper *how to monitor emotional states in Human-Computer Interaction with acceptable reliability, accuracy and granularity?* remains open and requires further research. However, the second question *what are the main challenges in automatic multimodal affect recognition?* , which was the main purpose of this study, was provided with some answers.

## VIII. Conclusion

Although there are some off-the-shelf solution for recognizing human affect (produced by Affectiva or Empathica) as well as many “smart” watches that track physiology, determining the actual emotional state of a human being is still a challenge, even for qualified psychologists. With all the complicated equipment and algorithms the only thing we could track is emotion’s external symptoms. Moreover the reliability and the accuracy of the provided estimate depends on many conditions: availability of the input channels, air conditions (temperature, humidity) and the context a human is in. Perhaps the internal phenomena of the emotion is what makes us really unpredictable, i.e. humans.

## References

[1] Szwoch W. (2013) Using physiological signals for emotion recognition, In Proc of HSI, Gdańsk, Poland, 556-561.

[2] Zeng Z, Pantic M, Roisman G, Huang T.S (2009) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(1), 39-58

[3] Kołakowska A. (2013) A review of emotion recognition methods based on keystroke dynamics and mouse movements, In Proc of HSI, Gdańsk, Poland, 548–555.

[4] Cambria, E, Schuller, B, Xia, YQ, Havasi, C (2013) New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 28: pp. 15-21

[5] Partala T., Surakka V. (2004) The effect of affective interventions in human-computer interaction, Interacting with Computers, 16, pp. 295-309

[6] Hazlett R., Benedek J. (2007) Measuring emotional valence to understand the user’s experience of software, Int. J. Human-Computer Studies, 65, 306-314.

[7] Zimmermann P., Gomez P., Danuser B., Schar S. (2006) Extending usability: putting affect into the user-experience, In Proc. of Nordic Conf. on Human-Computer Interaction, Oslo, pp 27-32.

[8] Kołakowska A, Landowska A, Szwoch M, Szwoch W, Wrobel M R (2013) Emotion Recognition and its Application in Software Engineering, In Proc of HSI, Gdańsk, Poland, 532–539.

[9] Wróbel M.R. (2013) Emotions in the software development process, In Proc of HSI, Gdańsk, Poland, 518-523.

[10] Binali H, Wu C, Potdar V (2009) A new significant area: Emotion detection in e-learning using opinion mining techniques. In: Proc. of 3rd IEEE International Conference on Digital Ecosystems and Technologies, 2009, 259-264

[11] Landowska A (2013) *Affective computing and affective learning – methods, tools and prospects*, EduAction. Electronic education magazine, 1(5), 16-31

[12] Landowska A. (2013) Affective computing and affective learning – methods, tools and prospects, EduAction. Electronic education magazine, 1, 5, pp. 16—31

[13] Chittaro L., Sioni R. (2014) Affective Computing vs. Affective Placebo: Study of a Biofeedback-Controlled Game for Relaxation Training. International Journal of Human-Computer Studies, 72, 8–9, pp. 663–73. doi:10.1016/j.ijhcs.2014.01.007.

[14] Gunes H., Schuller B. (2013) Categorical and dimensional affect analysis in continuous input: Current trends and future directions, Image and Vision Computing, 31, pp. 120-136

[15] Bailenson J.N., Pontikakis E.D., Mauss I.B., Gross J.J., Jabon M.E, Hutcherson C.A.C., Nass C., John O. (2008) Real-time classification of evoked emotions using facial feature tracking and physiological responses, International journal of human-computer studies, 66(5), 303-317

[16] Picard R, Daily S (2005) Evaluating affective interactions: Alternatives to asking what users feel. In CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches

[17] Ling H.S., Bali R, Salam R.A. (2006) Emotion detection using keywords spotting and semantic network, In Computing & Informatics, IEEE, 1-5

[18] Landowska A, Szwoch W, Szwoch M, (2015) Methodology of Affective Intervention Design for Intelligent Systems, Interactions with Computers (unpublished).

[19] Gunes H. and Piccardi M (2005) Affect Recognition from Face and Body: Early Fusion versus Late Fusion, Proc. IEEE International Conference on Systems, Man and Cybernetics, pp. 3437-3443.

[20] Hupont I; Ballano S; Baldassarri S.; Cerezo, E, (2011) Scalable multimodal fusion for continuous affect sensing, IEEE Workshop on Affective Computational Intelligence (WACI), pp.1,8, 11-15

[21] Landowska A.: Emotion monitoring - verification of physiological characteristics measurement procedures, Metrology and Measurement Systems Journal, Vol XXI, No. 4, 2014, pp. 719-732.

[22] Landowska A, Karpienko K, Wróbel M, Jędrzejewska-Szczerska M (2014) Selection of physiological parameters for optoelectronic system supporting behavioral therapy of autistic children, Proc. SPIE Vol. 9290, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments.

[23] Jędrzejewska-Szczerska M, Karpienko K, Landowska A (2015), System supporting behavioral therapy for children with autism, Journal of Innovative Optical Health Sciences Vol. 8, No. 3, 1541008

[24] Landowska A, Wrobel M (2015) Affective reactions to playing digital games, Int. conf. on Human-Systems Interaction, Warsaw, Poland, pp. 264-270

DOI: 10.15439/2015F384
ACSIS, Vol. 5

# Pawlak's flow graph extensions for video surveillance systems

Karol Lisowski*, Andrzej Czyżewski*
*Electronics, Telecommunications, and Informatics Faculty
Gdansk University of Technology, Narutowicza 11/12, PL80-233 Gdansk, Poland

***Abstract*—The idea of the Pawlak's flow graphs is applicable to many problems in various fields related to decision algorithms or data mining. The flow graphs can be used also in the video surveillance systems. Especially in distributed multi-camera systems which are problematic to be handled by human operators because of their limited perception. In such systems automated video analysis needs to be implemented. Important part of this analysis is tracking object within a single camera and between cameras' fields of vision. One of element needed to re-identify the single real object besides object's visual features and spatio-temporal dependencies between cameras is a behaviour model. The flow graph after some modifications, is a suitable data structure, which concept is based on the rough set theory, to contained as a behaviour model in it. Additionally, the flow graph can be used to predict the future movement of given object. In this paper a survey of authors research works related to employing flowgraphs in video surveillance systems is contained. The flow graph creation based on the paths of objects inside supervised area will presented. Moreover, a method of building a probability tree on the basis of the flow graph and a method for adapting the flowgraph to the changing topology of the camera network are also discussed.**

## I. Introduction

THE video surveillance systems have become common in public places and provided new possibilities (as well as challenges) in fields like security, crime prevention and automated video data processing. One of the main problems related to increasing number of cameras is that cameras' Fields of Vision (FOVs) do not overlap. In other words, there are locations which are not observed by any of the cameras. Thus the method for tracking object in a such adverse environment is needed. Therefore, this issue formed the basis of the authors' research work presented in this paper.

Tracking objects in a single camera is based on visual features of a moving object which differ from a background of a video image [1], [2]. Unfortunately such an approach is not suitable for the posed problem of re-identification of the same object in two different cameras. Therefore, some additional information related to statistical data analysis need to be obtained. There are two more types of premises (except of a comparison of visual features) which can be used to track objects more efficiently, that are:

- time of transition between given pair of the cameras
- probability that object will pass between a given pair of the cameras.

The first type of premise can be presented in the form of the weighted directed graph called topology graph. Edges of the topology graph determine physical possibility of transitions between cameras and describe time of these transitions. The description of transition time can be in form of:

- a single value (like average time of transition)
- a probability destiny function (e.g. Gaussian)
- a model (approximation) of time transition time (e.g. with Gaussian Mixture Model)

The second type of the premise (that is behaviour model) can be also described with a weighted directed graph but in this case it is also acyclic graph called flow graph. The idea of flow graphs was introduced by Pawlak and is based on the rough set theory [3], [4]. This paper is focused on a presentation of methods related to behaviour modelling with the Pawlak's flow graphs based on the data from video surveillance system. Utilization of the presented modified flow graph corresponds to tracking objects between cameras with non-overlapping fields of vision (FOVs). A general aim of this paper is presentation of the survey of the authors' works on the implementation of the Pawlak's flowgraph (as behaviour model) in the video surveillance systems. Certain paths of objects through observed area are more frequent than others and some transitions are more probable. The behaviour model can be considered as a container for knowledge about these patterns. In order to perform tracking object between cameras, three types of premises can be used to re-identify a single real object: visual features, time of transition and probability of choosing particular transition. The last premise is contained in the behaviour model (in the flow graph). The idea of Pawlak's flow graph is consistent with the rough set theory and Bayes' theorem. Description and definition of the flow graph can be found in the literature [5], [6], [7]. The flowgraph can be utilized in context of data mining and decision tree building [4]. Additionally, the flowgraph idea can be employed in processing of musical meta data [8]. Extensions and modifications of the flowgraph were also introduced [9], [10], [11].

The paper begins with a short presentation of the idea of flowgraphs with accordance to the rough set theory (see Sec. II). Next, in Sec. III modifications and extensions are introduced that were needed to apply, in order to use them with metadata obtained from the analysis of video data from surveillance system. Sec. IV presents an application of the extended flow graph in surveillance system and describes consecutive steps of authors' research work. The paper ends with summary and conclusion in Sec. V

## II. Pawlak's flow graphs and Rough Set Theory

In order to start a reflection on the flow graphs, some terms related to the rough set theory have to be presented according to literature [3]. Thus, a data set used in rough set theory is called an information system. The set of attributes denoted as $A$ must considered. Each attribute $a \in A$ may have values from a certain set $V_a$ (called the domain). If two disjoint subsets of attributes (called conditions $C$ and decisions $D$) are distinguished in the information system, then such a system becomes a decision system is denoted as:

$$S = (U, C, D),\ C \sqcup D \tag{1}$$

where $U$ is called the universe, $C$ is a set of condition attributes and $D$ is a set of decision attributes.

Based on the Pawlak publication [12] definition of flowgraph and their properties will be presented below. The flowgraphs are actually a kind of data structure suitable for containing a distribution of information flow and to present statistical features of objects from the mentioned universe $U$. Such an approach enables a new possibility of statistical data analysis belonging to intelligent methods.

A flow graph can be considered as a directed acyclic graph:

$$G = (N, E, \varphi) \tag{2}$$

where $N$ is a set of nodes, $E$ is a set of edges ($E \subseteq N \times N$) and $\varphi : E \to (R^+ \cup \{0\})$ is a flow function. Moreover, the idea of flow graph assumes the following notations and terminology:

- $(x, y)$ determines an edge with a node $x$ as an input and a node $y$ as an output, the edge $(x, y)$ must be contained in the set $E$;
- $I(x)$ is the set of all inputs of node $x$ and $O(x)$ is the set of all outputs of node $x$, while $x \in N$;
- also output and input of the whole flowgraph $G$ can be denoted as $I(G) = \{x \in N : I(x) = \emptyset\}$ and $O(G) = \{x \in N : O(x) = \emptyset\}$;
- input and output nodes are called external nodes and the rest of nodes are internal;
- $\varphi(x, y)$ is called throughflow from $x$ to $y$ which fulfils condition $\varphi(x, y) \neq 0$ for each edge $(x, y)$ in the set $E$.

Thus, for each node $x$ of a flow graph $G$, the inflow can be determined as:

$$\varphi_+(x) = \sum_{y \in I(x)} \varphi(y, x) \tag{3}$$

and the outflow can be defined as:

$$\varphi_-(x) = \sum_{y \in O(x)} \varphi(x, y) \tag{4}$$

In a similar way input and output of the whole flow graph can be formulated as:

$$\varphi_+(G) = \sum_{x \in I(G)} \varphi_-(x) \tag{5}$$

$$\varphi_-(G) = \sum_{x \in O(G)} \varphi_+(x) \tag{6}$$

Each internal node $x$ fulfil the condition:

$$\varphi_+(x) = \varphi_-(x) = \varphi(x) \tag{7}$$

where $\varphi(x)$ is called a throughflow of node $x$. For the whole flow graph $G$ the following formula is true:

$$\varphi_+(G) = \varphi_-(G) = \varphi(G) \tag{8}$$

where $\varphi(G)$ is a throughflow of the whole flow graph $G$. Hence, considering these assumptions all flows in the graph $G$ can be normalized with the value of $\varphi(G)$ as is presented in the formulae:

$$\sigma(x, y) = \frac{\varphi(x, y)}{\varphi(G)},\ \ 0 \leq \sigma(x, y) \leq 1 \tag{9}$$

$$\sigma(x) = \frac{\varphi(x)}{\varphi(G)},\ \ 0 \leq \sigma(x) \leq 1 \tag{10}$$

The value of $\sigma(x, y)$ is called the strength of edge $(x, y)$ and the value of $\sigma(x)$ is called the strength of node $x$.

Above defined normalized flows in the flowgraph allow for obtaining relative parameters assigned to the edges which are called certainty factor:

$$cer(x, y) = \frac{\sigma(x, y)}{\sigma(x)},\ \ \sigma(x) \neq 0 \tag{11}$$

and coverage factor:

$$cov(x, y) = \frac{\sigma(x, y)}{\sigma(y)},\ \ \sigma(y) \neq 0 \tag{12}$$

## III. Using flowgraph for behaviour modelling

In order to create behaviour model based on the flow graph a video surveillance system will be considered as in the previous paper [13]. Thus, a set of locations related to particular cameras is distributed on a certain area which can be described with the formula:

$$C = \{c_1, ..., c_N\} \tag{13}$$

where $c_i$ is camera with $i$ index and $N$ is a number of the cameras ($1 \leq i \leq N$). As it was mentioned a network of the cameras can be presented with the topology graph on which spatial dependencies between cameras are described (see Fig. 1)

Object which moved through the observed area creates a path which contains consecutive locations visited by the object. This path can be defined with the following formula:

$$p = \{(c_{id_1}, 1), ..., (c_{id_L}, L)\} \tag{14}$$

where $c_{id_1}$ and $c_{id_L}$ corresponds to camera which is visited by the object at the entrance to the observed area and at the exit from this area, respectively; numers from $id_1$ to $id_L$ define consecutive values of the index $i$ from Eq. 13. Thus, the set of paths can be described as presented by the following formula:

$$P = \{p1, ..., p_M\} \tag{15}$$

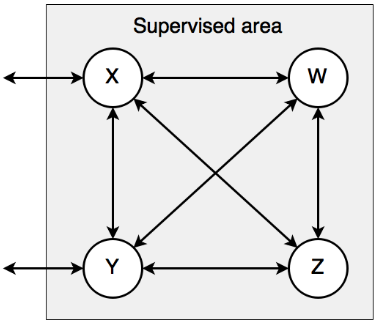


Fig. 1: The graph presenting topology of video surveillance system. According to Eq. 13: $C = \{w, x, y, z\}$

where $M$ is the number of paths in the set. Having the input data prepared in this way, next steps of creating the behaviour model can be carried out. However, in the beginning, the attributes and the domains must be reconsidered. Hence, the attribute used in the flow graph contains two parts:

- the index (number) that describes an order of this element in its path,
- the label of camera in which an object appeared.

For example, the attribute $X_1$ means that the given object was observed for the first time (on entrance to the observed area) in the camera 'X'. The consecutive domains are determined with the index mentioned above. The example flow graph created on the basis of the set of paths is presented in Fig. 2. Moreover, the certainty and coverage factors are also used in a specific way. In order to realize above, also parameters of the flowgraph need to be redefined as the following formulae show:

$$\begin{aligned} \sigma\left(x_i, y_{i+1}\right) &= \tfrac{\varphi(x_i, y_{i+1})}{\varphi(G)} \\ \sigma\left(x_i\right) &= \tfrac{\varphi(x_i)}{\varphi(G)} \\ cer\left(x_i, y_{i+1}\right) &= \tfrac{\sigma(x_i, y_{i+1})}{\sigma(x_i)} \\ cov\left(x_{j-1}, y_j\right) &= \tfrac{\sigma(x_{j-1}, y_j)}{\sigma(y_j)} \end{aligned} \tag{16}$$

where $\sigma(x_i, y_{i+1})$ defines the rate of objects passing from the camera $x$ in the step $i$ of the path to the camera $y$ in the next step $i + 1$ of the path, $\varphi(x_i, y_{i+1})$ determines the number of paths (in the set of path) which contain a transition from step $x_i$ to step $y_{i+1}$, the total number of paths taken into consideration while building the flow graph is denoted as $\varphi(G)$, and $\varphi(x_i)$ is the number of paths in the set of paths which contains the step (flow graph's node) $x_i$. Also the values of the certainty an coverage change their meanings. The certainty ($cer$) estimates the conditional probability that the object which left the camera $x$ in the step $i$ of its path will appear in the camera $y$ in the consecutive step $i + 1$, whereas the coverage ($cov$) determines estimation of the conditional probability that an object which appears in the camera $y$ in the step $j$ of the path was seen before, in the camera $x$ in the previous step $j - 1$ of the path. The certainty $cer$ is used to predict future movements of the object whereas the coverage $cov$ is useful in re-identification method. The $cov$ is utilized during the decision-making related to identification of the single object observed in two different cameras.

## IV. Flowgraphs in surveillance systems

The essential use case of the flow graph in the surveillance is related to prediction of object movements. The certainty factor estimates the probability of the future location that the object will visit, based on its previous route through the supervised area. Because of data concerning flows of objects through the observed locations, which is contained within the flow graph, probabilities of more than one location ahead can be predicted. As a result a probability tree is obtained. The formula which allows for creation of the probability tree (according to [14]) is shown in Eq. 17:

$$cer\left[x_{root}, \ldots, z_{end}\right] = \prod_{i=root}^{i=end} cer\left(x_i, x_{i+1}\right) \tag{17}$$

where the root of probability tree is denoted as $x_{root}$, the probability of the path from the vertex $x_{root}$ to $z_{end}$ is calculated as a product of probabilities of subsequent steps in the given path. An example probability tree created on the basis of the flow graph presented in Fig. 2 is shown in Fig. 3. The probability tree is created on the basis of the particular instance of the flow graph. It presents possible future transitions of the object observed in the certain camera in determined step of the object path.

Another important issue is the changing environment which is observed with cameras. It causes changes in the topology of the camera network. Some transitions may become physically impossible or new transitions appear. In such fluctuating conditions the created behaviour model may quickly become out of date and it will contain incorrect transition probability estimates. The flow graph is a quite slowly updating structure so a dedicated method for speed up adaptation to new conditions is needed. In order to solve this problem, an adaptation method employing some additional modifications, according to [15] is necessary:

- each path (obtained from video surveillance system) before adding to the set of paths and being used to build a recent behaviour model are weighted by so called importance factor;
- the importance factor is based on probability of the occurrence of the same path in the past;
- two instances of flowgraph are created: the first called $core\_$ and the second called $recent\_$
- measures of distance between two flow graphs also have to be introduced;
- values of two thresholds need to be determined: the first one called $learningThershold\_$ is created in order to enable making decision that flow graph $recent\_$ is a proper model of unchanging object behaviour, and the second one called $adaptiveThreshold\_$ is used to determine a moment in which adaptation method must be

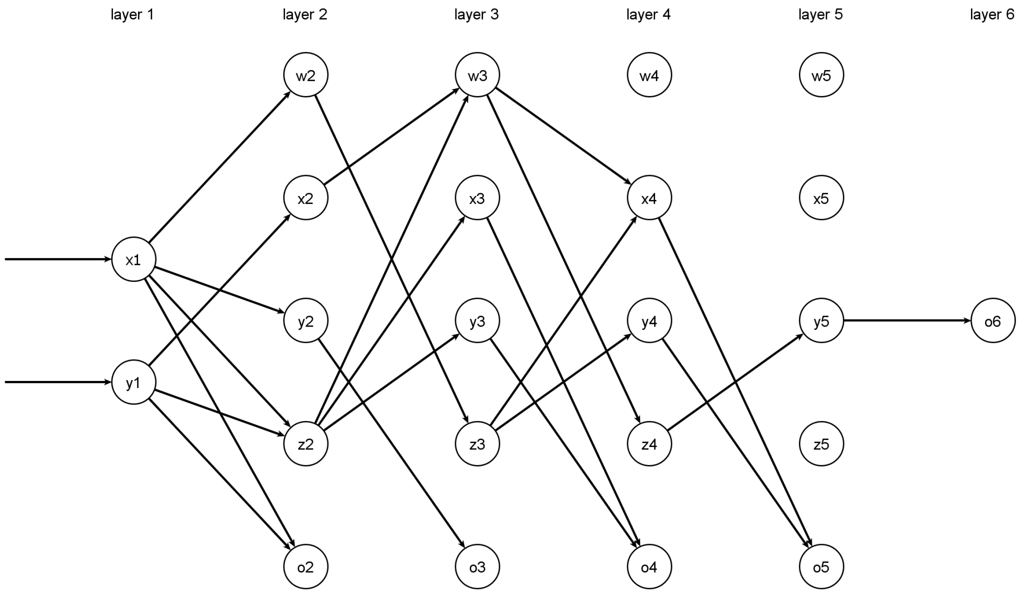


Fig. 2: The example flow graph obtained from the set of paths

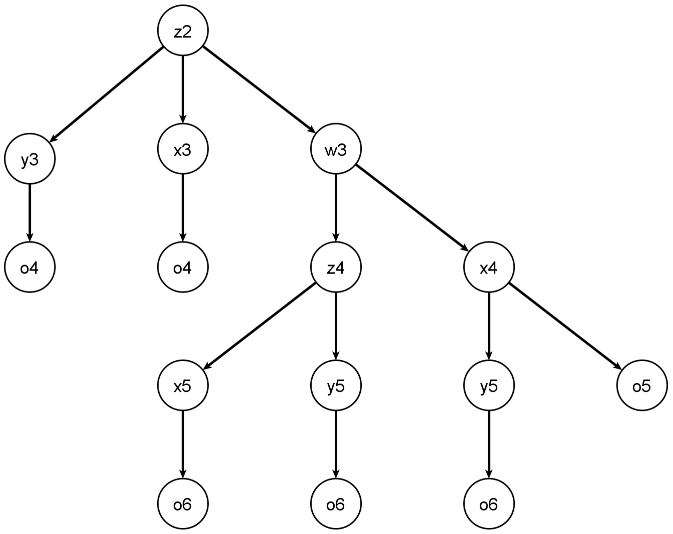


Fig. 3: Probability tree which begins in node $Z_2$ (what means that $x_{root} = Z_2$)

performed (when distance between $core\_$ and $recent\_$ is too large).

The adaptation method requires that some paths will be considered as more important ones than single paths. Moreover, a measure of importance of the path $imp$ needs to be introduced. In case of normal adding path to the flow graph, the importance factor $imp$ is equal to 1, but in the adaptation process $imp$ can be greater than 1. Hence, a weighting of each path must be performed and a weighted set of paths $w\hat{P}$ needs to be introduced, as follows:

$$\widehat{wP} = \{w\hat{p}_i\} \quad , \quad w\hat{p}_i = \langle \hat{p}_i, imp_i \rangle \qquad (18)$$

where $w\hat{p}_i$ is weighted path that contains path defined previously by Eq. 14 and $imp_i$ is importance factor of path $\hat{p}_i$.

The set of path $P$ is extended by probability of occurrence of a particular path in the past, upon fulfilling the condition:

$$Pr\left(\underline{Path} = p_i\right) = \frac{\|p_i\|}{\|P\|} \qquad (19)$$

where $\|p_i\|$ is number of instances of particular path in the set of paths $P$ and $\|P\|$ in number of all paths in set $P$.

The difference between two flowgraphs needs to be determined with distance metrics. The first metric is based on the average absolute deviation in certainty factor assigned to all edges of the flowgraph. This metric is called also conformity. In case that we have two flowgraphs denoted $A$ and $B$, it holds:

$$D = \frac{\sum\limits_{Edges_A} \left|cer_B\left(x,y\right) - cer_A\left(x,y\right)\right|}{\left\|Edges_A\right\|} \tag{20}$$

where $Edges_A$ is a set of edges in flowgraph $A$, $cer_A$ and $cer_B$ are certainty factors from flowgraph A and flowgraph B, respectively, $\|Edges\|$ is the number of edges in the flowgraph A.

The second metric utilizes probability distributions defined by Eq. 19. The metric is based on a coincidence index. Changes in the flowgraph are also related to changes in the probability distribution of paths $Pr\left(\underline{Path}\right)$. One of probability distributions must be determined as a reference (in this case it is $Pr_A$), whereas the second probability distribution is the modified one ($Pr_B$). The probability distribution $Pr_B$ comes from the flowgraph B that used more paths as input than the flowgraph A. The mentioned modification is implied as a possibility of appearance paths which were not present in the input set of the flowgraph A. In order to solve this problem a modification must be made of $Pr_B$. All instances of paths that are in probability distribution $Pr_A$ do not appear in probability distribution $Pr_B$ must be removed from the probability distribution $Pr_B$. Next, a renormalization of the probability distribution $Pr_B$ is made. The modified probability distribution $\hat{Pr}_B$ prepared in such a way can be used in the formula 21 as follows:

$$CI = \frac{\sum\limits_{\underline{Path}_A} \left[Pr_A\left(\underline{Path_A} = p_i\right) \cdot \hat{Pr}_B\left(\underline{Path_B} = p_i\right)\right]}{\left\|\underline{Path_A}\right\|} \tag{21}$$

where $\left\|\underline{Path_A}\right\|$ determines how many different instances of path is in the set of paths $P$ defined in Eq. 14.

The adaptation method operates in two phases. The first is creation of two new flow graphs ($core\_$ and $recent\_$) and adding paths to $recent\_$ in groups (of i.e. one hundred paths). After adding a group of paths, the distance between $core\_$ and $recent\_$ is calculated (using formula Eq. 20). If the distance is larger than $learningThershold\_$, then $recent\_$ is copied to the $core\_$ flowgraph and a next group of paths is added to the $recent\_$, otherwise the $core\_$ is considered as a proper behaviour model and the adaptation algorithm passes to the second phase. In this phase paths are still adding to the $recent\_$ in groups, but there is no copying of the $recent\_$ flowgraph to the $core\_$. After adding the group of paths the distance between the flow graph $core\_$ and the flow graph $recent\_$ is calculated (see Eq. 20) . If this distance is lower than $adaptiveThreshold\_$, the next group of paths is added to the $recent\_$ flowgraph, otherwise the importance weights (see Eq. 21) is calculated for the last group of paths and these paths are added to the $core\_$ flowgraph with appropriate weights. Next, consecutive groups of paths are added to the $recent\_$ flowgraph. The flowchart of this algorithm is presented in the Fig. 4.

In order to prove this concept, simulations were performed. The results of the simulations are presented in Fig. 5,

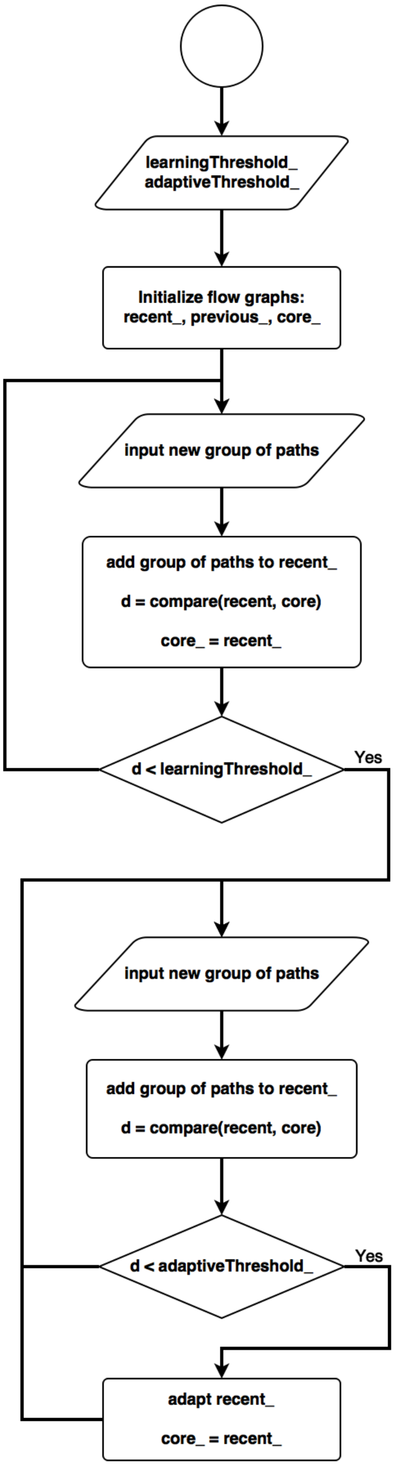
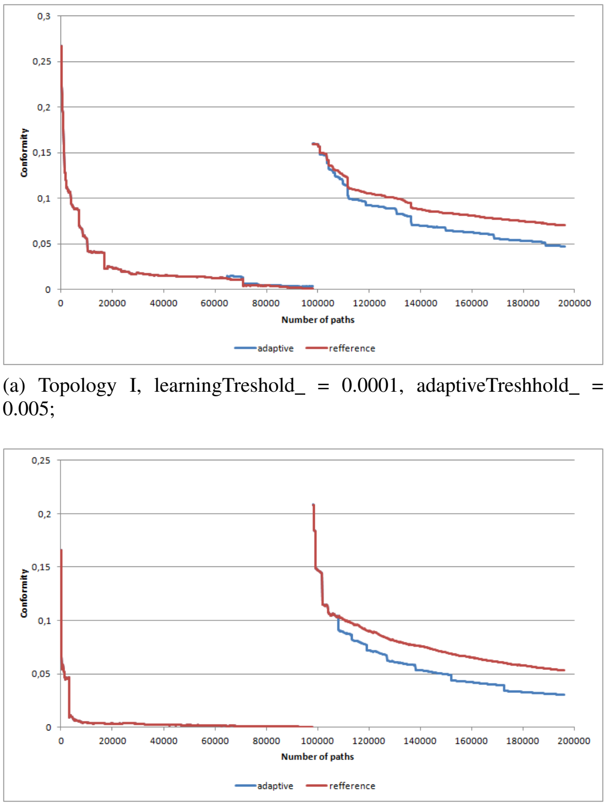


Fig. 4: Flowchart of the used adaptation algorithm

The simulation was carried out in the following way:

1) on the basis of real set of paths, which was quite small (about 1000 paths), a large set of path was generated (about 100 thousands paths),
2) the $real\_small\_$ flowgraph is built on the basis of this



(a) Topology I, learningTreshold_ = 0.0001, adaptiveTreshhold_ = 0.005;

(b) Topology II, learningTreshold_ = 0.0001, adaptiveTreshhold_ = 0.005;

Fig. 5: Result of preformed simulations for different topologies of video surveillance systems

small set of path,

3) building a flow graph enforced with adaptation method is performed with the large (generated) set of paths – adding paths in groups of 100,
4) the small set of path is modified in a way which simulates changing the topology of the cameras network,
5) the new version of $real_small\_$ flowgraph is built on the basis of this small modified set of path,
6) the second large set of paths is generated on the basis of this modified small set of paths,
7) adding the paths form the large set to the recent flowgraph occurs.

The charts from Fig. 5 present the distance (see Eq. 20) between the $real\_small\_$ flow graph and the $core\_$ flow graph (when adaptation method is in use). In order to show the difference, the $reference\_$ flow graph is added. This flow graph is built without any adaptation method. The drastic change in the center of the chart is related to the modification of the small set of the path (simulating change in the topology graph). In order to obtain real input data (a small set of object paths) the setup of 6 cameras, denoted as *Topology I*, was used. The *Topology II* was a group of 11 cameras. The $real\_small\_$ flow graph for both sets of camera was created basing on the analysis of 1,5 hour of video material.

## V. Conclusion and Future work

The flow graph is a suitable data structure to contain a behaviour model. It is prone to extensions, modifications and adaptation to various types of problems because of its transparency and simplicity. In case of video surveillance systems the flow graph is a container for knowledge concerning object behaviour which is easily to obtain and fast in use. The certainty and coverage factors can be clearly explained and easily applied to problems of object movement prediction (in correspondence to $cer$) or object re-identification (in correspondence to $cov$). Additionally, problem of changing conditions in video surveillance systems also can be also managed using the adaptation method presented above. This adaptation method allows for obtaining better conformity of the flow graph in case of modification of the topology of camera network. Moreover, the flow graph adapts faster to new conditions. The future works related to the flow graphs are concentrated on the application of them in tracking object in one camera in case when the object is obscured by another object or some other obstacles.

## Acknowledgment

This work has been partially funded by the Artemis JU and by the Polish National Centre for Research and Development (NCBR) as part of the COPCAMS project (http://copcams.eu) under GA number 332913.

## References

[1] A. Czyżewski, G. Szwoch, P. Dalka, S. P., C. A., E. D., M. T., L. K., K. L., and W. J., “Multi-stage video analysis framework,” in *Video Surveillance*, L. Weiyao, Ed. Intech, 2011, ch. 9, pp. 145–171. [Online]. Available: http://dx.doi.org/10.5772/16088

[2] A. Czyżewski and P. Dalka, “Moving object detection and tracking for the purpose of multimodal surveillance system in urban areas,” in *New Directions in Intelligent Interactive Multimedia*, ser. Studies in Computational Intelligence, G. Tsihrintzis, M. Virvou, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, 2008, vol. 142, pp. 75–84. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-68127-4_8

[3] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Norwell, MA, USA: Kluwer Academic Publishers, 1992. [Online]. Available: http://dx.doi.org/10.1016/s0967-0661(96)90021-0

[4] ——, “Transactions on rough sets iii,” J. F. Peters and A. Skowron, Eds. Berlin, Heidelberg: Springer-Verlag, 2005, ch. Flow Graphs and Data Mining, pp. 1–36. [Online]. Available: http://dx.doi.org/10.1007/11427834_1

[5] ——, “Decision algorithms, bayes theorem and flow graphs,” in *Neural Networks and Soft Computing*, ser. Advances in Soft Computing, L. Rutkowski and J. Kacprzyk, Eds. Physica-Verlag HD, 2003, vol. 19, pp. 18–24. [Online]. Available: http://dx.doi.org/10.1007/978-3-7908-1902-1_3

[6] ——, “Decision algorithms and flow graphs: A rough set approach.”

[7] ——, “Rough sets and flow graphs,” in *RSFDGrC (1)*, ser. Lecture Notes in Computer Science, D. Slezak, G. Wang, M. S. Szczuka, I. Dntsch, and Y. Yao, Eds., vol. 3641. Springer, 2005, pp. 1–11. [Online]. Available: http://dx.doi.org/10.1007/11548669_1

[8] B. Kostek and A. Czyzewski, “Processing of musical metadata employing pawlaks flow graphs,” in *Transactions on Rough Sets I*, ser. Lecture Notes in Computer Science, J. Peters, A. Skowron, J. Grzymala-Busse, B. Kostek, R. Swiniarski, and M. Szczuka, Eds. Springer Berlin Heidelberg, 2004, vol. 3100, pp. 279–298.

[9] P. Pattaraintakorn, "Entropy measures of flow graphs with applications to decision trees," in *Rough Sets and Knowledge Technology*, ser. Lecture Notes in Computer Science, P. Wen, Y. Li, L. Polkowski, Y. Yao, S. Tsumoto, and G. Wang, Eds. Springer Berlin Heidelberg, 2009, vol. 5589, pp. 618–625. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-02962-2_78

[10] J. Peters, D. Chitcharoen, and S. Ramanna, "Reasoning with near set-based digital image flow graphs," in *Multi-disciplinary Trends in Artificial Intelligence*, ser. Lecture Notes in Computer Science, S. Ramanna, P. Lingras, C. Sombattheera, and A. Krishna, Eds. Springer Berlin Heidelberg, 2013, vol. 8271, pp. 199–210. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-44949-9_19

[11] Z. Suraj and K. Pancerz, "Flow graphs as a tool for mining prediction rules of changes of components in temporal information systems," in *Rough Sets and Knowledge Technology*, ser. Lecture Notes in Computer Science, J. Yao, P. Lingras, W.-Z. Wu, M. Szczuka, N. Cercone, and D. lzak, Eds. Springer Berlin Heidelberg, 2007, vol. 4481, pp. 468–475. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72458-2_58

[12] Z. Pawlak, "Rough sets, decision algorithms and bayes theorem," *European Journal of Operational Research*, vol. 136, no. 1, pp. 181–189, 2002. [Online]. Available: http://dx.doi.org/10.1016/s0377-2217(01)00029-7

[13] A. Czyżewski and K. Lisowski, "Employing flowgraphs for forward route reconstruction in video surveillance system," *Journal of Intelligent Information Systems*, vol. 43, no. 3, pp. 521–535, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10844-013-0253-8

[14] Z. Pawlak, "Rough sets and flow graphs," in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, ser. Lecture Notes in Computer Science, D. Slezak, G. Wang, M. Szczuka, I. Duntsch, and Y. Yao, Eds. Springer Berlin Heidelberg, 2005, vol. 3641, pp. 1–11. [Online]. Available: http://dx.doi.org/10.1007/11548669_1

[15] A. Czyżewski and K. Lisowski, "Adaptive method of adjusting flowgraph for route reconstruction in video surveillance systems," *Fundam. Inf.*, vol. 127, no. 1-4, pp. 561–576, Jan. 2013. [Online]. Available: http://dx.doi.org/10.3233/FI-2013-927

DOI: 10.15439/2015F409
ACSIS, Vol. 5

# Representing Parametric Concepts with Situation Theory

Roussanka Loukanova
Department of Mathematics, Stockholm University
Email: rloukanova@gmail.com

***Abstract*—We use higher-order, type-theoretic Situation Theory to model semantic concepts as situation-theoretic objects consisting of parametric information. Situation Theory contributes by representing concepts as classes of parametric objects, in a computational way. We use concepts that are often expressed by human language in taxonomy classifications, as a demonstration of the situation theoretic-approach to model parametric information in abstract concepts.**

## I. Introduction

THE IDEAS of Situation Theory were originally introduced by Barwise [1], and then by Barwise and Perry [2], for modeling information in nature. The work emerged from decades of efforts by varieties of model-theoretic approaches for adequate computational semantics of human language and cognitive science. In search for adequate semantics of human language, with his extensive work in mathematics, model-theory, and admissible sets, Jon Barwise soon realized that semantic objects for human language are special cases of objects in a more general theory of meaning and information. Since then, Situation Theory has been under development as a powerful, highly expressive theory of finely-grained information that is partial, underspecified, and situational. Semantics of languages is one of the prominent applications of Situation Theory, know as Situation Semantics.

Our intensive efforts on modeling semantic information and concepts are currently in several, concurrent directions, for intelligent applications to information and language processing. In nature, information typically is partial, parametric, and dependent on situations, in most of its components. For adequate modeling of semantic objects, we need to represent these natural features of information and languages. Applications to many contemporary technologies, which are related to data science, information, and language processing, require models of information and information processing. These models need to reflect information flow in nature and, in the same time, to be computational.

On the side of the mathematical foundations of Situation Theory, Aczel non-well-founded set theory, see Aczel [3], has proved to be the most suitable set theory for modeling classes of situation-theoretic objects that are proper classes, i.e., which are not sets in the classic Zermelo-Fraenkel set theory ZFC, while they are non-well-founded sets in Aczel set theory. Aczel non-well-founded set theory is an axiomatic system consisting of the ZFC axioms, except the Axiom of Foundation, which is replaced with Aczel Anti-Foundation Axiom (AFA), see also Rathjen [4]. Situation Theory, when based on Aczel non-well-founded sets, models circular information and self-reference, including for concepts. It can model potentially large classes of situation-theoretic objects, which theoretically are proper non-well-founded sets. What is significant, for practical, intelligent applications, is that situation-theoretic objects, even when properly non-well-founded by the AFA, have finite, not necessarily large representations, e.g., visualized as cyclic graphs. Large objects can be limited for practical applications, e.g., by restrictions from specific domains of applications.

Situation Theory has been under development as a theory of the inherently relational and situational nature of information, in general, not only of linguistic meanings, by diverging from the traditional possible-world theories of semantics with type-theoretic settings, in particular from Montague's IL (see Montague [5]). Detailed discussions and motivations of the situation-theoretic objects, such as situation types similar to the ones introduced in this article, are given in Barwise and Perry [2]. For an informal introduction to Situation Theory and Situation Semantics for human language, with examples and intuitions, see Devlin [6]. Note that the typed situation-theoretic objects that we use in this paper extend the ideas of situated objects in the early works on Situation Theory, and in addition, are strictly defined objects of mathematical structures. Formal introduction, in the lines of our work here, is given in Loukanova [7]–[10]. These works include examples from human language, while Loukanova [11], [12] provides syntax-semantics constructions of human language expressions, by using phrase-structure syntax, which is the precursor and theoretic backbone of parsers in currently prevailing computational syntax.

One of the distinguished applications of Situation Theory has been Situation Semantics for semantic representations of human language in computational grammars. Head-driven Phrase Structure Grammar (HPSG) is one of the first practical grammar frameworks, based on formal syntax of human language by using typed, linguistic feature-value structures, see Pollard and Sag [13], [14], and Sag et al. [15]. Originally, HPSG was introduced by the ideas of Situation Theory for distribution of partial information throughout grammatical representations, via typed feature-value structures. Various, partly specified feature-structures can be combined according to grammar principles and constraints, by unification and expending them with new information. From start, HPSG came with ambitions to use Situation Semantics for including

978-83-60810-66-8/$25.00©2015, IEEE

semantic representations in syntactic analyses. Current HPSG systems have been successfully realizing such semantic representations with a specialized language, Minimal Recursion Semantics (MRS), for handling scope ambiguities, see, e.g., Copestake et al. [16]. Loukanova [17] shows that the concept of minimal recursion in MRS has a functional formalization by the formal language of acyclic recursion introduced in Moschovakis [18]. By considering the relational character of the predicate symbols used in elementary predications in MRS, we see MRS as an implementation of a special case of a formal language for Situation Theory, in the lines of Loukanova [10], while more work on the relationship is necessary. The original Situation Semantics inspired other work in linguistics. E.g., it was used for semantic analysis of questions, see Ginzburg and Sag [19]. Lambalgen and Hamm [20] used concepts of Situation Semantics for semantics of tense and aspect, from cognitive perspective.

Situation Theory is an open area of theoretic development, with potentials for varieties of applications. While it has established classic applications to computational semantics, as briefly summarized above, both Situation Theory and Situation Semantics are largely open areas, in theory and applications. Currently, Situation Theory has new significance as a theory of heterogeneous information, along with the proliferation of interdisciplinary technologies and applications, especially in Artificial Intelligence and other areas that involve intelligent computation.

This paper is on a specific task of using situated information, with parametric objects, to represent hierarchically linked classes of parametric concepts. We employ situated types that support linking parametric objects with restrictions. We introduce primary restrictions over parametric objects as types associated with the argument roles of relations and types. These restrictions are called appropriateness conditions over argument roles. The argument roles (commonly known as argument slots) of relations and types can be filled up only by objects satisfying the respective appropriateness conditions.

The notational symbolism that we use to designate abstract objects of Situation Theory reminds of expressions of a formal language, but by these notations, we do not define a formal language and do not use any such formal language per se. I.e., Situation Theory is a higher-order, typed, mathematical structure. In this work, we use Situation Theory as a model theory of information, by a focus on specific abstract objects, without formal language. On the other side, a formal language for Situation Theory can provide many advantages. E.g., it is important to have a formal language for situation-theoretic analyses of human and artificial languages, via semantic representations by formal terms (which are usually called logic forms). Introducing a formal language for Situation Theory is the topic of other work, see Loukanova [10], [21], for development of formal languages of Situation Theory. In this work, our focus is on introducing semantic domains of situation-theoretic objects, i.e., complex types with parameters and complex, restricted sets of linked parameters. In the second part of the paper, we use situation-theoretic objects to represent parametric concepts, i.e., concepts as model-theoretic objects with rich informative structure, where information can be parametric.

*Note 1:* The situation-theoretic objects are often designated by multi-line expressions, i.e., spread over several lines, for lack of space in a single line, but also to visualize the structure of the objects. We have tried to follow traditional indentation as in programming, wherever possible, but primarily, we try to follow the convention that the arguments of a given relation or type are vertically aligned.

## II. A Brief Overview of Basic Situation-Theoretic Notions

This section introduces situation-theoretic notions and objects that are fundamental for representation of information and essential for the following sections of the paper. Informally, the informational pieces, called *infons*, are basic and complex objects that have structure carrying information about relations and objects filling the arguments of the relations, at time-space locations. Recursively, basic and complex infons are constructed by starting with primitive relations, argument roles, primitive individuals filling the argument roles of the relations, basic space-time locations, and positive or negative polarity. The polarity of an infon carries the information about whether or not the objects in the infon are in its relation.

### A. Primitive Individuals

A collection (typically, a set) $\mathcal{A}_{\text{IND}}$ is designated as the set of primitive *individuals* of Situation Theory:

$$\mathcal{A}_{\text{IND}} = \{a, b, c, \dots\} \tag{1}$$

The objects in $\mathcal{A}_{\text{IND}}$ are set-theoretic objects, not necessary atomic urelements, which are considered as primitives in Situation Theory. In various versions of Situation Theory, designated for specific applications, some of the individuals in $\mathcal{A}_{\text{IND}}$ may be parts of other individuals in $\mathcal{A}_{\text{IND}}$, and as such, can be in respective *part-of* relations.

### B. Space-time Locations

Situation Theory make a substantial use of a given class $\mathcal{A}_{\text{LOC}}$ of space-time points, periods, and regions units. Note that $\mathcal{A}_{\text{LOC}}$ can be a proper class, or a set, depending on the version of Situation Theory that one can select to use in applications.

$$\mathcal{A}_{\text{LOC}} = \{l, l_0, l_1, \dots\} \tag{2}$$

The collection $\mathcal{A}_{\text{LOC}}$ is endorsed with relations of time precedence $\prec$, time overlapping $\circ_t$, space overlapping $\circ_s$, space-time overlapping $\circ$, and space and time inclusions $\subseteq_t$, $\subseteq_s$, $\subseteq$. In some versions of Situation Theory, the space-tile locations can be given by complex objects, e.g., as pairs of two components, one for space locations (regions), and one for time points or periods.

### C. Primitive Relations

Significantly, Situation Theory has a collection (which can be a set in practical applications, or a proper class) $\mathcal{A}_{\text{REL}}$ of abstract, primitive objects that are the primitive, i.e., basic, relations: $\mathcal{A}_{\text{REL}} = \{r_0, r_1, \dots\}$ The elements of $\mathcal{A}_{\text{REL}}$ are abstract representatives of real or virtual relations. For example, some of them can be abstract representatives of real properties of objects and relations between objects, in reality, or in virtual models, which humans are attuned to distinguish perceptually in reality, or cognitively, i.e., conceptually.

In typical set-theoretic practice, relations between set-theoretic objects are represented as sets of ordered tuples of the objects being in the relevant relations. On the contrary, Situation Theory takes the relations in $\mathcal{A}_{\text{REL}}$ as primitive, first-class objects. I.e., the objects in $\mathcal{A}_{\text{REL}}$, are primitive objects that are not themselves sets of tuples of individuals being in those relations. Set-theoretically, the primitive relations in $\mathcal{A}_{\text{REL}}$, as well as the other primitive objects in Situation Theory, such as individuals, properties, relations, and types, can be taken as urelements of the meta-theoretic set theory. E.g.,

$$\mathcal{A}_{\text{REL}} = \{\, man, woman, dog, run, like, \dots \} \tag{3}$$

By introducing more complex situation-theoretic objects, it is possible to define the notion of the *extension* of a given relation $r$, in a given situation $s$ as the set of the tuples of objects being in the relation $r$ in $s$. For example, the extension of the relation $read$ in a given, specific situation $s$ and a space-time location $l$, is the set of all pairs $\langle a, b\rangle$ of objects, such that the primitive relation of reading holds between $a$ as the reader and $b$ as the object that is read, at the location $l$, in the situation $s$.

### D. Primitive Types

Situation Theory has a collection (a relatively small, finite set) of objects, which are called *primitive* or *basic* types, that represent our intuitions, cognitive concepts of types, and type classifications of objects in specific areas of applications:

$$B_{\text{TYPE}} = \{\text{IND}, \text{LOC}, \text{REL}, \text{TYPE}, \text{POL}, \text{PAR}, \tag{4a}$$
$$\text{ARoles}, \text{INFON}, \text{SIT}, \text{PROP}, \models\} \tag{4b}$$

where IND is the type of individuals; LOC: of space-time locations; REL: of relations (primitive and complex); TYPE: of types (primitive and complex); POL: of two polarity objects (e.g., represented by the natural numbers 0 and 1); PAR: of parameters; ARoles: of abstract argument roles (primitive and complex); INFON: of situation-theoretic objects that are basic or complex information units; SIT: of situations; PROP: of abstract objects that are propositions; $\models$ is a type called "supports". Some of these types will be explained later.

### E. Primitive Parameters — Indeterminates

Situation Theory has a collection (a set) of primitive parameters, for each of the basic types, e.g.:

$$\mathcal{P}_{\text{IND}} = \{\dot{a}, \dot{b}, \dot{c}, \dots\}, \qquad \mathcal{P}_{\text{LOC}} = \{\dot{l}_0, \dot{l}_1, \dots\}, \tag{5a}$$
$$\mathcal{P}_{\text{REL}} = \{\dot{r}_0, \dot{r}_1, \dots\}, \qquad \mathcal{P}_{\text{POL}} = \{\dot{i}_0, \dot{i}_1, \dots\}, \tag{5b}$$
$$\mathcal{P}_{\text{SIT}} = \{\dot{s}_0, \dot{s}_1, \dots\}, \qquad \dots \tag{5c}$$

We assume that, for every type $\theta$ : TYPE, there is potential availability of parameters of that type $\theta$, (6a).

$$\mathcal{P}_\theta \text{ is a class of parameters, for every } \theta : \text{TYPE} \tag{6a}$$
$$p \in \mathcal{P}_\theta \quad \text{iff} \quad p : \theta \text{ and } p : \text{PAR} \tag{6b}$$

Thus, theoretically, the classes of types and parameters can be proper classes, which are Aczel non-well founded sets, see Aczel [3]. Note that in applications, for many types $\theta$ : TYPE, it can be the case that $\mathcal{P}_\theta = \emptyset$. Practically, it would be useful, to add classes $\mathcal{P}_\theta$ not in advance, but depending on needs, and to add fresh, new parameters to them "on-the-go".

Sometimes, but not always, we use a notation originally introduced in Situation Theory, to denote parameters with dotted letters, as in (5a). Marking letters with dots is a visual distinction of parameters from other individuals and objects. However, we should stress that this paper is about modeling domains of Situation Theory, not about a formal language. Letters, characters, and expressions that we use are notational means of denoting objects in situational domains.

In this paper, we start with the idea of situation-theoretic parameters as representing very primitive concepts that are distinguished only by their types. Thus, $\mathcal{P}_{\text{IND}}$ is the class of the primitive concepts of individuals, e.g., $\dot{a}$, $\dot{b}$, etc., are concepts of individuals. $\mathcal{P}_{\text{LOC}}$ is the class of the primitive concepts of space-time locations, e.g., $\dot{l}_0$, $\dot{l}_1$, etc., are concepts of space-time locations. $\mathcal{P}_{\text{REL}}$ is the class of the primitive concepts of relations, where any element $r \in \mathcal{P}_{\text{REL}}$ is a concept of a relation. E.e., $blue$, as a unary relation, i.e., a property of objects, is the concept of an object being blue in color, in space-time. $give$ is the concept of a relation between three objects, one being an individual giving an object to another individual, which takes place in space-time. Typically, relations between objects and properties of objects happen in space-time. The class $\mathcal{P}_{\text{SIT}}$ consists of primitive parameters that represent abstract concepts of situations. We consider that the biological nervous systems, at least those of humans and other advanced living organisms, are attuned to recognize both abstract entities and specific instances of abstract entities. In particular, human brain has inner facilities to form and comprehend concepts for individuals, relations, space-time locations, and situations, as well as specific representatives, i.e., instantiations, of the abstract concepts.

*Notation 1:* For any given type $T$ (primitive or complex) and an object $\Theta$, we write $(T : \Theta)$ to designate the proposition that $\Theta$ is of type $T$, and $T : \Theta$ iff $\Theta$ is of type $T$. An alternative notation of can be used, i.e., $\Theta : T$, as in some type systems, such as the intensional logics of Montague and Gallin. We allow both notations depending on convenience, i.e., given a type $T$ (primitive or complex) and an object $\Theta$, we write

$$\begin{gathered}(T : \Theta) \quad \text{iff} \quad (\Theta : T)\\ \text{i.e., the proposition that } \Theta \text{ is of type } T\end{gathered} \tag{7a}$$
$$T : \Theta \quad \text{iff} \quad \Theta : T \quad \text{iff} \quad \Theta \text{ is of type } T \tag{7b}$$

The alternative notations in (7a)–(7b) can be used depending on the context, which makes clear the usage. Note that $(T : \Theta)$

and $(\Theta : T)$ in (7a) both designate the proposition that $\Theta$ is of type $T$, while the alternatives in (7b) designate the verified proposition, when factually $\Theta$ is of type $T$.

### F. Primitive Argument Roles

We assume a collection (a set) of primitive objects $\mathcal{BA}_{\text{ARoles}}$ designated as *primitive argument roles*, which is a sub-collection of the class of $complexargumentroles$:

$$\mathcal{BA}_{\text{ARoles}} = \{\rho_0, \dots, \rho_n, \dots\} \subset \mathcal{A}_{\text{ARoles}} \tag{8}$$

A set of argument roles is associated with each of the primitive relations, and each of the primitive types, by a function $\textsc{ArgR}$, with domain and range: $Dom(\textsc{ArgR}) = \mathcal{A}_{\text{REL}} \cup B_{\text{TYPE}}$, and $Range(\textsc{ArgR}) \subseteq \textsc{type} \times \mathcal{A}_{\text{ARoles}}$. Thus the argument roles of each type and each relation $X$ (basic or complex, recursively for the complex ones) are associated with corresponding types that restrict what objects can fill up the argument roles. I.e., every relation or type $X$, is associated with argument roles:

$$\textsc{ArgR}(X) = \{T_1 : arg_1, \dots, T_n : arg_n\}, \tag{9a}$$

$$\text{where } arg_i : \text{ARoles and } T_i : \textsc{type},\ i = 1, \dots, n, \text{ for some } n \geq 0 \tag{9b}$$

The types $T_i$ are called *appropriateness constraints* of the corresponding argument roles $arg_i$, $i = 1, \dots, n$, of the relation (type) $X$. Complex relations and types are associated with argument roles and corresponding appropriateness constraints, according to recursive definitions given in what follows, supplemented by examples.

In what follows, we assume that if an argument role $arg_i$ of a relation or type $X$, (9) is restricted by a type $T_i : arg_i$, this argument role can be filled up by a situation-theoretic object of type $T_i$, including by parameters, e.g., by using (6b).

### G. Basic Infons

Basic infons can be represented by specialized, marked tuples

$$\langle infon, \gamma, \theta, \tau, i \rangle \tag{10}$$

where $\gamma \in \mathcal{R}_{\text{REL}}$ is a relation (primitive or complex), $\textsc{loc} : \tau$, $\textsc{pol} : i$, and $\theta$ is a function, called the *argument-role filling of* $\gamma$, which fills up the argument roles $arg_1, \dots, arg_n$ $(n \geq 0)$ of $\gamma$ with objects $\xi_1, \dots, \xi_n$ of respective types $T_1, \dots, T_n$, i.e.:

$$\theta = \{\langle T_1 : arg_1, \xi_1 \rangle, \dots, \langle T_n : arg_n, \xi_n \rangle\} \tag{11}$$

for some situation-theoretic objects $\xi_1, \dots, \xi_n$ satisfying the corresponding appropriateness constraints of the argument roles of the relation $\gamma$.

*Notation 2:* The basic infons (10), as well as some of the complex ones are denoted by (12):

$$\ll \gamma, \theta, \tau, i \gg \tag{12}$$

*Notation 3:* When the types of the argument roles are agreed, i.e., understood by the context, the function filling the argument roles is denoted by (13).

$$\theta = \{\langle arg_1, \xi_1 \rangle, \dots, \langle arg_n, \xi_n \rangle\} \tag{13}$$

### H. Complex Infons

Complex infons for representation of conjunctive and disjunctive information are formed by the operators conjunction and disjunction. In some earlier versions of Situation Theory, the operators conjunction and disjunction in the infon constructions were taken to be primitive relations between infons, for which locations are irrelevant.

*Notation 4:* For sake of space, by assuming that the order of the argument roles is agreed to avoid confusion, we shall adopt the traditional linear notations of the basic infons, with or without the type constraints over the argument roles, as in (14a)–(14d).

$$\ll \gamma, T_1 : arg_1 : \xi_1, \dots, T_n : arg_n : \xi_n, \textsc{loc} : \tau; i \gg \tag{14a}$$

$$\ll \gamma, arg_1 : \xi_1, \dots, arg_n : \xi_n, \textsc{loc} : \tau; i \gg \tag{14b}$$

$$\ll \gamma, T_1 : \xi_1, \dots, T_n : \xi_n, \tau; i \gg \tag{14c}$$

$$\ll \gamma, \xi_1, \dots, \xi_n, \tau; i \gg \tag{14d}$$

We denote the class of all infons, basic or complex, by $\mathcal{I}_{\text{INFON}}$.

## III. Basic Semantic Concepts

In this section, we introduce the idea of representing basic semantic concepts as situation-theoretic objects with parametric components. We use prototypical examples of semantic relations and parametric information peaces.

*Example 3.1:* Assume that $read_2 \in \mathcal{A}_{\text{REL}}$, i.e., $read_2$ is a primitive situation-theoretic object of type $\textsc{rel}$, i.e., by (7b), $read_2 : \textsc{rel}$. Assume also that the relation $read_2$ has two argument roles as in (15):

$$\textsc{ArgR}(read_2) = \{\textsc{ind} : reader, \textsc{ind} : readed\} \text{ for } reader, readed \in \mathcal{A}_{\text{ARoles}} \tag{15}$$

Then, we have the following infons in (16a)–(16d).

$$\ll read_2, \textsc{ind} : reader : a, \textsc{ind} : readed : b, \textsc{loc} : l; 1 \gg \tag{16a}$$

$$\ll read_2, \textsc{ind} : reader : \dot{a}, \textsc{ind} : readed : \dot{b}, \textsc{loc} : \dot{l}; 1 \gg \tag{16b}$$

$$\ll read_2, \textsc{ind} : reader : a, \textsc{ind} : readed : \dot{b}, \textsc{loc} : \dot{l}; 1 \gg \tag{16c}$$

$$\ll read_2, a, b, l; \dot{p} \gg \tag{16d}$$

Note that we use the “misspelled” notation $readed$ for the semantic argument role of the relation $read_2$, which is to be filled by the object that is being read. I.e., $readed \in \mathcal{A}_{\text{ARoles}}$ is an abstract argument role denoted by this “misspelled” variant of the past participle of the verb “read”. This notation is by a trend in the early versions of Situation Semantics, by which, the argument role for the actor of an activity, usually denoted by a verb, is represented by using the suffix “er”, and the argument role for the object acted upon by using the suffix “ed” or “en”. Thus, some argument roles can have “misspelled” notations. How the argument roles are denoted is a matter of agreement settings. We have chosen here that trend, to avoid indexing the argument roles with natural numbers, i.e., $arg_1, \dots, arg_n$, which carries connotations that the argument roles have been linearly ordered, which is not always the case. We would like to stress that, in general, there is no intrinsic order over the argument roles of relations and types, except in specific cases and for notational needs.

In (16a)–(16d), $a, b \in \mathcal{A}_{\text{IND}}$ are individuals, $l \in \mathcal{A}_{\text{LOC}}$ is a location, while $\dot{a}, \dot{b} \in \mathcal{P}_{\text{IND}}$ are parameters for individuals, $\dot{l} \in \mathcal{P}_{\text{LOC}}$ is a location parameter, and $\dot{p} \in \mathcal{P}_{\text{POL}}$ is a parameter for either of the two polarities $\{0, 1\}$. E.g., the unknown individual, which fills up the argument role of the material that is being read, is represented by a semantic parameter $\dot{b}$ that is restricted to be of type IND, by the constraint over the argument role that $\dot{b}$ fills up, i.e., $\text{IND} : \mathit{readed} : \dot{b}$. This constraint allows $\dot{b}$ to be of both types IND and PAR, i.e., $\text{IND} : \dot{b}$ and $\text{PAR} : \dot{b}$, by using (6b). Similarly, for the constraints over $\dot{a}, \dot{l}, \dot{p}$.

We stress that the parameters are not variables in a formal language. In Situation Theory, parameters are first-class model-theoretic objects. In this paper, we use the parameters $\dot{a}$, $\dot{b}$ as representing the abstract concept of individuals that are 'unknown', and $\dot{l}$ represents a concept of a space-time location, without being any specifically determined location. The two specific individuals $a$, $b$, the location $l$, and the 'confirming' polarity 1, are instantiations of the corresponding concepts of two individuals, a location, and a polarity represented by parameters.

Next, we give examples for other parametric infons, by using a relation of reading, which is an alternative to the relation $read_2$ in (16a)–(16d), for having an extra argument role for an intended listener, which could have been denoted by $listener$. To avoid the connotation that the object filling up this argument role listens (which might not be the case), we denote it by $readee$.

*Example 3.2:* Now, we assume that $read \in \mathcal{A}_{\text{REL}}$, i.e., $read$ is a primitive situation-theoretic object of type REL, i.e., by (7b), $read : \text{REL}$. The significant difference is that, unlike in Examle 3.1, here the relation $read$ has the following three argument roles.

$$\begin{aligned} \text{ARGR}(\mathit{read}) = \{&\text{IND} : \mathit{reader}, \text{IND} : \mathit{readed}, \\ &\text{IND} : \mathit{readee}\} \end{aligned} \tag{17}$$
$$\text{for } \mathit{reader}, \mathit{readed}, \mathit{readee} \in \mathcal{A}_{\text{ARoles}}$$

Now we consider the infon (18a)–(18e).

$$\ll \mathit{read},\ \text{IND} : \mathit{reader} : a, \tag{18a}$$
$$\text{IND} : \mathit{readed} : \dot{b}, \tag{18b}$$
$$\text{IND} : \mathit{readee} : \dot{c}, \tag{18c}$$
$$\text{LOC} : \mathit{Loc} : \dot{l}; \tag{18d}$$
$$\text{POL} : \mathit{Pol} : 1 \gg \tag{18e}$$

The infon (18a)–(18c) represents the information that an individual $a$ reads the unknown or undetermined material $\dot{b}$ (i.e., $\dot{b}$ is a semantic parameter) to the unknown or undetermined $\dot{c}$ (i.e., $\dot{c}$ s a semantic parameter), at the unknown or undetermined location $\dot{l}$ (i.e., $\dot{l}$ is a location parameter). The informational piece (18a)–(18e) is about the relation $read$ between specific objects $a$, $\dot{b}$, $\dot{c}$, taking place at the specific location $\dot{l}$. The difference is that $a \in \mathcal{A}_{\text{IND}}$ is explicitly given as known, determinedly picked up from the set $\mathcal{A}_{\text{IND}}$. While $\dot{b} \in \mathcal{P}_{\text{IND}}$, $\dot{c} \in \mathcal{P}_{\text{IND}}$, and $\dot{l} \in \mathcal{P}_{\text{LOC}}$, are also specific, but are either unknown or simply left indeterminate, i.e., as parameters.

*Example 3.3:*

$$\ll \mathit{read},\ T_a : \mathit{reader} : a, \tag{19a}$$
$$T_b : \mathit{readed} : \dot{b}, \tag{19b}$$
$$T_c : \mathit{readee} : \dot{c}, \tag{19c}$$
$$\text{LOC} : \mathit{Loc} : \dot{l}; \tag{19d}$$
$$\text{POL} : \mathit{Pol} : \dot{p} \gg \tag{19e}$$

In (19a)–(19e), by $T_a$, $T_b$ and $T_c$, we represent sets of types that constrain the argument roles $reader$, $readed$, $readee$ of the relation $read$, by undetermined types.

By using a parameter for polarity, the infon (19a)–(19e) represents the parametric information that the specific object *$a$ either reads or does not read the undetermined $\dot{b}$ to the undetermined $\dot{c}$, at the undetermined location $\dot{l}$.* The undetermined polarity is represented by a semantic parameter $\dot{p}$ that is restricted to be of type POL, by the constraint over the argument role that $\dot{p}$ fills up, i.e., $\text{POL} : \mathit{Pol} : \dot{p}$. This constraint allows $\dot{p}$ to be of both types POL and PAR, i.e., $\text{POL} : \dot{p}$ and $\text{PAR} : \dot{p}$, by using (6b). Similarly, for the constraint over $\dot{l}$.

The importance of using the polarity parameter $\text{POL} : \dot{p}$, in this example, is that we do not have an explicit disjunction — we still have a piece of information by (19a)–(19e) about the relation of reading concerning $a$ as a possible reader.

The reason for which we have chosen examples with the relation of reading is not only to demonstrate the definitions, but also because it is denoted by a verb that syntactically can have either one, two, or three syntactic arguments. With the Examples 3.1–3.3, we make a point for a distinction between syntactic arguments of lexemes, in this case of the verb "read", and the corresponding semantic argument roles of the semantic relations denoted by those lexemes. Typically, a sentence like (20b) can be rendered into a term having a component infon similar to one of the infons in (16a)–(16d), which may be combined with additional infons depending on the noun phrases A and B. Thus, the verb "read", co-occurring with two syntactic arguments, would be treated as denoting a relation $read_2$ associated with two semantic argument roles (15). Similarly, the verb "read" in a sentence like (20a) can be rendered into a relation $read_1$ associated with one semantic argument role.

$$\text{A reads.} \tag{20a}$$
$$\text{A reads B.} \tag{20b}$$
$$\text{A reads B to C.} \tag{20c}$$

A sentence like (20c) can be rendered into a term with a component infon similar to the one in (18a)–(18e). In this way, when the verb "read" co-occurs with three syntactic arguments, it is rendered into a relation $read$, which is different from $read_1$ and $read_2$, by having three semantic argument roles (17). As a choice, one may keep up with this line of introducing different, variant relations, depending on the number of argument roles. This choice is deficient in representing that there is a common semantic relation of reading that may exhibit only some of its semantic argument roles in language expressions like (20a)–(20c).

Here we point that by using semantic parameters, we can render the verbal lexeme "read", occurring in the three kinds of sentences like (20a)–(20c), into the same relation $read$, associated with three semantic argument roles (17). For sentences like (20a), read would be rendered into the relation $read$, by filling up each of the argument roles $readed$ and $readee$ with undetermined parameters. The role $reader$ can be filled up by a specific individual $a$, or by a parameter $\dot{a}$ along with additional information depending on the NP A. In sentences like (20b), read would be rendered into the same relation $read$, by filling up the argument role $readee$ with an undetermined parameter, while the roles $reader$ and $readed$ can be filled by specific individuals, or by parameters along with additional information depending on NPs A and B. In this way, we have the same semantic relation $read$, which may exhibits only some of the semantic information associated with it, explicitly in syntactic expressions like (20a)–(20c). Information that is not expressed in (20a)–(20c) and is not available by context, is kept parametric and underspecified.

Thus, by the relation $read$ associated with semantic argument roles (17), we model a general, semantic concept of reading. Its argument roles can be filled up by parameters or specific individuals. Furthermore, the parameters can be additionally specified as we show in the second part of the paper. We used $read$ as a prototypical example of a class of similar basic semantic concepts.

## IV. Propositions

*Definition 1 (Proposition):* A proposition is a semantic, situation-theoretic object, represented set-theoretically in Aczel non-well-founded sets, by the tuple $\langle \text{PROP}, \mathbb{T}, \theta \rangle$, where $\mathbb{T} \in \mathcal{T}_{\text{TYPE}}$ is a type that is associated with a set of argument roles (21)

$$\text{ArgR}(\mathbb{T}) = \{T_1 : arg_1, \ldots, T_n : arg_n\} \tag{21}$$

and $\theta$ is a function, called the *argument-role filling* of the type $\mathbb{T}$, which fills up the argument roles $arg_1, \ldots, arg_n$ $(n \geq 0)$, of $\mathbb{T}$ with objects $\xi_1, \ldots, \xi_n$ of respective types $T_1, \ldots, T_n$, i.e.:

$$\theta = \{\langle T_1 : arg_1, \xi_1 \rangle, \ldots, \langle T_n : arg_n, \xi_n \rangle\} \tag{22}$$

for some situation-theoretic objects $\xi_1, \ldots, \xi_n$ satisfying the corresponding appropriateness constraints of the argument roles of the type $\mathbb{T}$, i.e.:

$$T_1 : \xi_1, \ldots, T_n : \xi_n \tag{23}$$

*Notation 5:* Typical notations are

$$\begin{aligned} \langle \text{PROP}, \mathbb{T}, \theta \rangle &\equiv \langle \mathbb{T}, \theta \rangle && \text{(24a)} \\ &\equiv (\mathbb{T} : \theta) && \text{(24b)} \\ &\equiv (\theta : \mathbb{T}) && \text{(24c)} \end{aligned}$$

The notations (24a) and (24b) resemble the application operators, where "$\mathbb{T}$ applies to the argument(s) $\theta$". The notational variants (24b) and (24c) are used alternatively depending on the context and the specific types $\mathbb{T}$. The notation (24c) follows the verbal expression "the proposition that the object(s) filling up the argument role(s) of $\mathbb{T}$ are of type $\mathbb{T}$".

*Definition 2 (Situated propositions):* Situated proposition, instantiated in a situation $s$, is any proposition (25):

$$\langle \text{PROP}, \models, \text{SIT} : s, \text{INFON} : \sigma \rangle, \tag{25}$$

where $s$ is a situation parameter $s \in \mathcal{P}_{\text{SIT}}$ i.e., $\text{SIT} : s$ and $\sigma$ is a basic or complex infon, i.e., $\sigma \in \mathcal{I}_{\text{INFON}}$.

The type $\models$ (pronounced "supports") has two argument roles, (26):

$$ArgR(\models) = \{\text{SIT} : arg_{\text{SIT}},\ \text{INFON} : arg_{\text{INFON}}\} \tag{26}$$

A proposition (25) is pronounced "the proposition that the situation $s$ supports the infon $\sigma$", or "the proposition that $\sigma$ holds in the situation $s$".

*Notation 6:*

$$\begin{aligned} \langle \text{PROP}, \models, \text{SIT} : s, \text{INFON} : \sigma \rangle &\equiv \langle \text{PROP}, \models, s, \sigma \rangle && \text{(27a)} \\ &\equiv \langle \models, s, \sigma \rangle && \text{(27b)} \\ &\equiv (s \models \sigma) && \text{(27c)} \end{aligned}$$

## V. Complex Types and Relations

Situation Theory has an abstraction operator, which resembles the $\lambda$-abstraction in functional $\lambda$-calculi, but is model-theoretic, informational abstraction. The informational abstraction is not a syntactic construction of a $\lambda$-expression in a formal language. It defines abstract complex relations and complex types, with abstract argument roles. These abstract situation-theoretic objects can be modeled with set-theoretic objects, by choosing appropriate set theory, e.g., a classic set-theory for more restricted applications, while choosing Aczel non-well-founded set theory, such as in Aczel [3], for more sophisticated applications.

### A. Complex Relations

*Definition 3 (Complex relations and argument roles):* Let $\sigma$ be a given infon, and $\{\xi_1, \ldots, \xi_n\}$ a set of parameters, i.e., primitive or complex objects of type PAR, $\text{PAR} : \xi_i$, for $i = 1, \ldots, n$ $(n \geq 0)$, which may occur in $\sigma$ (when some $\xi_i$ does not occur in $\sigma$, the abstraction over $\xi_i$ is vacuous, but it adds an additional argument role to the complex relation). Let, for each $i \in \{1, \ldots, n\}$, $T_i$ be the union of the constraints over the argument roles filled up by $\xi_i$. Then $\lambda\{\xi_1, \ldots, \xi_n\}\sigma$ is a *complex relation*, with abstract argument roles denoted by

$[\xi_1], \ldots, [\xi_n]$ and having $T_1, \ldots, T_n$ as *appropriateness (type) constraints*, respectively, i.e.:

$$\big[T_1 : [\xi_1], \ldots, T_n : [\xi_n] \mid \sigma\big] \in \mathcal{T}_{\text{REL}}, \quad \text{and} \tag{28a}$$

$$\begin{aligned} &\text{ARGR}(\big[T_1 : [\xi_1], \ldots, T_n : [\xi_n]) \mid \sigma\big] \\ &\quad = \{\langle[\xi_1], T_1\rangle, \ldots, \langle[\xi_n], T_n\rangle\} \\ &\quad \equiv \{T_1 : [\xi_1], \ldots, T_n : [\xi_n]\} \end{aligned} \tag{28b}$$

Instead of (28a), we shall primarily use the notation (29b), by suppressing the types of the argument roles when they are understood. The notation (29a) may be useful too.

$$\big[T_1 : [\xi_1], \ldots, T_n : [\xi_n] \mid \sigma\big] \tag{29a}$$

$$\equiv \{\lambda(\xi_1), \ldots, \lambda(\xi_n)\}\sigma \tag{29b}$$

$$\equiv \lambda\{\xi_1, \ldots, \xi_n\}\,\sigma \tag{29c}$$

Upon agreed order, which is a typical practice in mathematics and in computer science, the argument roles (usually called argument slots) and/or the types constraints can be skipped, and only the objects filling up the argument roles (slots) are listed.

### B. Complex Types

In this subsection, we define the abstract objects complex types. They are significant for what follows in this paper.

*Definition 4 (Complex types and argument roles):* Assume that

1) $\Theta$ is a given situation-theoretic proposition, and $\xi_1, \ldots, \xi_n$ are parameters, i.e., PAR : $\xi_i$, for $i = 1, \ldots, n$ ($n \geq 0$).
2) For each $i \in \{1, \ldots, n\}$, $T_i$ is the union of all the appropriateness constraints of all the argument roles of constituents of $\Theta$ that are filled up by $\xi_i$.

Note that $\Theta$ may have various, components (constituents), which are either types or relations, with arguments roles that are filled by $\xi_i$. In addition, a single constituent type, or a constituent relation, may have more than one argument role filled by $\xi_i$.

The result of the *abstraction over the parameters* $\xi_1, \ldots, \xi_n$ from the proposition $\Theta$ is a *complex type* (30a), with argument roles $[\xi_i]$ that are associated with *appropriateness (type) constraints*, respectively, $(T_i : [\xi_i])$, for $i = 1, \ldots, n$ ($n \geq 0$), i.e., (30b):

$$\big[T_1 : [\xi_1], \ldots, T_n : [\xi_n] \mid \Theta\big] \in \mathcal{T}_{\text{TYPE}}, \quad \text{and} \tag{30a}$$

$$\begin{aligned} &\text{ARGR}([T_1 : [\xi_1], \ldots, T_n : [\xi_n] \mid \Theta]) \\ &\quad = \{\langle[\xi_1], T_1\rangle, \ldots, \langle[\xi_n], T_n\rangle\} \\ &\quad = \{T_1 : [\xi_1], \ldots, T_n : [\xi_n]\} \end{aligned} \tag{30b}$$

Thus the abstraction over the parameters $\xi_1, \ldots, \xi_n$, from the proposition $\Theta$, results in complex, abstract argument roles $[\xi_1]$, …, $[\xi_n]$ of the complex type $\big[T_1 : [\xi_1], \ldots, T_n : [\xi_n] \mid \Theta\big]$. The abstraction creates argument roles along with appropriateness type constraints $T_1 : [\xi_1], \ldots, T_n : [\xi_n]$.

Note that, as in the complex relations, the parameters $\xi_i$, $i = 1, \ldots, n$, are primitive or complex objects of type PAR, PAR : $\xi_i$, for $i = 1, \ldots, n$ ($n \geq 0$), which may occur in $\Theta$. When some $\xi_i$ does not occur in $\Theta$, the abstraction over $T_i : \xi_i$ is vacuous, but it adds an argument role of the complex type (30a). In the cases when $n = 0$, the complex type has no argument roles.

*Notation 7:* A complex type $[T_1 : [\xi_1], \ldots, T_n : [\xi_n] \mid \Theta]$ is alternatively denoted by (31b), when the argument roles of the result of the abstraction operation are suppressed; and with (31c) when the corresponding type constraints over the complex roles are suppressed too.

$$[T_1 : [\xi_1], \ldots, T_n : [\xi_n] \mid \Theta] \tag{31a}$$

$$\equiv [T_1 : \xi_1, \ldots, T_n : \xi_n \mid \Theta] \tag{31b}$$

$$\equiv [\xi_1, \ldots, \xi_n \mid \Theta] \tag{31c}$$

$$\equiv \lambda\{\xi_1, \ldots, \xi_n\}\Theta \tag{31d}$$

Similarly to the complex relations in (29a)–(29c), complex types might sometimes be denoted by the $\lambda$-notation $\lambda\{\xi_1, \ldots, \xi_n\}\Theta$ in (31d). Whether a situation-theoretic object $[T_1 : \xi_1, \ldots, T_n : \xi_n \mid \vartheta]$, as an abstraction over the parameters PAR : $\xi_1$, …, PAR : $\xi_n$ in an object $\vartheta$, is a complex relation or complex type depends on whether $\vartheta$ is an infon or a proposition, not on what notation we use for it per se. Nevertheless, we shall primarily use the notations in (29b)–(29c) for relations, and (31a)–(31c) for types, in order to make clear distinction between (1) the abstract complex relations, which are abstractions over parameters from infons; and (2) the abstract complex types, which are abstractions over parameters from propositions.

## VI. Complex Concepts

In this section, we demonstrate how to use complex parametric types to model taxonomic concepts. The concepts can be at different level of abstraction over various parameters.

### A. A Sample Concept

Here we demonstrate the general ideas of concepts in taxonomic classifications, e.g., the class of *odd-toed, ungulate* entities as a subclass of *animate* entities. Then, we demonstrate how to instantiate the concept of an odd-toed, ungulate entity with a specific representative $a$ having a complex property, e.g., walking in a space-time location and sleeping in another space-time location.

$$T_{animate} \equiv \tag{32a}$$

$$\begin{aligned} &[\text{IND} : [x] \mid (\dot{s} \models \ll animate, \text{IND} : arg : x, \\ &\qquad\qquad \text{LOC} : \dot{l}; 1 \gg)] \end{aligned} \tag{32b}$$

The type $T_{otu}$ defined in (33a)-(33c) is the type of odd-toed ungulate individuals in some (parametric) situation $\dot{s}$ and location $\dot{l}$:

$$T_{otu} \equiv [T_{animate} : [x] \mid (\dot{s} \models \tag{33a}$$

$$\ll \textit{odd-toed}, T_{animate} : arg : x, \text{LOC} : \dot{l}; 1 \gg \land \tag{33b}$$

$$\ll ungulate, T_{animate} : arg : x, \text{LOC} : \dot{l}; 1 \gg)] \tag{33c}$$

### B. Instantiation of Concepts

The type (34a) is a two-argument type of a situation and a location, in which a specific odd-toed ungulate individual $a$ walks. Note that the type (34a) is defined only in case the proposition $(a : T_{otu})$ is true, i.e., $a : T_{otu}$, which is so in case $a$ of type $T_{otu}$.

$$[\text{SIT} : [\dot{s}_0], \text{LOC} : [\dot{l}_0] \mid (\dot{s}_0 \models \tag{34a}$$

$$\ll \mathit{walk}, T_{otu} : \mathit{walker} : a, \text{LOC} : \dot{l}_0; 1 \gg)] \tag{34b}$$

The type (35a)-(35d) is a three-argument type of situations $\dot{s}_0$ and locations $\dot{l}_0$, $\dot{l}_1$, such that, in the situation $\dot{s}_0$, the odd-toed ungulate individual $a$ walks through the location $\dot{l}_0$, and sleeps through the location $\dot{l}_1$, and where the locations $\dot{l}_0$ and $\dot{l}_1$ are not overlapping in time.

$$[\text{SIT} : [\dot{s}_0], \text{LOC} : [\dot{l}_0], \text{LOC} : [\dot{l}_1] \mid \tag{35a}$$

$$(\dot{s}_0 \models \ll \mathit{walk}, T_{otu} : \mathit{walker} : a, \text{LOC} : \dot{l}_0; 1 \gg \wedge \tag{35b}$$

$$\ll \mathit{sleep}, T_{otu} : \mathit{sleeper} : a, \text{LOC} : \dot{l}_1; 1 \gg) \wedge \tag{35c}$$

$$(\dot{l}_0 \not\circ_t \ \dot{l}_1)] \tag{35d}$$

The type (36a)-(36d) is a four-argument type of situations $\dot{s}_0$, $\dot{s}_1$ and locations $\dot{l}_0$, $\dot{l}_1$, such that, in the situation $\dot{s}_0$, the odd-toed ungulate individual $a$ walks through the location $\dot{l}_0$, and, in the situation $\dot{s}_1$, $a$ sleeps through the location $\dot{l}_1$, where the locations $\dot{l}_0$ and $\dot{l}_1$ are not time overlapping.

$$[\text{SIT} : [\dot{s}_0], \text{LOC} : [\dot{l}_0], \text{SIT} : [\dot{s}_1], \text{LOC} : [\dot{l}_1] \mid \tag{36a}$$

$$(\dot{s}_0 \models \ll \mathit{walk}, T_{otu} : \mathit{walker} : a, \text{LOC} : \dot{l}_0; 1 \gg) \wedge \tag{36b}$$

$$(\dot{s}_1 \models \ll \mathit{sleep}, T_{otu} : \mathit{sleeper} : a, \text{LOC} : \dot{l}_1; 1 \gg) \wedge \tag{36c}$$

$$(\dot{l}_0 \not\circ_t \ \dot{l}_1)] \tag{36d}$$

### C. Concept Subdivision

A conceptual class can be subdivided into subclasses by using subtypes. E.g., we demonstrate the technique, by the type $T_{nocturnal}$ in (37a) defined as a subtype of the type $T_{otu}$ defined in (33a)-(33c).

$$T_{nocturnal} \equiv [\text{IND} : [x] \mid \tag{37a}$$

$$(\dot{s} \models \ll \mathit{nocturnal}, \text{IND} : x, \text{LOC} : \dot{l}; 1 \gg)] \tag{37b}$$

Now, the type $T_{nocturnal\text{-}otu}$ in (38a)-(38f), has a single argument role $[\dot{a}]$:

$$T_{nocturnal\text{-}otu} = [\{T_{nocturnal}, T_{otu}\} : \dot{a} \mid \tag{38a}$$

$$(\dot{s}_2 \models \ll \mathit{healthy}, T_{nocturnal} : \mathit{arg} : \dot{a}, \tag{38b}$$

$$\text{LOC} : \dot{l}_2; 1 \gg) \wedge \tag{38c}$$

$$(\dot{s}_0 \models \ll \mathit{walk}, T_{otu} : \mathit{walker} : \dot{a}, \text{LOC} : \dot{l}_0; 1 \gg) \wedge \tag{38d}$$

$$(\dot{s}_1 \models \ll \mathit{sleep}, T_{otu} : \mathit{sleeper} : \dot{a}, \text{LOC} : \dot{l}_1; 1 \gg) \wedge \tag{38e}$$

$$(\dot{l}_0 \not\circ_t \ \dot{l}_1) \wedge (\dot{l}_0 \subset \dot{l}_2) \wedge (\dot{l}_1 \subset \dot{l}_2)] \tag{38f}$$

The type $T_{nocturnal\text{-}otu}$ in (38a)-(38f), that has a single argument role $[\dot{a}]$ with the argument constraint $\{T_{nocturnal}, T_{otu}\}$ for appropriate filling, i.e.:

$$\text{ARGR}(T_{nocturnal\text{-}otu}) = \{\, \{T_{nocturnal}, T_{otu}\} : [\dot{a}] \,\} \tag{39}$$

Now, while the types in Examples VI-A-VI-C share the parameters for a situation $s$ and a location $l$, we might consider them as related by them, as long as these types are used together, as if in a "package". In case these types are "separated", the parameters $s$ and $l$ can be anchored, i.e., instantiated, with unrelated objects of the respective types.

### D. Instances of Concepts in Conceptual Sub-divisions

Let $c_1$ be a parameter assignment for the type $T_{nocturnal\text{-}otu}$ in (38a)-(38f), such that:

$$c_1(\dot{s}_0) = s_0, \quad c_1(\dot{l}_0) = l_0, \tag{40a}$$

$$c_1(\dot{s}_1) = s_1, \quad c_1(\dot{l}_1) = l_1, \tag{40b}$$

$$c_1(\dot{s}_2) = s_2, \quad c_1(\dot{l}_2) = l_2. \tag{40c}$$

Then the following propositions have the same truth values:

$$(c_1(T_{nocturnal\text{-}otu}) : a) \text{ is true} \tag{41a}$$

$$\Longleftrightarrow$$

$$(c_1(T_{nocturnal}) : a) \wedge \tag{41b}$$

$$(c_1(T_{otu}) : a) \wedge \tag{41c}$$

$$(s_2 \models \ll \mathit{healthy}, T_{nocturnal} : \mathit{arg} : a, \tag{41d}$$

$$\text{LOC} : l_2; 1 \gg) \wedge$$

$$(s_0 \models \ll \mathit{walk}, T_{otu} : \mathit{walker} : a, \tag{41e}$$

$$\text{LOC} : l_0; 1 \gg) \wedge$$

$$(s_1 \models \ll \mathit{sleep}, T_{otu} : \mathit{sleeper} : a, \tag{41f}$$

$$\text{LOC} : l_1; 1 \gg) \wedge$$

$$(l_0 \not\circ_t \ l_1) \wedge (l_0 \subset l_2) \wedge (l_1 \subset l_2) \tag{41g}$$

In general, in order to model parametric assignments and cognitive concepts, including for semantics of natural language, and in systems of taxonomic information, we allow the sub-propositions $(c_1(T_{nocturnal}) : a)$ and $(c_1(T_{otu}) : a)$ may be true by "preserving" parametric information via assignment $c_1(\dot{s}) = \dot{s}$ and $c_1(\dot{l}) = \dot{l}$. Then, the following propositions are truth equivalent:

$$(c_1(T_{nocturnal}) : a) \iff (T_{nocturnal} : a) \tag{42a}$$

$$\iff (c_2(T_{nocturnal}) : a) \tag{42b}$$

In the above types, there is not explicit requirement that further update of information, by a new parameter assignment $c_2$, have to agree on the parameters $\dot{s}$ and $\dot{l}$ (which in certain cases may be desirable). Let $c_2$ be a parameter assignment for the type $T_{nocturnal}$ in (37a), such that:

$$c_2(\dot{s}) = s', \quad c_2(\dot{l}) = l'. \tag{43}$$

Then:

$$(c_2(T_{nocturnal}) : a) \quad \text{is true} \tag{44a}$$

$$\begin{aligned} \iff (s' \models \ll \mathit{nocturnal}, \text{IND} : \mathit{arg} : a, \\ \text{LOC} : l'; 1 \gg) \quad \text{is true} \end{aligned} \tag{44b}$$

$$\begin{aligned} \iff s' \models \ll \mathit{nocturnal}, \text{IND} : \mathit{arg} : a, \\ \text{LOC} : l'; 1 \gg \end{aligned} \tag{44c}$$

The parameter assignment $c_1$ may not agree with $c_2$, for example, in case $s \neq s'$ and $s \neq s'$. There is a need for explicitly expressing that $c_1$ should be a parameter assignment for the types $T_{nocturnal}$ in (37a), $T_{otu}$ in (33a)-(33c), and $T_{animate}$ in (32b), so that they agree on $\dot{s}$ and $\dot{l}$, as in (45):

$$c_1(\dot{s}) = s, \quad c_1(\dot{l}) = l, \tag{45}$$

The effect of (41a) is represented in (46a)–(46h):

$$(c_1(T_{nocturnal\text{-}otu}) : a) \text{ is true} \tag{46a}$$

$$\Longleftrightarrow$$

$$(s \models \ll nocturnal, \text{IND} : arg : a, \text{LOC} : l; 1 \gg) \wedge \tag{46b}$$

$$\begin{aligned}(s \models \ll odd\text{-}toed, T_{animate} : arg : a, \text{LOC} : l; 1 \gg \wedge \\ \ll ungulate, T_{animate} : arg : a, \text{LOC} : l; 1 \gg) \wedge\end{aligned} \tag{46c}$$

$$(s \models \ll animate, \text{IND} : arg : a, \text{LOC} : l; 1 \gg) \wedge \tag{46d}$$

$$(s_2 \models \ll healthy, T_{nocturnal} : arg : a, \text{LOC} : l_2; 1 \gg) \wedge \tag{46e}$$

$$(s_0 \models \ll walk, T_{otu} : walker : a, \text{LOC} : l_0; 1 \gg) \wedge \tag{46f}$$

$$(s_1 \models \ll sleep, T_{otu} : sleeper : a, \text{LOC} : l_1; 1 \gg) \wedge \tag{46g}$$

$$(l_0 \not\circ_t \; l_1) \wedge (l_0 \subset l_2) \wedge (l_1 \subset l_2) \tag{46h}$$

## VII. Linking Parameters in Informational Structures

The following examples are patterns of how to achieve effects of linking argument roles of relations and types and their filling. The results are complex informational structures, as in (41a)–(46h), that represent general, informational patterns.

### A. Complex, Interrelated Types

The type $T_{sitan}$ in (47a)-(47b) is a three-argument type, with argument roles for an individual, $\text{IND} : [\xi]$, a situation, $\text{SIT} : [\dot{s}']$, and a location, $\text{LOC} : [\dot{l}']$.

$$T_{sitan} = [\text{IND} : [\xi], \text{SIT} : [\dot{s}'], \text{LOC} : [\dot{l}'] \;\mid \tag{47a}$$

$$\begin{aligned}(\dot{s}' \models \ll animate, \text{IND} : arg : \xi, \\ \text{LOC} : \dot{l}'; 1 \gg)]\end{aligned} \tag{47b}$$

Thus, the type $T_{sitan}$ is a type having three argument roles: the argument $[\xi]$ that can be filled up by an animate individual; the argument $[\dot{s}']$ can be filled up only by a situation where the individual filling up the role $[\xi]$ is animate; and $[\dot{l}']$ can be filled up only by the corresponding location in that situation. E.g., as in the proposition (48b) that is constituent of the type $T_{sitotu}$ in (48a)-(48c).

The type $T_{sitotu}$ in (48a)-(48c) is a three-argument type, with argument roles for an individual, $\text{IND} : [x]$, a situation, $\text{SIT} : [\dot{s}]$, and a location, $\text{LOC} : [\dot{l}]$, and is a type of odd-toed ungulate individuals $x$ in a situation $\dot{s}$ and a location $\dot{l}$:

$$T_{sitotu} = [\text{IND} : [x], \text{SIT} : [\dot{s}], \text{LOC} : [\dot{l}] \;\mid \tag{48a}$$

$$(T_{sitan}, \text{IND} : [\xi] : x, \text{SIT} : [\dot{s}'] : \dot{s}, \text{LOC} : [\dot{l}'] : \dot{l}) \wedge \tag{48b}$$

$$\begin{aligned}(\dot{s} \models \ll odd\text{-}toed, \text{IND} : arg : x, \text{LOC} : \dot{l}; 1 \gg \wedge \\ \ll ungulate, T_{sitan} : arg : x, \text{LOC} : \dot{l}; 1 \gg)]\end{aligned} \tag{48c}$$

The constituent proposition $(T_{sitan}, \text{IND} : [\xi] : x, \text{SIT} : [\dot{s}'] : \dot{s}, \text{LOC} : [\dot{l}'] : \dot{l})$ in (48b) states that the objects $x$, $\dot{s}$, and $\dot{l}$ (filling correspondingly the argument-roles $[\xi]$, $[\dot{s}']$, and $[\dot{l}']$ of the type $T_{sitan}$) have to be of type $T_{sitan}$.

### B. Interrelated Generalizations

Now, we demonstrate the generalization of the type $T_{nocturnal}$ in (37a) to a situated type that has three arguments:

$$T_{sitnoct} = [\text{IND} : [\zeta], \text{SIT} : [\dot{s}''], \text{LOC} : [\dot{l}''] \;\mid \tag{49a}$$

$$(\dot{s}'' \models \ll nocturnal, \text{IND} : \zeta, \text{LOC} : \dot{l}''; 1 \gg)] \tag{49b}$$

The type $T_{sitnoctotu}$ in (50a)-(50e) is a three-argument type, with argument roles for an individual, $\text{IND} : [x]$, a situation, $\text{SIT} : [\dot{s}]$, and a location, $\text{LOC} : [\dot{l}]$, and is a type of odd-toed ungulate individuals $x$, that are nocturnal, in a situation $\dot{s}$ and a location $\dot{l}$.

As in (48a)-(48c), the constituent proposition $(T_{sitan}, \text{IND} : [\xi] : x, \text{SIT} : [\dot{s}'] : \dot{s}, \text{LOC} : [\dot{l}'] : \dot{l})$ in (50b) states that the object parameters $x$, $\dot{s}$, and $\dot{l}$ (filling correspondingly the argument-roles $[\xi]$, $[\dot{s}']$, and $[\dot{l}']$ of the type $T_{sitnoct}$) have to be of type $T_{sitnoct}$. Any objects respectively instantiating these parameters have to be of these types too. The constituent $(T_{sitnoct}, \text{IND} : [\zeta] : x, \text{SIT} : [\dot{s}''] : \dot{s}, \text{LOC} : [\dot{l}''] : \dot{l})$ in (50c) is a proposition, which stating that the same object parameters $x$, $\dot{s}$, and $\dot{l}$ (filling correspondingly the argument-roles $[\zeta]$, $[\dot{s}'']$, and $[\dot{l}''])$ have to be of type $T_{sitnoct}$, also.

$$T_{sitnoctotu} = [\text{IND} : [x], \text{SIT} : [\dot{s}], \text{LOC} : [\dot{l}] \;\mid \tag{50a}$$

$$(T_{sitan}, \text{IND} : [\xi] : x, \text{SIT} : [\dot{s}'] : \dot{s}, \text{LOC} : [\dot{l}'] : \dot{l}) \wedge \tag{50b}$$

$$\begin{aligned}(T_{sitnoct}, \text{IND} : [\zeta] : x, \text{SIT} : [\dot{s}''] : \dot{s}, \\ \text{LOC} : [\dot{l}''] : \dot{l}) \wedge\end{aligned} \tag{50c}$$

$$(\dot{s} \models \ll odd\text{-}toed, \text{IND} : arg : x, \text{LOC} : \dot{l}; 1 \gg \wedge \tag{50d}$$

$$\ll ungulate, T_{sitan} : arg : x, \text{LOC} : \dot{l}; 1 \gg)] \tag{50e}$$

Therefore, the truth conditions for the proposition (51a) are equivalent to those for the conjunction (51b)–(51d), where $a$, $s$, and $l$, are either specific objects or parameters, respectively for an individual, situation, and space-time location. I.e., the proposition (51a) stating that $a$, $s$, $l$ are of the type $T_{sitnoctotu}$ is true iff the conjunctive proposition (51b)–(51d) is true by instantiating the parameters $x$, $\dot{s}$, $\dot{l}$ with $a$, $s$, $l$, correspondingly:

$$(T_{sitnoctotu}, \text{IND} : [x] : a, \text{SIT} : [\dot{s}] : s, \text{LOC} : [\dot{l}] : l) \tag{51a}$$

$$\Longleftrightarrow$$

$$(T_{sitan}, \text{IND} : [\xi] : x, \text{SIT} : [\dot{s}'] : \dot{s}, \text{LOC} : [\dot{l}'] : \dot{l}) \wedge \tag{51b}$$

$$\begin{aligned}(T_{sit\text{-}nocturnal}, \text{IND} : [\zeta] : x, \text{SIT} : [\dot{s}''] : \dot{s}, \\ \text{LOC} : [\dot{l}''] : \dot{l}) \wedge\end{aligned} \tag{51c}$$

$$\begin{aligned}(\dot{s} \models \ll odd\text{-}toed, \text{IND} : arg : x, \text{LOC} : \dot{l}; 1 \gg \wedge \\ \ll ungulate, T_{sitan} : arg : x, \text{LOC} : \dot{l}; 1 \gg)\end{aligned} \tag{51d}$$

$$\text{is true for } c(x) = a, \; c(\dot{s}) = s, \; c(\dot{l}) = l, \tag{51e}$$

where $c$ is a function for parameter assignment, i.e., instantiating, the parameters $x$, $\dot{s}$, $\dot{l}$ with $a$, $s$, $l$, correspondingly. In what follows, we shall use simply the equality sign for direct parameter assignment, without using the assignment function *c*. The above information can be expressed all together in the following way:

$$\begin{aligned}&(T_{sitnoctotu}, \text{IND} : [x] : a,\ \text{SIT} : [\dot{s}] : s,\\ &\qquad \text{LOC} : [\dot{l}] : l\ )\end{aligned} \tag{52a}$$

$$\iff (p_1 \wedge p_2 \wedge p_3) \quad \text{is true} \tag{52b}$$

where

$$\begin{aligned}p_1 = \big(\big[[x], [\dot{s}], [\dot{l}]\ \mid \big(&T_{sitan},\ \text{IND} : [\xi] : x,\\ &\text{SIT} : [\dot{s}'] : \dot{s},\\ &\text{LOC} : [\dot{l}'] : \dot{l}\big)\big],\end{aligned} \tag{53a}$$

$$\text{IND} : [x] : a,\ \text{SIT} : [\dot{s}] : s,\ \text{LOC} : [\dot{l}] : l\big), \tag{53b}$$

$$\begin{aligned}p_2 = \big(\big[[x], [\dot{s}], [\dot{l}]\ \mid \big(&T_{sitnoct},\ \text{IND} : [\zeta] : x,\\ &\text{SIT} : [\dot{s}''] : \dot{s},\\ &\text{LOC} : [\dot{l}''] : \dot{l}\big)\big],\end{aligned} \tag{53c}$$

$$\text{IND} : [x] : a,\ \text{SIT} : [\dot{s}] : s,\ \text{LOC} : [\dot{l}] : l\ \big), \tag{53d}$$

$$\begin{aligned}p_3 = \big(\big[&[x], [\dot{s}], [\dot{l}]\ \mid\\ &\big(\dot{s} \models \ll \textit{odd-toed},\ \text{IND} : \textit{arg} : x,\\ &\qquad \text{LOC} : \dot{l}; 1 \gg \wedge\\ &\ll \textit{ungulate},\ T_{sitan} : \textit{arg} : x,\\ &\qquad \text{LOC} : \dot{l}; 1 \gg \big)\big],\end{aligned} \tag{53e}$$

$$\text{IND} : [x] : a,\ \text{SIT} : [\dot{s}] : s,\ \text{LOC} : [\dot{l}] : l\ \big) \tag{53f}$$

Now, instead of repeating the informational patterns in (52b)-(53f), for various specific types and properties, we can use sub-typing, by type and relation parameters, which can be instantiated as necessary.

$$(T_{sitnoctotu}, \text{IND} : [x] : a,\ \text{SIT} : [\dot{s}] : s,\ \text{LOC} : [\dot{l}] : l\ ) \tag{54a}$$

$$\iff (p_1 \wedge p_2 \wedge p_3) \tag{54b}$$

where

$$\begin{aligned}p_1 = \big(\big[&[x], [\dot{s}], [\dot{l}]\ \mid\\ &\big(T_1,\ \text{IND} : [\xi] : x,\ \text{SIT} : [\dot{s}'] : \dot{s},\\ &\qquad \text{LOC} : [\dot{l}'] : \dot{l}\ \big)\big],\\ &\text{IND} : [x] : a,\ \text{SIT} : [\dot{s}] : s,\ \text{LOC} : [\dot{l}] : l\ \big),\end{aligned} \tag{55a}$$

$$\begin{aligned}p_2 = \big(\big[&[x], [\dot{s}], [\dot{l}]\ \mid\\ &\big(T_2, \text{IND} : [\zeta] : x, \text{SIT} : [\dot{s}''] : \dot{s},\\ &\qquad\qquad \text{LOC} : [\dot{l}''] : \dot{l}\big)\big],\\ &\text{IND} : [x] : a, \text{SIT} : [\dot{s}] : s, \text{LOC} : [\dot{l}] : l\big),\end{aligned} \tag{55b}$$

$$\begin{aligned}p_3 = \big(\big[&[x], [\dot{s}], [\dot{l}]\ \mid\\ &\big(\dot{s} \models \ll r,\ \text{IND} : [x] : x,\\ &\qquad \text{LOC} : [l] : l; 1 \gg \big)\big],\\ &\text{IND} : [x] : a,\ \text{SIT} : [\dot{s}] : s,\ \text{LOC} : [\dot{l}] : l\ \big),\end{aligned} \tag{55c}$$

$$\begin{aligned}r = \big[[x], [\dot{l}]\ \mid &\ll r_u,\ \text{IND} : [x] : x,\\ &\qquad \text{LOC} : [l] : l; 1 \gg \wedge\\ &\ll r_t,\ \text{IND} : [x] : x,\ \text{LOC} : [l] : l; 1 \gg \big],\end{aligned} \tag{55d}$$

$$\begin{aligned}r_u = \big[&\text{IND} : [x],\ \text{LOC} : [\dot{l}] : \dot{l}\ \mid\\ &\ll u,\ T_1 : \textit{arg} : x,\ \text{LOC} : \dot{l}; 1 \gg \big],\end{aligned} \tag{55e}$$

$$\begin{aligned}r_t = \big[&\text{IND} : [x],\ \text{LOC} : [\dot{l}] : \dot{l}\ \mid\\ &\ll t,\ \text{IND} : \textit{arg} : x,\ \text{LOC} : \dot{l}; 1 \gg \big],\end{aligned} \tag{55f}$$

$$T_1 = T_{sitan},\ T_2 = T_{sitnoct},\ u = \textit{ungulate}, \tag{55g}$$

$$t = \textit{odd-toed}, \tag{55h}$$

$$a = a_0,\ s = s_0,\ l = l_0 \tag{55i}$$

Now, the parameter $t$ in (54b)–(55i) can be left underspecified, by dropping out the instantiation $t = \textit{odd-toed}$ in (55h). However, one may still need to restrict the possible instantiations of $t$. E.g., that can be done by replacing (55f) and (55h) with (56a)–(56b).

$$\begin{aligned}r_t = \Big[&\text{IND} : [x],\ \text{LOC} : [\dot{l}] : \dot{l}\ \mid\\ &\ll t_o,\ \text{IND} : \textit{arg} : x,\ \text{LOC} : \dot{l}; 1 \gg\ \vee\\ &\ll t_e,\ \text{IND} : \textit{arg} : x,\ \text{LOC} : \dot{l}; 1 \gg \Big],\end{aligned} \tag{56a}$$

$$t_o = \textit{odd-toed},\ t_e = \textit{even-toed} \tag{56b}$$

Another possibility to introduce alternative instantiations, i.e., alternative parameter assignments, is by using sets and *membership instantiations*:

$$\begin{aligned}p \in M,\ \text{for given } &T : \text{TYPE}, p \in \text{PAR}, p : T,\\ &\text{and a set } M \text{ of objects of type } T\end{aligned} \tag{57}$$

I.e., in (54b)–(55i), we can replace the assignment $t = \textit{odd-toed}$ in (55h) with (58).

$$t \in \{\ \textit{odd-toed}, t = \textit{even-toed}\ \} \tag{58}$$

The truth values of the proposition in (51a), i.e., in (52a) and (54a), can be determined, i.e. calculated, from (51b)–(51e), alternatively from (52b)–(53f), or (54b)–(55h). The informational structures in (52b)–(53f) and in (54b) reveal detailed informational compounds and how they are "linked" in general informational patterns. In particular, the informational structure of (54b) reveals:

1) The informational structures of the propositions $p_1$ and $p_2$ have the same pattern given in (59):

$$\big(T, \text{IND} : [\xi] : x, \text{SIT} : [\dot{s}'] : \dot{s}, \text{LOC} : [\dot{l}'] : \dot{l}\big) \tag{59}$$

Note that the distinctions between $\dot{s}'$ and $\dot{s}''$, and respectively, between $\dot{l}'$ and $\dot{l}''$, are inessential. The propositions $p_1$ and $p_2$ differ in the specific instantiation of the type parameter $T$ with $T_{sitan}$ and $T_{sitnoct}$, respectively.

2) $p_3$ is the general information pattern for a situated, propositional content, where $r$ is a relation parameter, which can be instantiated, i.e., anchored to by a parameter assignment, or equality like (55d), to any unary relation (without counting the location argument, which is specially designated). In this case it is instantiated by a complex conjunction infon. The parameter $r$ can be considered as a relation parameter $r$ with any number of arguments, $n \geq 0$:

$$\begin{aligned}\big(\dot{s} \models \ll r,\ &T_1 : \textit{arg}_1 : x_1,\ \ldots, T_1 : \textit{arg}_n : x_n,\\ &\text{LOC} : [l] : l; 1 \gg \big)\end{aligned} \tag{60}$$

## VIII. General Argument Structure in Complex Relations

The relation parameters $r_t$ and $r_u$ are instantiated to complex relations by (55f) and (55e) respectively, both of which have the same informational structure as $r_{2,R_2}$ in (61):

$$\begin{aligned} r_{2,R_2} = [&\textsc{ind} : [x],\ \textsc{loc} : [\dot{l}] : \dot{l} \ \mid \\ &\ll R_2,\ \textsc{ind} : \mathit{arg} : x,\ \textsc{loc} : \dot{l}; 1 \gg] \end{aligned} \tag{61}$$

The relation $r_{m,R_n}$, in (62), is the generalization of the two-argument, parametric relation $r_{2,R_2}$ in (61) to a parametric relation $r_{m,R_n}$ of several complex argument roles $[x_{j_1}]$, ..., $[x_{j_m}]$, $[\dot{l}]$, by abstraction over the location parameter $\dot{l}$, and a set of parameters, $T_{j_1} : x_{j_1}, \dots, T_{j_m} : x_{j_m}$, filling some of the argument roles of an $n$-argument relation parameter $R_n$:

$$\begin{aligned} r_{m,R_n} = \big[&T_{j_1} : [x_{j_1}],\ \dots,\ T_{j_m} : [x_{j_m}],\ \textsc{loc} : [\dot{l}] : \dot{l} \ \mid \\ &\ll R_n,\ T_1 : \mathit{arg}_1 : x_1,\ \dots, T_n : \mathit{arg}_n : x_n, \\ &\quad \textsc{loc} : \dot{l}; 1 \gg \big] \end{aligned} \tag{62}$$

## IX. Conclusions and Future Work

### A. Conclusions

The paper covers the topic of linking parametric information in taxonomic concepts that are constructed by using relations and types having complex argument roles. In general, the argument roles of complex relations and types are constrained by situational types. The argument roles can only be saturated by appropriate situational objects of the corresponding types. The objects that fill-up the complex argument roles in parametric objects can be parametric too. Constituents of information can be linked via recursively constructed, parametric components and instantiations.

We have shown how informational patterns can be productively re-used by different instantiations, with updated situational constituents, which can be specific or again parametric. The constraints over abstract argument roles are expressed recursively with complex types, via abstraction operator.

The presented situation-theoretic approach is primarily intended for modeling information that is typically partial, parametric, and depends on context and situations, as in concepts. In addition, in the contemporary explosion of technologies for large databases of uncertain information, we find that Situation Theory can be used for efficient representation of large collections of data and information streams. Information can be hierarchically organized according to informational types in classes and subclasses. Instead of repeating larger or smaller amounts of details, concepts represent parametric generalizations, which can be instantiated depending on needs and specific situations. Typically, parametric instantiations are dependent on situations. Situation-theoretic types, which are used in concepts with components that are parameters for situations, carry informational content that is "placed" in abstract, "unknown" situations.

### B. Future Developments

Future work, which is related to the topic of this paper, is linking parametric objects via complex restricted parameters, for representing underspecified cognitive concepts. Situation Theory with similar parametric objects has been used for semantics of attitude expressions and quantifier ambiguities (e.g., see Loukanova [11], [12], [22]).

A related topic, on which we work concurrently, is development of formal languages for Situation Theory, see Loukanova [10], [21]. Forthcoming work is on association of such formal languages with denotational and algorithmic semantics. The denotational structures of such formal languages are situation-theoretic domains, as in Loukanova [7]–[9], and in this paper. The varieties of formal languages depend on the areas of application, coverage of semantic concepts, and the situation-theoretic domains of objects. A formal language of Situation Theory can be useful for expressing general semantic concepts such as the ones introduced in this paper.

As we explained in Section I, Aczel non-well-founded set theory provides a theoretic limitation of the situation-theoretic objects, including situations, to objects that conform to Aczel Anti-Foundation Axiom (AFA). These objects can be large as non-well-founded sets, and can be circular, but they have finite representations, e.g., by finite graphs, which, in case of circular information, are circular graphs. Specific applications can have additional restrictions on the objects in their domain and still allow circular information, if needed. In typical applications, circularity is undesirable by leading to circular algorithms, which may not end. Excluding circular situation-theoretic objects from the domains is complicated at model-theoretic level. Using a formal language of Situation Theory, as in Loukanova [10], provides a useful tool by formulation of acyclicity constraints over formal terms denoting situation-theoretic objects.

A primary area of applications of Situation Theory is to computational semantics of languages. Typical approaches to computational semantics encounter problems due to progressive expansion of ambiguities, and multiple, unknown, or undetermined interpretations, which actually depend on partial, and underspecified semantic information. Situation-theoretic objects, as introduced in this paper, model informational content of concepts by using semantic parameters. which can be instantiated depending on context, situations, events, and other resources.

Among the target areas of applications of the presented approach are neuroscience of language and cognitive science, where cognitive concepts are essential. Of particular interests are situation-theoretic models of forming, developing, productivity, and efficiency of language and cognitive concepts and universals.

An open area of work is on formal languages of Situation Theory, e.g., as in Loukanova [10], [21], expanding them with syntax-semantics inferences, checks for consistency, and in addition, implementing them as computerized systems. Such practical implementations require work on developments of specialized algorithms, their classification with respect to complexity, developments of databases, and techniques for evaluations.

### C. Comparative Studies

Detailed comparative study of the technique introduced in this paper with other techniques for representation of concepts is an unexplored area. The technique introduced here shares ideas with work from fuzzy networks, see Yager [23]. Another direction of comparative study, by considering prospects for new developments, is with the approach of Rough Neural Computing (RNC), see Pal et al. [24]. Comparison with RNC would be of especial interest because RNC is a development of modeling neural networks for computations by using lexemes from human language. Similarly, situation-theoretic representations of concepts, as introduced here, have essential components consisting of semantic representations of lexical items and phrases from human language. Comparative studies may open possibilities for enhanced integration of approaches. We have a preliminary view that approaches that use fuzzy logic and fuzzy sets can benefit by further developments integrating Situation Theory and formal languages for it. This will enhance representation of fuzzy informational concepts and other informational units, by enriching them with content that is structured information, even if partly known and otherwise underspecified. On the other hand, Situation Theory and formal languages for it, may benefit by incorporation of fuzzy set theory and weighted parametric information, for practical applications.

A promising direction of developments is for relational databases with Situation Theory and semantic representations, by using formal languages of it. This is essential for domains such as health and medical sciences, where semantic information is typically partial, underspecified, and dependent on situations. E.g., situation-theoretic approach introduced here can contribute to enhanced databases systems introduced in Ślęzak et al. [25], and Sosnowski and Ślęzak [26].

## REFERENCES

[1] J. Barwise, "Scenes and other situations," *Journal of Philosophy*, vol. 78, pp. 369–397, 1981.

[2] J. Barwise and J. Perry, *Situations and Attitudes*. Cambridge, MA:MIT press, 1983, republished as [27].

[3] P. Aczel, *Non-well-founded Sets*, ser. CSLI Lecture Notes. Stanford, California: CSLI Publications, 1988, vol. 14.

[4] M. Rathjen, "Predicativity, circularity, and anti-foundation," in *One hundred years of Russelll's paradox (De Gruyter Series in Logic and Its Applications)*, G. Link, Ed. Walter de Gruyter, Berlin, New York, 2004, vol. 6, pp. 191–219.

[5] R. H. Thomason, Ed., *Formal Philosophy: Selected Papers of Richard Montague*. New Haven, Connecticut: Yale University Press, 1974.

[6] K. Devlin, "Situation theory and situation semantics," in *Handbook of the History of Logic*, D. Gabbay and J. Woods, Eds. Elsevier, 2008, vol. 7, pp. 601–664.

[7] R. Loukanova, "Situated Propositions with Constraints and Restricted Parameters," in *Proceedings of the 6th International Workshop on Constraints and Language Processing*, ser. Computer Science Research Reports, P. Blache, H. Christiansen, V. Dahl, and J. Villadsen, Eds., no. 134, October 2011, pp. 44–55.

[8] ——, "Situated Agents in Linguistic Contexts," in *Proceedings of the 5th International Conference on Agents and Artificial Intelligence*, J. Filipe and A. Fred, Eds., vol. 1. Barcelona, Spain: SciTePress — Science and Technology Publications, 2013, pp. 494–503.

[9] ——, "Situation Theory, Situated Information, and Situated Agents," in *Transactions on Computational Collective Intelligence XVII*, ser. Lecture Notes in Computer Science, N. T. Nguyen, R. Kowalczyk, A. Fred, and F. Joaquim, Eds. Springer Berlin Heidelberg, 2014, vol. 8790, pp. 145–170.

[10] ——, "Underspecified Relations with a Formal Language of Situation theory," in *Proceedings of the 7th International Conference on Agents and Artificial Intelligence*, S. Loiseau, J. Filipe, B. Duval, and J. van den Herik, Eds., vol. 1. SCITEPRESS — Science and Technology Publications, Lda., 2015, pp. 298–309.

[11] ——, "Quantification and Intensionality in Situation Semantics," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin / Heidelberg, 2002, vol. 2276, pp. 32–45.

[12] ——, "Generalized Quantification in Situation Semantics," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin / Heidelberg, 2002, vol. 2276, pp. 46–57.

[13] C. Pollard and I. A. Sag, *Information-Based Syntax and Semantics, Part I*, ser. CSLI Lecture Notes. CSLI Publications, 1987, no. 13.

[14] ——, *Head-driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press, 1994.

[15] I. A. Sag, T. Wasow, and E. M. Bender, *Syntactic Theory: A Formal Introduction*. Stanford, California: CSLI Publications, 2003.

[16] A. Copestake, D. Flickinger, C. Pollard, and I. Sag, "Minimal recursion semantics: an introduction," *Research on Language and Computation*, vol. 3, pp. 281–332, 2005.

[17] R. Loukanova, "From Montague's Rules of Quantification to Minimal Recursion Semantics and the Language of Acyclic Recursion," in *Biology, Computation and Linguistics — New Interdisciplinary Paradigms*, ser. Frontiers in Artificial Intelligence and Applications, G. Bel-Enguix, V. Dahl, and M. D. Jiménez-López, Eds. Amsterdam; Berlin; Tokyo; Washington, DC: IOS Press, 2011, vol. 228, pp. 200–214.

[18] Y. N. Moschovakis, "A logical calculus of meaning and synonymy," *Linguistics and Philosophy*, vol. 29, pp. 27–89, 2006.

[19] J. Ginzburg and I. A. Sag, *Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives*. Stanford, California: CSLI Publications, 2000.

[20] M. Van Lambalgen and F. Hamm, *The Proper Treatment Of Events*, ser. Explorations in Semantics. Oxford: Wiley-Blackwell, 2004.

[21] R. Loukanova, "A Formalization of Generalized Parameters in Situated Information," (to appear).

[22] ——, "Russellian and Strawsonian Definite Descriptions in Situation Semantics," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin / Heidelberg, 2001, vol. 2004, pp. 69–79.

[23] R. Yager, "Concept Representation and Database Structures in Fuzzy Social Relational Networks," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 40, no. 2, pp. 413–419, March 2010.

[24] S. K. Pal, L. Polkowski, and A. Skowron, *Rough Neural Computing: Techniques for Computing with Words*. Springer, 2004.

[25] D. Ślęzak, A. Janusz, W. Świeboda, H. S. Nguyen, J. G. Bazan, and A. Skowron, "Semantic analytics of PubMed content," in *Information Quality in e-Health*. Springer, 2011, pp. 63–74.

[26] Ł. Sosnowski and D. Ślęzak, "How to Design a Network of Comparators," in *Brain and Health Informatics*, ser. Lecture Notes in Computer Science, K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe, and N. Zhong, Eds. Springer International Publishing, 2013, vol. 8211, pp. 389–398.

[27] J. Barwise and J. Perry, *Situations and Attitudes*, ser. The Hume Series. Stanford, California: CSLI Publications, 1999.

DOI: 10.15439/2015F283
ACSIS, Vol. 5

# Lessons Learnt From Designing Indoor Positioning System Using 868 MHz Radios and Neural Networks

Michał Meina
Faculty of Mathematics,
Informatics and Mechanics
University of Warsaw
Warsaw, Poland
mail:mich@mimuw.edu.pl

Bartosz Celmer
Section of Computer Science
The Main School of Fire Service
Warsaw, Poland
mail:bart.celmer@gmail.com

Krzysztof Rykaczewski
Faculty of Mathematics,
Informatics and Mechanics
University of Warsaw
Warsaw, Poland
mail:k.rykaczewski@mimuw.edu.pl

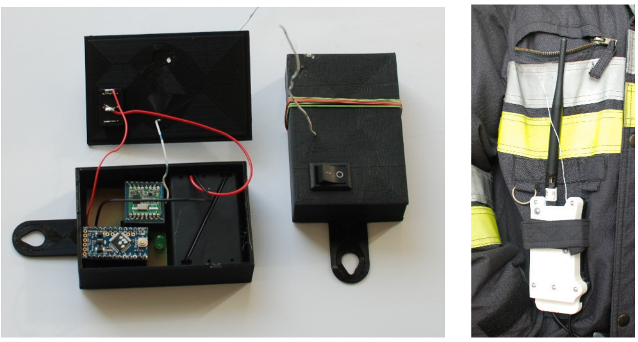
***Abstract*—This paper summarizes our approach and experimental evaluation of infrastructure-based Indoor Positioning System (IPS) designed to be used by First Responders. We are using 868 MHz single channel, power-efficient radio markers and RSSI (Receiver Signal Strength Indicator) fingerprinting. Artificial Neural Network translates vectors of RSSI constructed using mobile units into position. Special preprocessing needs to be applied to on-line signal to construct a vector for classification.**

## I. Introduction

Positioning system in GPS-denied environments (such as large buildings, tunnels etc.) can play crucial role in the Search and Rescue operations. Indoor Positioning Systems (IPSs) can improve not only safety of First Responders, e.g. Smoke Divers who have to search the building and safely withdraw having very limited oxygen supply, but also system like that could assist in decision making and risk management at rescue scene [1] by enhancing situation awareness of the Incident Commander. Yet, reliable information about deployment of resources (both humans and equipment) in the dynamical changing, decision-demanding environment is very challenging. This paper summarizes our approach for infrastructure-based indoor localization designed to be a part of risk management system for Incidents Commanders.

IPS can significantly optimize performance of fire brigade at incident scene in various aspects. Firstly, communication can be significantly enhanced as it was discussed in [4]—basing on interviews with experts and some on-field experiments the number of voice communication could be significantly reduced. Secondly, one of the most significant factor that is known to be source of accident (or near-miss incidents) at a fire scene is the lack of situational awareness[1]. Information about deployment of personnel is a key-factor at incident place. At the same time, it should be advised how to present the information. Amount of information generated at the scene can easily overwhelm [1] the Incident Commander.

Most reliable IPSs involve usage of infrastructure-based techniques, in which various transmitters (or beacons) are deployed in buildings beforehand. That enables us to position the receiver node carried by the subject. Different techniques can be used for such setup (see survey [2] for extensive summary). Following this survey, we have chosen to exploit radio Received Signal Strength Indicator (RSSI) fingerprinting, because of two main reasons: (1) relatively cheap and easy to deploy, and (2) accuracy can be easily tuned by adjusting number of anchor nodes.



(a) Anchor node (b) Receiver node

Fig. 1: Radio Nodes: anchor nodes are deployed at known locations around the building while Receiver node is carried by the First Responder.

RSSI fingerprinting is an empirical technique based on measuring the intensity (strength) of received signal at known positions. Those measurements form features (fingerprints) of signal attenuation of different radios influenced mostly by walls and steel constructions of the building. Positioning can be seen as a problem of classification of incoming signal in order to find “best-match” from known database.

In our setup we also have taken into consideration issues that are specific for search and rescue operation. Firstly,

The research was supported by the Polish National Centre for Research and Development (NCBiR)—Grant No. O ROB/0010/03/001 in the frame of Defence and Security Programmes and Projects: “Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in the buildings”.

[1] *The National Fire Fighter Near Miss Reporting. Annual Report 2008.*

978-83-60810-66-8/$25.00©2015, IEEE

the algorithm needs to be fault-tolerant: damage of a single radio should not increase significantly the positioning error. Secondly, it is known that fire environment could alter signal propagation significantly [3].

The contributions of this paper are as follows:

- experimental evaluation of IPS based on RSSI fingerprinting and Artificial Neural Networks,
- discussion on hardware design of radio markers (anchor nodes).

## II. Related Work

Due to the growing demand for indoor positioning systems, wireless location is an important area of research in recent years. Many studies have been published concerning different types of localization techniques.

Both Personal Dead Reckoning (PDR) and Foot-mounted Pedestrian Navigation System use Inertial Measurement Units (IMUs) for path estimation. However, since the IMU position error accumulate during the procedure of walking, a lot of attention is paid to systems based on a pre-installed infrastructure. Nowadays, the most frequently used technology is based on radios, e.g.: pseudo-satellite transmitters, Radio Frequency Identification (RFID) markers and Ultra-Wideband (UWB) radars. This has many advantages, radio is not limited by the line-of-sight condition as radio signals can penetrate walls and diffract around objects [5].

Several methods have been proposed to estimate the location using sensor networks. Usually, the approach is based on reference nodes (beacons, anchor nodes), which positions are known. The position of the receiver is calculated from the information it receives from the beacons.

The position can be derived from distance estimates between the beacons and receiver node. Most radio receivers in a wireless system have the ability to measure the Radio Signal Strength Indicator (RSSI). This can be later translated to a distance by using a path loss model. Generally, the relation between RSSI and distance is determined by the following formula

$$\mathrm{RSSI}(d) = P - R - 10\alpha \log_{10} d, \tag{1}$$

where $P$ is the transmitted power, $\alpha$ is the path loss exponent which falls linearly and $R$ is a constant that depends on the conditions of the environment, $d$ is the distance from transmitting end [6].

Generally, three main methods are used for the problem of localization: trilateration, multilateration and fingerprinting. Trilateration and multilateration are based on the propagation model, conceptually simpler, but difficult to calculate in a complex environment—firstly distance needs to be estimated accurately, which involves usage of more expensive transceivers (e.g. UWB radio that uses ToF model[2]) and, secondly, environment (and its changes over time) modelling can be challenging. In contrast, fingerprinting is empirical method in which signal attenuation in the building are measured and, therefore neither the signal propagation model is not used nor the building plan does not to be known.

Trilateration technique uses properties of triangles to determine the location, therefore it usually requires at least three access points on the surface. While using this technique precise distance needs to be measured (which usually is not achievable using RSSI). Precise distance is measured using different physical techniques: Angle of Arrival (AoA) is a method that locates the user by measuring the angle of incoming signal, Time of Arrival (ToA) is a technique based on the Time of Flight (ToF). While using this method the clocks of all physical units must be precisely synchronized and clock drift compensated. All this makes the final system more complex.

Multilateration is a navigation technique based on measurement of the difference between the distances to two or more stations located at known locations that transmit a signal in the indicated times [7]. It differs from trilateration in that it does not use absolute measurement of Time of Flight, but its differences (TDoA, Time Difference of Arrival). Position is then estimated by the intersection of hyperboloids which are places consisting of points having equal TDoAs. In this case, the problem can be represented as an optimization problem and solved using, for example, the method of the least squares or gradient descent method.

Due to the fact that certain signals can be disturbed by presence of obstacles, some extensions (like, e.g., multiwalls model) to above-mentioned methods were introduced.

Fingerprinting is another method frequently used in indoor positioning. In this technique radio signal strength is measured at different locations beforehand. During the first (training or off-line) stage signal strength data is collected is the physical location (usually up to $50 \times 50\, m^2$) to the training/labelling database, or to a non-linear mapping. In the second (on-line) stage of the mobile unit measures RSSI and compares its value to values held in the database. In result location with similar matching is returned. The location of the fingerprint technique requires an adequate number of reference devices and stable environment before calibration, because the result is sensitive to environmental changes, such as moving objects in a building that may have an impact on the properties of the signal. Fingerprinting can obtain good performance, since the noise arising from all obstacles is already included in the map. Therefore, we do not have to add to it any additional model.

The widely used basic matching algorithm used in fingerprinting is the $k$-Nearest Neighbour ($k$-NN) [8]. In the on-line positioning step the $k$-NN algorithm is searching for $k$ neighbour closest (in the sense of the Euclidean distance) between classes of fingerprint database and the real-time RSSI values to determine the location. The Support Vector Regression (SVR) [9] as well as, Artificial Neural Network (ANN) are in widespread use as well [10], [11], [12], [13]. Comparison of different architectures of neural networks can be found in [12].

Moreover, there are attempts to combine such estimation with dead-reckoning navigation using foot-strapped inertial measurement units [14].

[2] http://www.decawave.com/

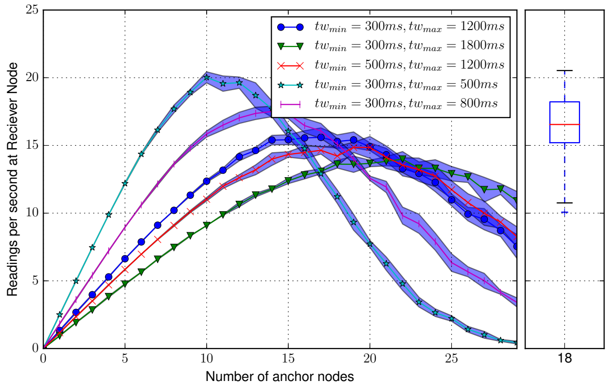


Fig. 2: Simulation of radio sampling vs density of radio displacement. On the right results for empirical data (for $tw_{min} = 300$ ms and $tw_{max} = 1200$ ms) is enclosed.

## III. Hardware setup

We use transceivers based on RFM22 Hope Microelectronics co. silicon with 17 cm coper wire antenna (half-wave long). RFM22 have ability to work in very different modes (modulation, frequency, transmit power). Additionally, our setup involves computations of neural network on mobile device (Odroid-U3), sending it via ZigBee 867 Mbps (2.4 GHz) and displaying actual positioning using Recon Jet head-mounted display[3].

The relatively low frequency (868 MHz) was chosen because of the high penetration ability of signal comparing to power consumption. Also, the noise from different devices operating in this particular frequency band is expected to be smaller (comparing to e.g. 2.4 Ghz). It is worth to note that higher penetration gives us possibility to build sparse node networks, which drastically lowers the cost of overall system. Another issue is signal attenuation in smoke and fire environment which is known issue (see [3]) but there is not enough comparable results to choose the best operating band for this purpose.

Two type of radio nodes was constructed: anchor node (see Fig. 1a) and receiver node (Fig. 1b). Anchors send periodically small portions of data including identifier and message number. Receiver Node is gathering those packages while establishing RSSI and reports it to processing unit.

Configuration of RFM22 was as follows: 868 MHz frequency transmission band, FSK modulation without Manchester encoding (error detection technique—disabled for shorter time of transmission) and +17 dBm mode (transmit power). During initial test we confirmed sufficient wall penetration and expected RSS loss.

Every anchor node operates on exactly the same frequency and, because of that, two radios which transmit their signal can drown each other. Two or more radios that are in mutual coverage area cannot transmit their data in the same time. Therefore, transmission synchronization needs to be performed to overcome problem of mutual jamming. Nevertheless, direct clock synchronization is very complicated in Wireless Sensor Networks (see [15] for overview of the problem), especially in indoor environment (where GPS-based synchronization is unavailable).

Straightforward, node-independent mutual jamming prevention technique based on randomized transmission was implemented. Node number $i$ transmits its mark which usually lasts for about 15 ms, and then radio goes into sleep mode for $T_i$ ms. Idle time is picked randomly after each transmission from the interval

$$tw_{min} < T_i < tw_{max}. \qquad (2)$$

This way, idle time is long enough to allow other radios to transmit their data and short enough to retransmit packet, if it was dropped while mobile is not moved far. Due to the high noise in RSSI estimation it is important to get as many readings as possible. The data is later preprocessed using moving window technique (see Section V-A).

Figure 2 shows simulation result for selected values of $tw_{min}$ and $tw_{max}$ with regard of the number of anchor nodes. Increased number of anchor nodes obviously lead to increased sampling rate at Receiver Node. This, however, can be done only to some extent, after which sampling rate is degrading (because of the collisions in transmission). Peeks at Fig. 2 indicates more or less "optimal setup". Having estimated number of nodes that can jam each other on deployed building interval of sleeping time should be adjusted using this simulation.

Sampling rate estimated from empirical data (which was $16.3 \pm 2.5$ rps) in our experiment is higher than expected (depicted by the box plot on the right-hand side of Fig. 2). In our experimental deployment (see Fig. 6) distant nodes are on the edge communication reach and, therefore, they have limited possibility to drown each other.

## IV. System Description

Overall system processing schema is illustrated in Fig. 3. Operation of IPS that is based on fingerprinting is divided into two phases: off-line and on-line. In the first fingerprints are collected and learning procedure is performed. In on-line phase, on the other hand, incoming signal is processed "on-the-fly" and algorithm outputs position.

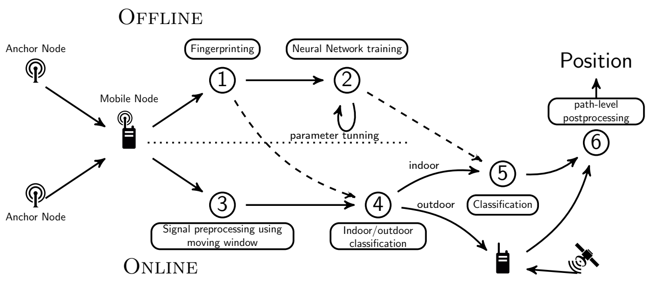


Fig. 3: Processing steps of the Localization Algorithm

At step ① fingerprints are collected (numbers in circles refers to Figure 3). Radio signal strength is recorded using

[3] http://www.reconinstruments.com/products/jet/

mobile device at predefined places (with known GPS position) and used to construct a feature vector (fingerprint) for particular point.

In the on-line situation at the beginning (③) RSSI signal needs to be preprocessed to create a test vector for ANN. Due to the fact that RSSI values from different radios did not arrive at the same time, sliding window technique is exploited—due to the small sampling rate fast movements of mobile unit can degrade precision of test vector formulation (windowing technique introduces lags). Later we remap entries of this vector to the interval $[0,1]$ in order to use the neural network at step ②.

Node ④ is used to resolve the method that will be used for positioning. It turns out that the simple thresholding of signals is inadequate, because some radios can be transmit with lower RSSI value outside than inside. However, it is true that the signal strength with decrease with the distance from the building. Indoor/outdoor detection is beyond the scope of this paper.

Classification (mapping RSSI to position) is performed at step ⑤. Just after it simple mapping to GPS coordinates is performed.

At the end ⑥ path needs to be post-processed (expected high RSSI noise introduces a lot of distortions on path). Kalman Filter is a good candidate for solving this problem, because of the motion model that can be expressed by it.

## V. Neural network-based fingerprinting

This section describes our framework for IPS based on Artificial Neural Network.

Most commonly used type of Artificial Neural Network (ANN) consists of several layers: the input layer is connected to layers of hidden units, which provide information to the output layer. For learning ANN the most commonly used method is the backpropagation algorithm. It tries to adjust weights of each neuron in order to reduce the error between the desired and calculated output. In this way, the neural network learns how to map input to output. The aim of the network is not only to restore the training input data but also to generalize the data to new situations (by interpolating capabilities). The number of input nodes and hidden layers depends on the design issue and depends on the number of base stations deployed in the environment.

### A. Input data preparation and training set construction

The data collection process involves marking of the reference points on the floor and making measurements for a 30–60 s. All points laid on the same plane. In this way at every point we received a number of recordings consisting of RSSI which comes from different radios. Therefore, we obtained a RSSI fingerprint log consisted of triplets

$$(i, \rho_i^{j,k}, P_j), \tag{3}$$

where $i = 1,\ldots,N$, denotes radio number, $P_j = (\text{latitude}, \text{longitude}, \text{elevation})$ is a position of $j$th fingerprint and $\rho_i^{j,k}$ is a $k$th RSSI value recorded at point $P_j$.

Since all radios may not be visible at once (due to interference and momentary jamming of radios), in order to obtain a vector of RSSI signals from all radios (which will be used as an input to network) we had to aggregate recordings at a given point. Therefore, as input corresponding to point $P_j$ and reading $k$ we took set

$$\left\{\operatorname{avg}\left(\{\rho^{j,k_l}\}_{l=1}^{K_j/2}\right) \mid M \text{ rand } \{k_l\}_{l=1}^{K_j/2} \subset \{1,\ldots,K_j\}\right\}, \tag{4}$$

i.e. we averaged random subsamples of recordings (by taking half of the recordings) for point $P_j$. In order to get rid of the noise we do it $M$ times. Therefore, $M$ can be interpreted as an aggregation (folding) parameter.

Let us define point signals in time $t$ from $N$ radios in $j$th point by the following

$$\rho^j(t) = [\rho_1^j(t), \ldots, \rho_N^j(t)]. \tag{5}$$

Such signals consist of RSSI recordings aggregated and averaged as above.

Due to issues described below, which are related to neural networks, we need to map GPS coordinates using affine scaling into interval $[0,1]$. Similarly, RSSI values from radio are converted into $[0,1]$.

### B. Network architecture

Neural Network that we use for IPS is depicted in Fig. 4. It consists of input layer constructed from RSSI vector for actual reading, output layer denotes position in three dimensional space (latitude, longitude and elevation) and the $L$ number of hidden layers. We have used sigmoid activation function, therefore units in input and output vectors need to be mapped by scaling into interval $[0,1]$.

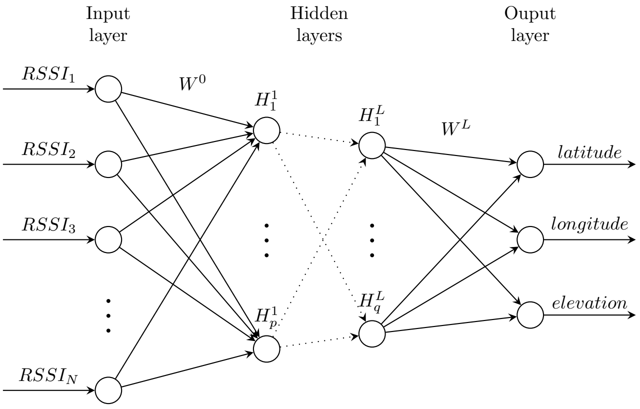


Fig. 4: Neural network setup.

The number of hidden layers and number of neurons at each layer should be tuned accordingly to specific task. Nevertheless, it is important to say that this particular architecture depends significantly on number of anchor nodes used for IPS system. It is hard to estimate time complexity of the learning algorithm. Estimation for the worst case scenario is

$$MNL\prod_{i=0}^{L}\#W^i, \tag{6}$$

where $\#W^i$ is the number of weights in layer $i$ and $L$ is the number of layers in the network. We see that it depends strictly on the size of the building, since additional reference nodes need to be added.

Learning takes quite long, depending on the permissible error. For example, for our problem with permissible error equal to 0.01 and folding parameter equal to 20 (which means that the number of samples grows by the factor of 20) it took between 4 and 10 h on a single core. The number of iterations was set to 10000. It is important to notice that quality of network classification depends on the number of iterations and goal set.

### C. On-line signal classification

Due to the possible interference and temporary signal deficiencies input data in on-line classification need to be preprocessed first. During the movement the mobile unit can receive readings from different anchors in different times (radio reading are sparse and unevenly sampled), therefore we used sliding window technique to receive feature vector at given time.

In order to impute missing values of signal strength we can use local linear approximation as follows

$$\begin{aligned}\rho(t+\Delta t) &= \rho(t) + \rho'(t)\Delta t + \mu \\ &= \rho(t) + \beta d'(t)\log_{10} d(t) + \mu, \end{aligned} \tag{7}$$

where $\beta$ is some constant and $\mu$ denotes higher order terms with small magnitude. Assuming that locally velocity of the subject is constant ($d'(t) = v_{\mathrm{loc}} = \mathrm{const}$) we use it to smooth the signal. The result is shown in Fig. 10 and was not so impressive as supposed. Therefore, we used linear regression.

Given fingerprints as in (3) we want to obtain input signal for the network as in (5).

Since at given time not all radios may be visible we define moving window $W$ of length $\tau$ for time series $s$ (till time $t$) as

$$W_\tau(s)(t) := [s(t-\tau), \dots, s(t)] \tag{8}$$

which will collect the signal strength in a short period.

To this end, we performed moving average on RSSI from the last two seconds. If there was no radio signal from a given radio in the given window, then we put 0. In short,

$$\hat{\rho}(t) := \big[\mathrm{MA}_k\big(W_\tau(\rho_1)(t)\big), \dots, \mathrm{MA}_k\big(W_\tau(\rho_N)(t)\big)\big], \tag{9}$$

where $\mathrm{MA}_k$ stands for moving average operator of length $k$. Signal prepared in this way can be used as an input to the neural network. For post-processing we used the Kalman filter. Therefore, path are smoother and better corresponds to reality.

### D. Network architecture tuning

We found out that standard heuristics, like taking two hidden layers with number of neurons in the second layer being half of that in the first layer (see [16] for more information) works quite fine. We used very large first layer since we need embed data in high dimension and obtain overfitting in order to discriminate it. Next layer is smaller but we get better generalization properties and avoid overfitting. Such architecture can be obtained when analysing in detail Fig. 7.

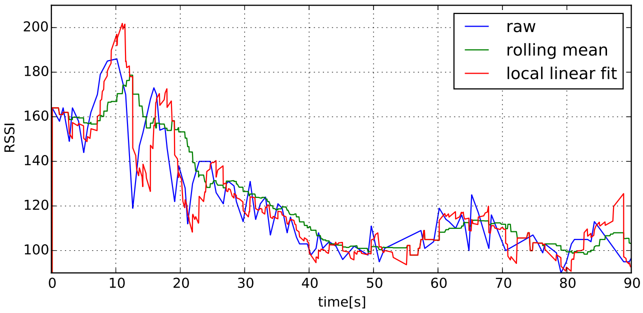


Fig. 5: RSSI (anchor number 4) while moving (only showing first 1.5 min of the recording). Rolling mean (moving average) have very good smoothing capabilities but it introduces certain lag while local linear regression seems to work on-line but does not lead to errors cancellation.

## VI. Experimental Section

We conducted experiment using 18 anchor nodes deploying them on the one floor of approximately 30 × 30 m building. Fig. 6 shows displacement of radios and fingerprints. Note that the data was collected only in selected rooms and passages. Additionally, the figure show bicubic interpolation of RSSI signal strength for 5th radio basing on fingerprint features.

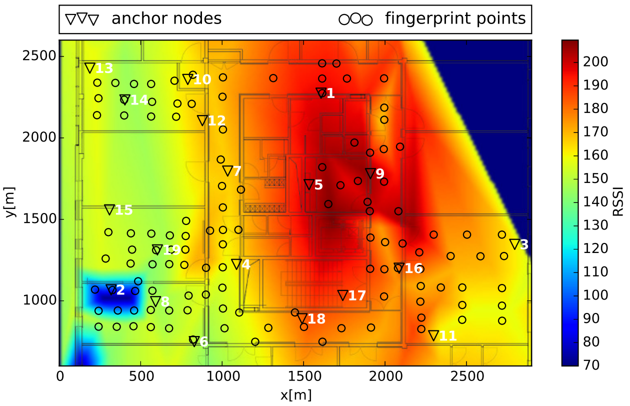


Fig. 6: Interpolated RSSI map for 5th radio. Blue colour in the right upper corner means that there were no data.

We performed collection at 119 points gathering 5991 records (samples) overall. Each position was recorder for about $18.1 \pm 4.2$ seconds which gives us $297.3 \pm 90.2$ records on average. Points were not distributed uniformly, but they were chosen in such a way that there were no large area without samples. Moreover, significant error was introduced while we manually collected GPS positions of reference points.

Two experiments were conducted: stationary position estimation and movement/path reconstruction. For training the neural networks we have used folding parameter $M = 20$. It may seem rather small, but that enables us to perform parameter sweeping through different network architectures in reasonable time (see Fig. 7).

### A. *Position in stationary points*

In the first experiment we tested position estimated at stationary points (without movement). Recorded fingerprints were divided into two separate sets in ratio 70/30. Points was chosen manually and are depicted in Fig. 6. Test sample was prepared according to Equation (4).

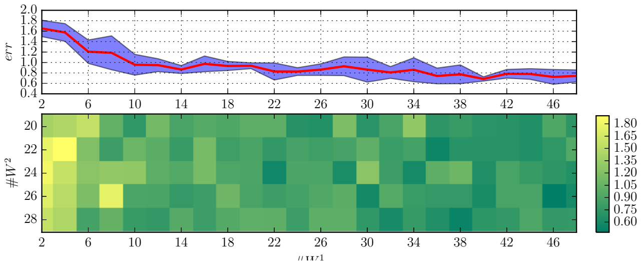


Fig. 7: Position estimation error using ANN with $\#W^1$ and $\#W^2$ neurons on the first and the second layer respectively (lower figure). Upper figure shows column-wise summed up results for $H^1$ layer.

As it can be seen, our heuristics works quite good. For example, when there is about 10–14 neurons in the first layer the average error falls below 0.5 m. Larger networks do not guarantee an improvement of the results (even though some of them were significantly better).

### B. *Paths*

Second experiment was focused on path reconstruction. We asked a subject to perform a walk-through the building with receiver node. Results are depicted in Fig. 8–10. True path was marked on figures using video camera recordings.

As it was already written, most of the errors in the motion is introduced by the low frequency of refreshing rate and, thus, there is the necessity to use of sliding windows.

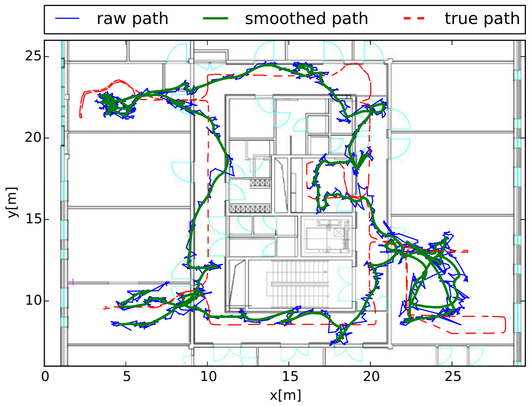


Fig. 8: Path for network with [18, 13, 21, 3] perceptrons in layers. Moving average size of $k = 5$ s.

The differences on the path are more apparent, where overfitting dominates interpolation ability of the network. For example, the path in the corridor, surrounded by a small amount of training points, is better suited than in other places.

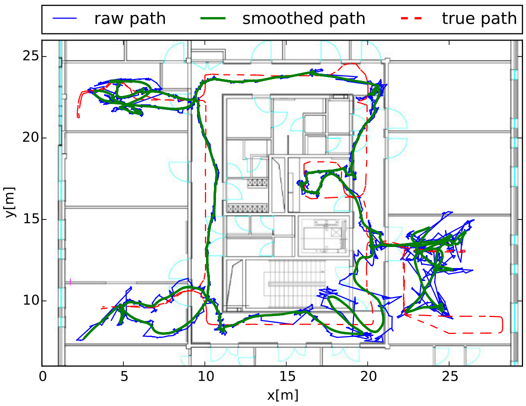


Fig. 9: Path for network with [3, 40, 26, 3] perceptrons in layers. Moving average size of $k = 5$ s.

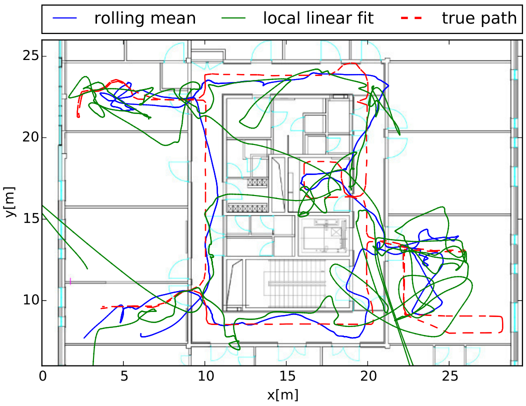


Fig. 10: RSSI smoothing comparison.

In smaller networks we can see that path is being pulled off (Fig. 9).

Another issue is the necessity of usage of sliding window technique—it introduces a certain lag into the feature vector construction. Simply speaking, different RSSIs are recorded at different places therefore we do not know signal strength precisely in time. The problem is illustrated on both previous mentioned figures at the path in lower-right room.

Our initial idea was to use forecasting on RSSI signal that can minimize lagging issues with standard rolling window techniques. Ability of forecast the signal strength lies on the assumption that RSSI is dependent on motion (acceleration and decelerations are not random), see Equation (7). Therefore, we tested the idea using local linear regression instead of rolling mean. Results are depicted in Fig. 10 and show clearly that this simple forecasting is not working due to the large impact of noise to path estimation. Nevertheless, the idea of compensating the signal sampling with motion prediction is worth pursuing in the future works.

### C. *Discussion*

Fingerprint coverage does not have to be dense (neural networks have very good interpolation capabilities), but we noticed a problem with the estimation of places “on edges,” where the paths are pulled towards more places with more dense fingerprint coverage. Some special case should be applied in order to overcome that.

It is worth noting that more anchor nodes not necessarily means better performance—sparse node deployment allow higher signal sampling rate and, therefore, allow better estimation of high-velocity motions. Observe, however, that we cannot assume that radios have different sampling, because we do not know whether the situation when radio sees only few neighbours will change in the future. During fire some wall may broke down and opens new way for the signal.

We also noticed a negative correlation between the velocity of movement of the subject and location accuracy. This is obviously due to the fact that we get only a few samples of RSSIs for a given position. In result the path can oscillate around the correct location.

Moreover, it can be observed that iron oxygen cylinder carried by the firefighters can distort the signal very badly. For example, there were situations where orientation (rotation along the axis) lead to tremendous improvement of the position.

## References

[1] A. Krasuski, A. Jankowski, A. Skowron, and D. Ślęzak, "From sensory data to decision making: A perspective on supporting a fire commander," *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 3, pp. 229–236, 2013.

[2] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 6, pp. 1067–1080, Nov 2007.

[3] C. M. Dissanayake, M. N. Halgamuge, K. Ramamohanarao, B. Moran, and P. Farrell, "The signal propagation effects on ieee 802.15.4 radio link in fire environment," in *5th International Conference on Information and Automation for Sustainability (ICIAFs), 2010*, Dec 2010, pp. 411–414.

[4] M. Scholz, D. Gordon, L. Ramirez, S. Sigg, T. Dyrks, and M. Beigl, "A concept for support of firefighter frontline communication," *Future Internet*, vol. 5, no. 2, pp. 113–127, 2013. [Online]. Available: http://www.mdpi.com/1999-5903/5/2/113

[5] V. Honkavirta, T. Perala, S. Ali-Loytty, and R. Piché, "A comparative survey of WLAN location fingerprinting methods," in *Positioning, Navigation and Communication, 2009. WPNC 2009. 6th Workshop on*. IEEE, 2009, pp. 243–251.

[6] C. Papamanthou, F. P. Preparata, and R. Tamassia, "Algorithms for location estimation based on RSSI sampling," in *Algorithmic Aspects of Wireless Sensor Networks*. Springer, 2008, pp. 72–86.

[7] S. Adler, S. Schmitt, Y. Yang, Y. Zhao, and M. Kyas, "Experimental evaluation of indoor localization algorithms," in *2014 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2014, 27th–30th October 2014*, 2014, pp. 1–9.

[8] F. Lemic, A. Behboodi, V. Handziski, and A. Wolisz, "Experimental decomposition of the performance of fingerprinting-based localization algorithms," in *2014 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2014, 27th–30th October 2014*, 2014, pp. 1–10.

[9] W. Farjow, A. Chehri, H. T. Mouftah, and X. N. Fernando, "Support vector machines for indoor sensor localization," in *Wireless Communications and Networking Conference (WCNC), 2011 IEEE*, March 2011, pp. 779–783.

[10] S. Outemzabet and C. Nerguizian, "Accuracy enhancement of an indoor ann-based fingerprinting location system using particle filtering and a low-cost sensor," in *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*. IEEE, 2008, pp. 2750–2754.

[11] R.-C. Hwang, P.-T. Hsu, J. Cheng, C.-Y. Chen, C.-Y. Chang, and H.-C. Huang, "The indoor positioning technique based on neural networks," in *Signal Processing, Communications and Computing (ICSPCC), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–4.

[12] L. Yu, M. Laaraiedh, S. Avrillon, and B. Uguen, "Fingerprinting localization based on neural networks and ultra-wideband signals," in *Signal Processing and Information Technology (ISSPIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 184–189.

[13] M. Altini, D. Brunelli, E. Farella, and L. Benini, "Bluetooth indoor localization with multiple neural networks," in *Wireless Pervasive Computing (ISWPC), 2010 5th IEEE International Symposium on*, May 2010, pp. 295–300.

[14] M. Nilsson, J. Rantakokko, M. A. Skoglund, , and G. Hendeby, "Indoor positioning using multi-frequency RSS with foot-mounted INS," in *2014 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2014, 27th–30th October 2014*, 2014, pp. 1–10.

[15] I.-K. Rhee, J. Lee, J. Kim, E. Serpedin, and Y.-C. Wu, "Clock synchronization in wireless sensor networks: An overview," *Sensors*, vol. 9, no. 1, pp. 56–85, 2009. [Online]. Available: http://www.mdpi.com/1424-8220/9/1/56

[16] S. Walczak and N. Cerpa, "Heuristic principles for the design of artificial neural networks," *Information and Software Technology*, vol. 41, no. 2, pp. 107–117, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950584998001165

DOI: 10.15439/2015F388
ACSIS, Vol. 5

# Rule Quality Measures Settings in a Sequential Covering Rule Induction Algorithm - an Empirical Approach

Marcin Michalak*, Marek Sikora*†, Łukasz Wróbel†*
*Institute of Informatics, Silesian University of Technology
ul. Akademicka 16, 44-100 Gliwice, Poland
Email: {Marcin.Michalak, Marek.Sikora}@polsl.pl
†Institute of Innovative Technologies EMAG
ul. Leopolda 31, 40-189 Katowice, Poland
Email: {Marek.Sikora, Lukasz.Wrobel}@ibemag.pl

*Abstract*—**The paper presents the results of research related to the efficiency of the so called rule quality measures which are used to evaluate the quality of rules at each stage of the rule induction. The stages of rule growing and pruning were considered along with the issue of conflicts resolution which may occur during the classification. The work is the continuation of research on the efficiency of quality measures employed in sequential covering rule induction algorithm. In this paper we analyse only these quality measures (9 measures) which had been recognised as effective based on previously conducted research.**

## I. Introduction

THE SEQUENTIAL covering rule induction algorithms can be used both for classification and descriptive purposes [1, 2, 3, 4, 5, 6, 7]. The main principle of the sequential covering rule induction algorithms is unchanged, despite the development of increasingly sophisticated versions of such algorithms. It involves the induction of rules in two phases: growing phase and pruning phase. In the growing phase, the elementary conditions occurring in the rule premise are specified. In the pruning phase, some of these conditions are removed. In comparison with other induction methods, rule sets obtained by the covering algorithms are characterised by good classification as well as descriptive capabilities. Taking into consideration only the classification abilities, better results can be often obtained using other methods, for example neural-fuzzy networks [8, 9] support vector machines [10], or ensemble of classifiers [11]. Models obtained in this way, however, are characterised by much less interpretability than classification (decision) rules. In the case of rule learning for descriptive purposes, the algorithms for induction of rules satisfying certain minimum quality criteria are most commonly used [12, 13, 14]. Induction of all rules that satisfy the given quality criteria leads to the induction of a very large number of rules which then must be limited by the filtering algorithms using so-called rule interestingness measures [15, 16, 17]. If the primary objective is to describe data, it is possible to use modified covering algorithms (so-called rule-based subgroup discovery [18, 13, 19]). These algorithms aim at the induction of statistically significant rules which cover regions in the feature space of as small union as possible. Rule sets generated in this way are characterised by worse classification abilities than in the standard covering algorithms. If the analysis objective is to create a classification system that uses an interpretable data model, application of sequential covering rule induction algorithms is the most sensible solution. The quality of the rule set obtained by the covering algorithm depends on the quality measure [20, 1, 21, 22, 23, 24, 25, 26] used in the growing and pruning phases. The used quality measure is one of the factors affecting the classification accuracy, the number of rules induced and their other characteristics (e.g. the statistical significance).

The objective of the paper is to present the results of research related to the efficiency of using the combinations of different measures at each stage of the rule induction. The sequential covering approach and the top-down method were applied as the basic induction algorithm. In this approach one can distinguish the stages of the rule growing and pruning. In these stages it is possible to use different optimisation criteria which control the rule creation process. These criteria are commonly called rule quality measures or search heuristics. Different quality measures can also be used at the stage of classification conflicts resolution - this way the mechanism of weighted voting gets modified. The paper features an analysis of the combination of 9 measures which were recognised as effective (this choice will be explained further in the paper). The analysis was conducted based on 34 data sets. In addition, the paper contains some supplementary information about the applied quality measures.

## II. Related Works

Measures used for the rule evaluation are divided into two categories [16]: objective and subjective measures. Most of the objective measures are defined on the basis of the contingency

This work was supported by Polish National Centre for Research and Development (NCBiR) grant PBS2/B9/20/2013 in frame of Applied Research Programmes.

table that characterises the quality of a rule in the context of a fixed (usually training or validation) data set [27, 23]. The task of the subjective measures is to evaluate the rules according to subjective user's preferences. The subjective measures evaluate such features of rules as: unexpectedness, surprisingness, novelty, utility or actionability [16]. The method of measuring these features depends on the application domain and the purpose of analysis. Our range of interest includes objective rule quality measures, also known as evaluation or search heuristics [22, 28, 29]. At present, in the literature over 50 objective measures can be found [20, 22, 15, 23, 28, 24, 26, 29]. A lot of measures used for evaluation of decision (classification) rules have wider significance - they function also as rule interestingness (attractiveness) measures [15, 30, 16, 17, 31]. Interestingness measures are used for evaluation of already induced rules, both decision and association. In the next part of the paper we will use the common term quality measures. The quality measure applied in decision rule induction is significant for the quality of an output set of rules. This is confirmed by empirical research [32, 20, 21, 23, 25, 26].

In [22, 23] Fürnkranz and Janssen conducted an analysis of the efficiency of measures which contain one or several parameters. These authors present a methodology to adapt the values of parameters contained in quality measures to the specifics of the analysed data set. In addition, they demonstrate that, as the parameters change, certain measures begin to behave similarly to others (such an analysis was conducted for m-estimates and Klösgen measures). In our articles [26, 33], in turn, we are analysing the measures which do not contain parameters and we demonstrate their efficiency on a data set similar to the one quoted in Fürnkranz's work. However, we use an algorithm which generates the rule set instead of a rule list. This results in the necessity to use the mechanism for classification conflicts resolution, i.e. there is another place (apart from rule growing and pruning) where a rule quality measure can be applied. This way it is possible to achieve better accuracy of the classification, particularly higher specificity and sensitivity of the classifier [33].

There are also works that deal with the decision rule length. In the paper [34] the analysis of relationships between length and the coverage of the decision rule is described. Relationships are considered as the influence of the length on the rule coverage and also as the reversed: the influence of the rule coverage on its length. In the other paper [35] a Dominance Based Rough Set Approach generated rules, and their length, determines the relevance of attributes, what becomes a criterion for a feature selection. There was also an attempt to characterise rule induction as a process specified by domain ontology [36], describing all data mining algorithms.

## III. Rule Induction

Let us assume that a finite set $Tr$ of examples is given. Each training example is described by a set $A \cup \{d\}$ of features, i.e. $a : Tr \to Va$ for each $a \in A \cup \{d\}$. The set $Va$ is called the range of the attribute $a$. Elements of $A$ are called conditional attributes, the variable $d$ is called a decision attribute, and its value is identified with the assignment of an example to a specific concept (decision class). Conditional attributes can be of symbolic (discrete-valued) or of numeric (real-valued) type. The decision attribute is of symbolic type. Each example $x$ belonging to the $Tr$ set can be written as a vector $x = (x_1, x_2, \ldots, x_{|A|}, y)$, where $a_i(x) = x_i$ for each $i \in \{1, 2, \ldots, |A|\}$ and $y = d(x)$. The conditional expression of the following form is called a decision rule:

$$IF\ w_1\ AND\ w_2\ AND\ \ldots\ AND\ w_k\ THEN\ d = v$$

An example satisfying all elementary conditions $w_i$ is assigned to the concept indicated in the rule conclusion. Construction of elementary conditions $w_i$ may be various and depends on a rule induction algorithm. Any elementary condition is most often the expression of the one of the two following forms: $w_i \equiv a_i\ op\ Z_i$, where $a_i$ is the name of the conditional attribute, $op$ is one of the relation operators $\{=, \leq, \geq, >, <\}$, $Z_i$ is the element of the range of the attribute or just $w_i \equiv a_i \in Z_i$ where $Z_i$ is a interval in the range $Va_i$: $Z_i \subset Va_i$. The positive examples are those belonging to the decision class pointed in the rule conclusion. The negative examples are the remaining ones.

Let $p$ denote the number of positive examples covered by the rule ($P$ stands for all positive examples in the training set), and let $n$ denote the number of negative examples covering the rule ($N$ stands for all negative examples). On the basis of this notation a contingency table for the rule can be built (Table I).

TABLE I
The contingency table for the rule covering $p$ positive and $n$ negative examples.

| | | |
|---|---|---|
| $p$ | $n$ | $p + n$ |
| $P - p$ | $N - n$ | $P + N - p - n$ |
| $P$ | $N$ | $P + N$ |

The aim of majority of rule induction algorithms is to find the minimal set of classification rules which cover and correctly predict decision classes of a given set of examples and which additionally are characterised by high values of $p$ and low values of $n$. Finding the optimal solution for such a problem is a computationally expensive task, therefore most of the rule induction algorithms use some heuristics. One of the most common approaches, used also in our implementation, is the sequential covering (known also as separate-and-conquer) strategy [4, 6].

To put it briefly, this strategy consists in learning a rule which covers some part of training examples. Next, the examples covered by the learnt rule are removed from the training set and the rule learning process starts recursively for the remaining examples.

Our implementation of the rule induction algorithm also works in the separate and conquer fashion. The outcome of the algorithm is a set of rules describing each decision class of a training set. The process of induction of a single rule consists of two phases: growing and pruning. In the growing

```
RuleInduction(examples, ruleQualityMeasure)
the set of generated rules (initially empty)
ruleSet := $\emptyset$
decisionClass is a set of examples which have the same value of
decision attribute
foreach decisionClass in examples do
    uncoveredPositives := decisionClass
    while uncoveredPositives $\neq \emptyset$ do
        rule with empty antecedent and consequent pointing current
        decision class
        rule := $\emptyset$
        the set of examples covered by the rule (initially empty rule
        covers all examples)
        covered := examples
        rule growing phase
        do
            conditions := PossibleElementaryConditions(covered,
            uncoveredPositives)
            bestQuality := $-\infty$
            bestCondition := $\emptyset$
            foreach c in conditions do
                evaluate the quality of rule with c condition added
                to its antecedent
                quality := Evaluate(rule c, ruleQualityMeasure)
                if quality > bestQuality then
                    bestQuality := quality
                    bestCondition := c
                end
            end
            add selected elementary condition to the antecedent of
            the rule
            rule := rule $\cup$ bestCondition
            covered := Covered(rule, examples)
        while until (stop criterion);
        rule pruning phase
        Prune(rule, ruleQualityMeasure)
        covered := Covered(rule, examples)
        uncoveredPositives := uncoveredPositives $\setminus$ covered
        ruleSet := ruleSet $\cup$ rule
    end
end
return ruleSet
```

**Algorithm 1:** The algorithm of rule induction

phase, the elementary conditions are added one by one to the premise of the rule. In the case of nominal attributes, the elementary condition can take the form of $(a = v)$, and for the numerical attributes it can take one of two forms: $(a < v)$ or $(a \geq v)$. For the numerical attributes the value v is the arithmetic mean between two successive values from the range of attribute $a$. The set of all the possible elementary conditions which might be added to the rule is created on the basis of examples currently covered by the rule. It means that the domains of the attributes are narrowed to the values which are taken by the examples covered by the currently formed rule. Moreover, the elementary condition is tested only if its addition to the rule causes that such a refined rule covers at least one positive example not covered by rules generated so far. The refinement that has the highest value of the specified rule quality measure among all refinements possible in a single step is selected as the final one. If several refinements obtain the same value of the rule quality measure, the refinement covering more uncovered positive examples is chosen.

During the induction of rules it should be taken into account that some attribute values of individual examples are not always known. In the literature, several propositions of strategies for handling unknown values of the attributes in rule learning algorithms can be found [37]. In our implementation it is assumed that the elementary condition always returns truth value false for the unknown value of the attribute. It means that if the example has unknown values, it can be covered only by the rule which does not contain attributes with unknown values for this example.

The process of rule growing is terminated if the rule is accurate (i.e. it covers none of the negative examples) or if the addition of the next conditions to the rule no longer increases its precision (which may take place if examples with identical attributes values but different decisions exist in the training set). After the rule growing phase, the rule is pruned. The rule pruning algorithm uses a hill-climbing strategy. At each iteration, it deletes the elementary condition without which the rule has the highest improvement in the value of quality measure. After each removal, the value of rule quality measure is recalculated on the entire set of training examples and the deletion of conditions is repeated until it does not cause decrease in the current value of quality measure. The pruned rule is added to the final set of rules, and then the process of rule induction starts again for the rest of the uncovered positive examples.

The assumed heuristic of rule building therefore focuses on induction of rules characterised by high values of specified quality measure and thus constitutes a generic framework for studying behaviour of various quality measures.

## IV. Rule Quality Measures

Quality measures strongly associated with decision rules evaluation are collected in surveys [27, 21, 23, 28, 29].

The papers [27, 21] focus mainly on studying measures influence' on conflicts resolution during classification. Janssen and Fürnkranz [23] present research on parametrised quality measures and the procedure of data driven selection of their parameters. Papers by Yao and Zhong [29] and Lavrac, Flach, and Zupan [28] focus on defining quality measures in the language of probability. Interestingly enough, the mentioned authors' considerations concern disjoint sets of measures.

Our choice of nine decision rule quality measures is provided in Table II. The selection is a result of a previous experiments — in already quoted works [26, 33], about 30 measures were analysed with respect to their efficiency. To check the efficiency, the measures were used in the rule induction algorithm - the same measure was used at each induction stage and during the resolution of classification conflicts. Then the classification abilities of the obtained rule classifier were checked. The classification abilities were analysed from the point of view of overall classification accuracy and balanced accuracy. In the case of two decision classes and the discrete classifier, this measure is equivalent to the AUC criterion [38]. In the case of a bigger number of classes, the balanced accuracy informs about the average accuracy of each decision class. This measure is suitable to verify the classifier which works on unbalanced data. Additionally, the number of rules was analysed with respect to the interpretability of the obtained rule sets. Without going deeper in the measure selection methodology, which is presented for example in [33], 9 evaluation measures were selected. These measures achieved good results, at least from the point of view of one of the three above mentioned criteria - statistically they performed better than other measures in the majority of data sets.

TABLE II
THE LIST OF SELECTED OBJECTIVE RULE QUALITY MEASURES.

$$\begin{aligned}
\mathbf{g} &= \frac{p}{p+n+2}\\
\mathbf{wLap} &= \frac{(p+1)(P+N)}{(p+n+2)P}\\
\mathbf{LS} &= \frac{pN}{nP}\\
\mathbf{Rss} &= \frac{p}{P}-\frac{n}{N}\\
\mathbf{MS} &= \frac{p}{n+P}\\
\mathbf{C1} &= Coleman\cdot\frac{2+Cohen}{3}\\
\mathbf{C2} &= Coleman\cdot 0.5\left(1+\frac{p}{P}\right)\\
\mathbf{corr} &= \frac{pN-Pn}{\sqrt{PN(p+n)(P-p+N-n)}}\\
\mathbf{s} &= \frac{p}{p+n}-\frac{P-p}{P-p+N-n}
\end{aligned}$$

$$\begin{aligned}
Cohen &= \frac{(P+N)(\frac{p}{p+n})-P}{\frac{P+N}{2}\frac{p+n+P}{p+n}-P}\\
Coleman &= \frac{(P+N)\frac{p}{p+n}-P}{N}
\end{aligned}$$

Now, we will focus on a brief discussion of the measures and presentation of their properties. The $g-measure$ ($g$ = 2) was proposed by Fürnkranz and Flach [22] and is a simplified version of the Laplace estimate. The $g-measure$ ($g$ = 2) assumes that a rule covering only one positive example is assigned true precision equal to 0.33. Many experiments (e.g. in papers [39, 23]) indicate that the precision of a rule evaluated on a training set is too optimistic. This especially concerns rules covering a small number of positive examples. If a rule covers large number of positive examples, the correction introduced by the number 2 has less and less meaning.

The $Weighted\ Laplace$ ($wLap$) measure is derived from decision tree induction algorithms and it is the Laplace estimate multiplied by $(P+N)/P$. So it takes into account the distribution of examples between the current positive class ($P$) and the remaining classes ($N$).

The $Logical\ sufficiency$ ($LS$) measure is applied to association rule evaluation as well as to decision rule induction [29]. It belongs to the group of measures which emphasise precision of rules during their evaluation and pay less attention to the number of examples covered by the evaluated rule.

The RSS measure is a measure equivalent, in terms of the generated rule order, to the well-known weighted relative accuracy (WRA) measure used by Lavrac, Flach, and Zupan [28], both in rule induction and subgroup discovery. However, RSS and WRA have different ranges of values. The use of RSS for classification conflicts resolving leads to better results than the use of the WRA measure.

The $MutualSupport$ ($MS$) measure is, in context of decision rule evaluation, presented in the paper by Yao and Zhong [29]. Let us assume that a rule of the form $\varphi \rightarrow \psi$ is given. MS measures the strength of dependencies not only between $\varphi \rightarrow \psi$, but also between $\varphi \leftarrow \psi$. It follows that we can consider it as a measure evaluating the strength of the double implication $\varphi \leftrightarrow \psi$. Therefore, the measure prefers rules characterised by high coverage.

The $Correlation$ ($Corr$) measure computes the correlation coefficient between the predicted and the target labels. It was applied to classification rule induction algorithms as well as to subgroup discovery and evaluation of association rules [15, 23, 40].

The $s-Bayesian\ confirmation$ ($s$) measure has been proposed by Christensen [41] and Joyce [42] as a confirmation measure. The first component of the measure evaluates the rule precision and the second is responsible for decrease in the quality of rules that cover small number of positive examples. Its application in evaluation of decision rules obtained by the rough set theory was considered in papers [43, 30].

Measures C1 and C2 have been proposed by Bruha [27, 21]. The measures are a combination of two quality measures known as the Coleman and Cohen (see Table II) measures. Bruha noticed that the Coleman measure prefers rules with high precision and small coverage, while the Cohen measure prefers rules with high coverage. Hence, C1 and C2 join the assessment made by the Coleman and Cohen measures

## V. Experiments and Results

In this paper, the performance of the rule-based classifier is evaluated by the two criteria: overall classification ( $Acc$ ) accuracy and balanced accuracy ( $BAcc$ ). The overall accuracy is the ratio of the number of correctly classified examples to the number of all examples. This is one of the most common criteria for assessing a classifier. However, in the case of unbalanced distribution of examples between decision classes, higher value of overall accuracy is often achieved at the cost of low accuracy of minority classes, therefore in such a case the balanced accuracy is more appropriate. It calculates the classification accuracy of each decision class and then takes an average over all classes. In the case of 2-class problems, balanced accuracy is equivalent to Area Under the ROC Curve (AUC) criterion [38]. One can meet with the interpretation that balanced accuracy calculated for a number of classes is a generalization of AUC for multi-class problems [44].

The used rule induction algorithm generates an unordered set of rules, therefore during the classification of examples it may happen that the test example is covered by rules pointing at different decision classes. In that case, a strategy

TABLE III
CHARACTERISTICS OF DATA SETS USED IN EMPIRICAL STUDIES.

| dataset | cl. | obj. | maj. | attributes | | dataset | cl. | obj. | maj. | attributes | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | class | all | nom. | | | | class | all | nom. |
| anneal | 5 | 898 | 76.17 | 38 | 32 | hepatitis | 2 | 155 | 79.35 | 19 | 13 |
| audiology | 24 | 226 | 25.22 | 69 | 69 | horse-colic | 2 | 368 | 63.04 | 22 | 15 |
| auto-mpg | 3 | 398 | 62.56 | 7 | 2 | hungarian-heart-disease | 2 | 294 | 63.95 | 13 | 7 |
| autos | 6 | 205 | 32.68 | 25 | 10 | iris | 3 | 150 | 33.33 | 4 | 0 |
| balance-scale | 3 | 625 | 33.33 | 4 | 0 | mammography-masses | 2 | 961 | 53.69 | 5 | 2 |
| breast-cancer | 2 | 286 | 70.28 | 9 | 9 | mushroom | 2 | 8124 | 51.80 | 22 | 22 |
| breast-w | 2 | 699 | 65.52 | 9 | 0 | prnn-synth | 2 | 250 | 50.00 | 2 | 0 |
| car | 4 | 1728 | 70.02 | 6 | 6 | segment | 7 | 2310 | 14.29 | 19 | 0 |
| cleveland | 5 | 303 | 54.13 | 13 | 7 | sick-euthyroid | 2 | 3772 | 93.88 | 29 | 22 |
| contact-lenses | 3 | 24 | 62.50 | 4 | 4 | sonar | 2 | 208 | 53.37 | 60 | 0 |
| credit-g | 2 | 1000 | 70.00 | 20 | 13 | soybean | 19 | 683 | 13.47 | 35 | 35 |
| cylinder-bands | 2 | 540 | 57.78 | 35 | 18 | titanic | 2 | 2201 | 67.70 | 3 | 3 |
| diabetes | 2 | 768 | 65.10 | 8 | 0 | vehicle | 4 | 846 | 25.77 | 18 | 0 |
| echocardiogram | 2 | 131 | 67.18 | 11 | 2 | vote | 2 | 435 | 61.38 | 16 | 16 |
| ecoli | 8 | 336 | 42.56 | 7 | 0 | wine | 3 | 178 | 39.89 | 13 | 0 |
| flag | 4 | 194 | 46.91 | 28 | 18 | yeast | 10 | 1484 | 31.20 | 8 | 0 |
| heart-statlog | 2 | 270 | 55.56 | 13 | 0 | zoo | 7 | 101 | 40.59 | 16 | 15 |

for resolving such conflicts has to be chosen. The most popular one is known as the voting scheme and consists in assigning a numeric value called the confidence degree to each rule. Confidence degrees of the rules covering the test example are summed up for each decision class and then the class with a maximal confidence degree is picked. In this paper the voting scheme is also used during classification. Initially, we assumed that the confidence degree of each rule is equal to its value of some quality measure used during the rule induction. If the test example is not covered by any of the rules, it is treated as wrongly classified.

All presented results stand for average values obtained based on the analysis of 34 data sets from the UCI Repository [45]. The analysis of each set was conducted with the use of the stratified 10-fold cross validation strategy. Comparisons were made on the same partitions of the data sets. The characteristics of these sets are presented in Table III. As it can be seen, the sets with different characteristics were selected for the analysis, particularly in relation to the number of decision classes, the distribution of examples between the classes and types of the attributes. The following abbreviations mean: *cl.* – number of classes, *obj.* – number of objects, *maj. class* – a majority class fraction in %, *nom.* – a number nominal attributes.

Tables IV and V contain the results of the classification of rule based classifiers which were obtained in such a way that the same quality measure was used at each stage of the rule induction and during the resolution of classification conflicts.

TABLE IV
RESULTS OF THE MEASURES COMPARISON ON 34 DATA SETS.

| measure | rules | $Acc[\%]$ | $BAcc[\%]$ |
|---|---|---|---|
| g | 144 | 81.4 | 74.2 |
| wLap | 156 | 81.1 | 77.8 |
| LS | 172 | 80.7 | 75.8 |
| RSS | 34 | 78.2 | 72.3 |
| MS | 32 | 77.8 | 70.5 |
| C1 | 151 | 82.0 | 73.0 |
| C2 | 127 | 82.3 | 76.3 |
| Corr | 44 | 79.9 | 74.3 |
| s | 86 | 80.3 | 74.9 |

TABLE V
CHARACTERISTICS OF THE INDUCED RULE SETS.

| Measure | Classification conflicts | | Rules | | | |
|---|---|---|---|---|---|---|
| | all | wrong | number | elementary conditions | Avg Prec. | Avg Cov. |
| g | 32 | 10 | 144 | 2.9 | 0.96 | 0.32 |
| wLap | 29 | 9 | 156 | 2.6 | 0.98 | 0.26 |
| LS | 29 | 8 | 172 | 2.4 | 0.98 | 0.19 |
| RSS | 58 | 18 | 34 | 3.6 | 0.77 | 0.74 |
| MS | 59 | 19 | 32 | 3.8 | 0.73 | 0.77 |
| C1 | 29 | 8 | 151 | 2.5 | 0.98 | 0.28 |
| C2 | 33 | 9 | 127 | 2.8 | 0.96 | 0.36 |
| Corr | 53 | 16 | 44 | 3.7 | 0.81 | 0.68 |
| s | 51 | 15 | 86 | 2.6 | 0.96 | 0.34 |

It can be seen that the C2 measure leads, on average, to the highest overall classification accuracy, wLap to the highest balanced accuracy, and RSS to the smallest number of rules but with significantly worse classification qualities (in the paper [33] it was proven that this worsening is statistically relevant). What is more, the Corr measure can be a reasonable compromise between the number of generated rules and their classification abilities. A more detailed analysis of the measures efficiency can be found in [26, 33]. Table V presents yet unpublished results which characterise the rule sets generated by means of particular measures. Here we can find the following interesting regularities:

- the highest values of $Acc$ and $BAcc$ take (with a few exceptions) the measures leading to induction of more rules, the average precision of the rule (denoted as $AvgPrec$) obtained by such measures is greater than 0.95, the average coverage (denoted as $AvgCov$) is not greater than 0.36; if the induction of rules aims at classification, it is appropriate to use measures attaching the utmost importance to the rule precision;
- in the classifiers obtained based on measures leading

TABLE VI
AVERAGE RANKING OF DIFFERENT MEASURE APPLICATION ON THE CLASSIFICATION PHASE.

| induction | criterion | C1 | C2 | Corr | g | LS | MS | Rss | s | wLap |
|---|---|---|---|---|---|---|---|---|---|---|
| g | $Acc$ | 5.0 | 4.5 | 5.1 | 5.1 | 4.5 | 4.5 | 4.1 | 5.9 | 6.4 |
| | $BAcc$ | 5.0 | 5.1 | 5.0 | 6.3 | 5.2 | 5.5 | 4.9 | 4.2 | 3.8 |
| C1 | $Acc$ | 4.5 | 3.8 | 4.4 | 5.3 | 5.4 | 4.8 | 4.7 | 5.6 | 6.5 |
| | $BAcc$ | 4.7 | 4.9 | 4.2 | 6.8 | 5.6 | 5.2 | 5.0 | 4.3 | 4.3 |
| C2 | $Acc$ | 4.1 | 3.9 | 4.6 | 4.9 | 5.9 | 5.9 | 6.0 | 4.7 | 5.1 |
| | $BAcc$ | 4.5 | 4.9 | 4.6 | 6.2 | 5.3 | 6.1 | 6.1 | 4.1 | 3.3 |
| Corr | $Acc$ | 4.3 | 4.1 | 4.7 | 5.0 | 5.0 | 5.6 | 5.7 | 4.4 | 6.1 |
| | $BAcc$ | 4.8 | 4.7 | 4.9 | 6.4 | 3.2 | 6.8 | 5.3 | 5.0 | 4.2 |
| wLap | $Acc$ | 4.5 | 4.3 | 4.1 | 5.3 | 5.0 | 5.0 | 4.8 | 5.8 | 6.3 |
| | $BAcc$ | 4.6 | 5.0 | 4.2 | 6.5 | 5.3 | 5.9 | 5.5 | 4.2 | 3.8 |
| LS | $Acc$ | 4.0 | 3.9 | 4.9 | 4.6 | 4.9 | 5.0 | 4.8 | 6.2 | 6.7 |
| | $BAcc$ | 4.5 | 4.9 | 4.5 | 6.6 | 5.7 | 5.5 | 5.1 | 4.5 | 3.6 |
| MS | $Acc$ | 4.7 | 4.3 | 4.8 | 4.7 | 5.9 | 4.7 | 5.9 | 4.1 | 5.9 |
| | $BAcc$ | 4.6 | 4.7 | 4.9 | 6.1 | 3.5 | 6.5 | 5.3 | 5.4 | 4.1 |
| Rss | $Acc$ | 4.3 | 4.3 | 4.9 | 5.6 | 3.9 | 5.5 | 5.6 | 4.9 | 6.1 |
| | $BAcc$ | 4.9 | 5.1 | 4.9 | 6.6 | 3.2 | 6.2 | 5.0 | 4.9 | 4.3 |
| s | $Acc$ | 4.3 | 4.4 | 4.8 | 4.6 | 5.4 | 5.4 | 5.8 | 4.5 | 5.9 |
| | $BAcc$ | 4.2 | 4.6 | 5.0 | 6.2 | 4.7 | 6.1 | 5.9 | 4.1 | 4.3 |
| avg | $Acc$ | 4.4 | **4.2** | 4.7 | 5.0 | 5.1 | 5.2 | 5.3 | 5.1 | 6.1 |
| | $BAcc$ | 4.6 | 4.9 | 4.7 | 6.4 | 4.6 | 6.0 | 5.3 | 4.5 | **4.0** |

TABLE VII
CLASSIFICATION IMPROVEMENT COMPARISON.

| Induction | Class. | Acc [%] | Wilcoxon test p-value |
|---|---|---|---|
| g | g | 81.37 | 0.1271 |
| | C2 | 82.24 | |
| C1 | C1 | 81.95 | 0.1885 |
| | C2 | 82.06 | |
| corr | corr | 79.77 | 0.0572 |
| | C2 | 80.28 | |
| wLap | wLap | 81.02 | 0.0154 |
| | C2 | 82.80 | |
| LS | LS | 80.80 | 0.3684 |
| | C2 | 81.02 | |
| MS | MS | 77.90 | 0.5723 |
| | C2 | 77.14 | |
| Rss | Rss | 78.07 | 0.0383 |
| | C2 | 79.08 | |
| s | s | 80.39 | 0.9357 |
| | C2 | 79.43 | |

| Induction | Class. | BAcc [%] | Wilcoxon test p-value |
|---|---|---|---|
| g | g | 74.17 | 0.0050 |
| | wLap | 78.15 | |
| C1 | C1 | 76.91 | 0.0512 |
| | wLap | 77.81 | |
| C2 | C2 | 76.22 | 0.0066 |
| | wLap | 78.35 | |
| corr | corr | 74.21 | 0.2206 |
| | wLap | 76.01 | |
| LS | LS | 75.67 | 0.0161 |
| | wLap | 77.04 | |
| MS | MS | 70.67 | 0.0090 |
| | wLap | 74.57 | |
| Rss | Rss | 72.21 | 0.0672 |
| | wLap | 74.06 | |
| s | s | 75.03 | 0.6811 |
| | wLap | 74.92 | |

to the induction of a smaller number of more general rules (the average precision of induced rules is not higher than 0.96, the average coverage is not less than 0.34), there is a large number of classification conflicts (Classification conflicts); a large number of these conflicts is resolved incorrectly, which explains the lower quality of the classifiers;

- the larger the rule coverage, the greater the number of positive examples covered by them uniquely;
- the rule average precision and rule average coverage are correlated with the number of induced rules;
- the average number of elementary conditions (Elementary conditions) contained in the rule premises is correlated with the number of induced rules and their average coverage.

The experiments show that the rules generated by means of measures which lead to the induction of a large number of precise rules (precision $> 0.95$) contain fewer elementary conditions than the rules obtained with the use of measures which promote more general rules. This means that the ranges of elementary conditions, set by means of MS and RSS measures, are wide. This, in turn, implies that each individually considered condition covers a large number of examples, including negative ones. In order to have more detailed rules with "wide" elementary conditions, it is necessary to place more conditions in the rule premise. The measures which are oriented towards the induction of precise rules (wLap, C1, C2) make conditions with narrower ranges, thus the number of conditions in the rule premise is smaller.

The first experiment to check whether it is sensible to use the combination of measures aimed at selecting a measure to resolve classification conflicts. The experiment was conducted in such a way that for the set measure (the same at the stages of growing and pruning) applied in the rule induction, the measure for resolving classification conflicts was changed. The results are presented in Table VI.

The first column of this table denotes the quality measure used in the rule induction. The second column denotes the quality criterion of a final classification phase. The following nine columns correspond to nine quality measures used during the voting in the classification. Each of these nine values is an average rank of a specific classifier (ranks for the classifiers should be read by rows) — the lower is the value of the ranking, the better are the results achieved by the measure (value equal to 1 would mean that the measure was always first in the ranking). Experiments were performed on 34 mentioned datasets in a stratified 10 fold cross validation model. A Friedman test with the significance level $\alpha = 0.05$ does not show statistical differences only for the $s$ measure and the criterion $Acc$ . If post–hoc analysis is performed and the critical distance is calculated it occurs that most of measures behave in a similar way: there are no significant differences between classification accuracy between different measures application. However, it is worth to notice that for the $Acc$ criterion $C2$ measure is the right choice and for the $BAcc$ criterion the $wLap$ is the right choice. $C2$ measure (according to $Acc$ criterion) is the best in 5 of 9 cases (measures) and is always in the best three places. $wLap$ measure (according to $BAcc$ criterion) is four time the best one and not worse than third in other cases. Additionally, an averaging row is also included in the bottom of the table. Emphasised averaged values confirm the above conclusions.

The influence of application of C2 measure during classification instead of a measure used in the rule induction is presented on the left side of the Table VII. The first column points the measure used in a rule induction, second column points the measure used during the classification and the third column means the $Acc$ of a classification. Additional column remarks the p-value of the Wilcoxon test, which null hypothesis is that there is no statistical significant difference between the algorithm which uses the same measure during classification and during the rule induction and the algorithm which uses the C2 measure during classification. Analogical summary for the wLap measure applied in the classification with the $BAcc$ classification quality criterion is presented on the right side of the Table VII.

As it can be observed in the most of cases the final classification quality is improved as the result of application a specified rule quality measure. In the case of a C2 measure this result is rather expected as this measure is strictly dedicated for the classification conflict resolution purposes [27, 21]. wLap measure also improves classification results and the statistical significance of this improvement is very high in the most of cases. This is a new and important observation.

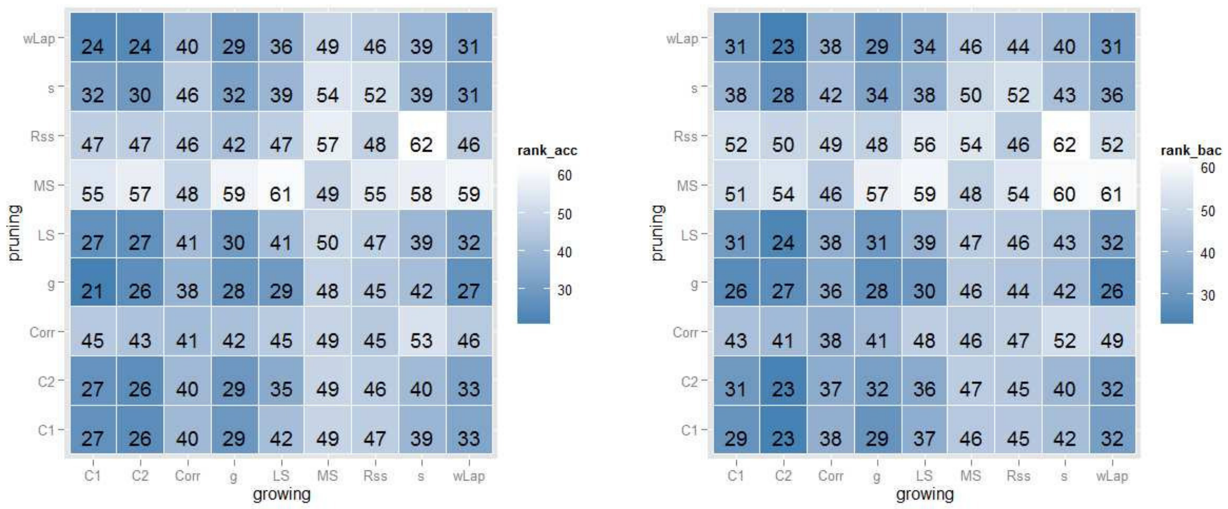
Having the selected rule quality measures for two criteria of classification quality ( $Acc$ and $BAcc$ ) we performed experiments on application of different rule quality measures during the rule growing and pruning phase. This time rankings were calculated on the basis of 81 experiments (nine measures for the growing and nine for the pruning). Results are presented on Fig. 1.

TABLE VIII
COMPARISON OF RESULTS OF THE COMBINATION OF QUALITY MEASURES FOR THE $Acc$ .

| Growing/Pruning/ Classification | Acc | Rules | Wilcoxon |
|---|---|---|---|
| 1. C2-C2-C2 | 82.3 | 127 | 1 - 2 (+)<br>1 - 3 (+)<br>1 - 4 (+) |
| 2. C2-Corr-C2 | 80.1 | 47 | 2 - 3 (-)<br>2 - 4 (-) |
| 3. Corr-C2-C2 | 80.4 | 46 | 3 - 4 (+) |
| 4. Corr-Corr-Corr | 79.8 | 44 | |
| 1. wLap-wLap-C2 | 81.8 | 156 | 1 - 2 (+)<br>1 - 3 (+)<br>1 - 4 (+) |
| 2. wLap-Corr-C2 | 79.4 | 47 | 3 - 2 (+)<br>2 - 4 (-) |
| 3. Corr-wLap-C2 | 80.4 | 46 | 3 - 4 (+) |
| 4. Corr-Corr-Corr | 79.8 | 44 | |
| 1. C2-C2-C2 | 82.3 | 127 | 1 - 2 (+)<br>1 - 3 (+)<br>1 - 4 (+) |
| 2. C2-RSS-C2 | 78.7 | 34 | 2 - 3 (-)<br>2 - 4 (-) |
| 3. RSS-C2-C2 | 79.3 | 37 | 3 - 4 (+) |
| 4. RSS-RSS-RSS | 78.2 | 34 | |
| 1. wLap-wLap-C2 | 81.8 | 156 | 1 - 2 (+)<br>1 - 3 (+)<br>1 - 4 (+) |
| 2. wLap-RSS-C2 | 78.3 | 30 | 3 - 2 (+)<br>2 - 4 (-) |
| 3. RSS-wLap-C2 | 79.4 | 37 | 3 - 4 (+) |
| 4. RSS-RSS-RSS | 78.2 | 34 | |

It features a matrix which reflects the ranking of the given combination of measures. As it was already mentioned, 81 combinations of measures were checked. For the $Acc$ criterion, the conflicts resolving measure was always the C2 measure, while for $BAcc$ - the wLap measure. The matrices show that it is not possible to draw straightforward conclusions from such a general comparison. It is interesting that the values presented in antidiagonal (X-X - showing the case when the same measure is used in both phases of the rule induction) are very close to the lowest values in row X and column X. A more detailed analysis was conducted for the C2 and wLap measures (as the best due to their classification abilities) as well as Corr and RSS (as the best due to the number of generated rules). In Tables VIII and IX an analysis of the combination of these measures was conducted for the $Acc$ (Table VIII) and $BAcc$ (Table IX) criteria. The first column of both tables includes the combination of measures, while the second one - the $Acc$ ( $BAcc$ ) value obtained by the given combination. Finally, the third column features the achieved number of rules. The fourth column contains the results of the paired comparison of particular combinations (the Wilcoxon test was used; p-value=0.1) of four models. X-Y (+) means that the combination X is statistically better than Y. The (-) symbol means that there is no such a difference.

Let us assume that the precise measure (C2, wLap) is marked P, while the general measure (Corr, RSS) - C. Additionally, the P-C inscription means that in the phase of the rule growing the precise measure was used, while in the



| pruning \ growing | C1 | C2 | Corr | g | LS | MS | Rss | s | wLap |
|---|---|---|---|---|---|---|---|---|---|
| wLap | 24 | 24 | 40 | 29 | 36 | 49 | 46 | 39 | 31 |
| s | 32 | 30 | 46 | 32 | 39 | 54 | 52 | 39 | 31 |
| Rss | 47 | 47 | 46 | 42 | 47 | 57 | 48 | 62 | 46 |
| MS | 55 | 57 | 48 | 59 | 61 | 49 | 55 | 58 | 59 |
| LS | 27 | 27 | 41 | 30 | 41 | 50 | 47 | 39 | 32 |
| g | 21 | 26 | 38 | 28 | 29 | 48 | 45 | 42 | 27 |
| Corr | 45 | 43 | 41 | 42 | 45 | 49 | 45 | 53 | 46 |
| C2 | 27 | 26 | 40 | 29 | 35 | 49 | 46 | 40 | 33 |
| C1 | 27 | 26 | 40 | 29 | 42 | 49 | 47 | 39 | 33 |

| pruning \ growing | C1 | C2 | Corr | g | LS | MS | Rss | s | wLap |
|---|---|---|---|---|---|---|---|---|---|
| wLap | 31 | 23 | 38 | 29 | 34 | 46 | 44 | 40 | 31 |
| s | 38 | 28 | 42 | 34 | 38 | 50 | 52 | 43 | 36 |
| Rss | 52 | 50 | 49 | 48 | 56 | 54 | 46 | 62 | 52 |
| MS | 51 | 54 | 46 | 57 | 59 | 48 | 54 | 60 | 61 |
| LS | 31 | 24 | 38 | 31 | 39 | 47 | 46 | 43 | 32 |
| g | 26 | 27 | 36 | 28 | 30 | 46 | 44 | 42 | 26 |
| Corr | 43 | 41 | 38 | 41 | 48 | 46 | 47 | 52 | 49 |
| C2 | 31 | 23 | 37 | 32 | 36 | 47 | 45 | 40 | 32 |
| C1 | 29 | 23 | 38 | 29 | 37 | 46 | 45 | 42 | 32 |

Fig. 1. Matrices of average rankings of models with different quality measures used in growing and pruning but a common classification quality: $Acc$ (left) and $BAcc$ (right).

TABLE IX
COMPARISON OF RESULTS OF THE COMBINATION OF QUALITY MEASURES FOR THE $BAcc$.

| Growing/Pruning/ Classification | Acc | Rules | Wilcoxon |
|---|---|---|---|
| 1. wLap-wLap-wLap | 77.8 | 156 | 1 - 2 (+)<br>1 - 3 (+)<br>1 - 4 (+) |
| 2. wLap-Corr-wLap | 74.7 | 47 | 3 - 2 (+)<br>2 - 4 (-) |
| 3. Corr-wLap-wLap | 76.1 | 46 | 3 - 4 (+) |
| 4. Corr-Corr-Corr | 74.4 | 44 | |
| 1.C2-C2-wLap | 78.4 | 127 | 1 - 2 (+)<br>1 - 3 (+)<br>1 - 4 (+) |
| 2. C2-Corr-wLap | 75.4 | 47 | 2 - 3 (-)<br>2 - 4 (-) |
| 3. Corr-C2-wLap | 76.1 | 46 | 3 - 4 (+) |
| 4. Corr-Corr-Corr | 74.4 | 44 | |
| 1. wLap-wLap-wLap | 77.8 | 156 | 1 - 2 (+)<br>1 - 3 (+)<br>1 - 4 (+) |
| 2. wLap-RSS-wLap | 72.3 | 30 | 3 - 2 (+)<br>2 - 4 (-) |
| 3. RSS-wLap-wLap | 74.5 | 37 | 3 - 4 (+) |
| 4. RSS-RSS-RSS | 72.3 | 34 | |
| 1. C2-C2-wLap | 78.4 | 127 | 1 - 2 (+)<br>1 - 3 (+)<br>1 - 4 (+) |
| 2. C2-RSS-wLap | 73.2 | 34 | 2 - 3 (-)<br>2 - 4 (-) |
| 3. RSS-C2-wLap | 74.4 | 37 | 3 - 4 (+) |
| 4. RSS-RSS-RSS | 72.3 | 34 | |

rule pruning - the general one. Both tables demonstrate the following dependencies:

- the P-C combination, when compared with P-P, leads to a significantly worse quality of the classifier and to a decreasing number of rules,
- the C-P combination, when compared with C-C, leads to a significant increase of the classification quality with, simultaneously, a small increase in the number of rules,
- it is not significant for the number of determined rules whether we use the C-P or P-C sequence, however it is recommended to use C-P for classification abilities as it allows to significantly raise these classification abilities (with respect to C-C).

Moreover, the C2-C2-C2 sequence allowed to achieve the highest classification accuracy. No other combination allowed to statistically improve the $Acc$ value. On the other hand, the C2-C2-wLap combination leads to significantly higher $BAcc$ values, than the wLap-wLap-wLap combination.

## VI. CONCLUSIONS

When real problems are solved (analysed), the measure is usually selected with respect to the data set specifics. The automatic method for the measures selection is based on internal cross-validation. Our earlier works allowed to limit the number of measures which are considered by the automatic method [33], which significantly speeds up the calculations. The results of this work extend the knowledge about the combinations of measures which should be tested first in the automatic method. Our results can also be useful for the user for whom the maximisation of the classification accuracy is not the most important criterion, while it is more important to have proper balancing of the model complexity (number of rules) with its classification abilities.

Currently the RapidRuleInduction plug-in is prepared for the RapidMiner environment. The plug-in will contain the possibilities to induce the rules by means of combining different quality measures - similarly to the plug-in we developed for the R environment [46].

## References

[1] J. Błaszczyński, R. Słowiński, and M. Szeląg, "Sequential covering rule induction algorithm for variable consistency rough set approaches," *Information Sciences*, vol. 181, no. 5, pp. 987–1002, 2011. doi: 10.1016/j.ins.2010.10.030. [Online]. Available: {http://dx.doi.org/10.1016/j.ins.2010.10.030}

[2] W. Cohen, "Fast effective rule induction," in *Proc. of the 12th Int. Conference ICML95*, 1995, pp. 115–123.

[3] T. Fawcett, "Prie a system for generation rulelist to maximize roc performance," *Data Mining and Knowledge Discovery*, vol. 17, pp. 207–224, 2008. doi: 10.1007/s10618-008-0089-y. [Online]. Available: {http://dx.doi.org/10.1007/s10618-008-0089-y}

[4] J. Fürnkranz, "Separate-and-conquer rule learning," *Artificial Intelligence Review*, vol. 13, no. 1, pp. 3–54, 1999. doi: 10.1023/A:1006524209794. [Online]. Available: {http://dx.doi.org/10.1023/A:1006524209794}

[5] J. Grzymała-Busse and W. Ziarko, *Data Mining Opportunities and Challenges*. Idea Group Publishing, 2003, ch. Data mining based on rough sets, pp. 142–173.

[6] K. Kaufman and R. Michalski, "Learning in inconsistent world, rule selection in star/aq18," Machine Learning and Inference Laboratory, Tech. Rep. Report P99-2 1999, 1999.

[7] J. Stefanowski, "Rough set based rule induction techniques for classification problems," in *Proc. Of the 6th European Congress of Intelligent Techniques and Soft Computing*, 1998, pp. 107–119.

[8] E. Czogała and J. Łęski, *Fuzzy and Neuro-Fuzzy Intelligent Systems*, ser. Studies in Fuzziness and Soft Computing. Physica-Verlag, 2000, vol. 47. [Online]. Available: {http://dx.doi.org/10.1007/978-3-7908-1853-6}

[9] O. Nelles, A. Fink, R. Babuska, and M. Setnes, "Comparison of two construction algorithms for takagi-sugeno fuzzy models," *International Journal of Applied Mathematics and Computer Science*, vol. 10, no. 4, pp. 835–855, 2000.

[10] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. of the 5th Annual ACM Workshop on Computational Learning Theory*, 1992. doi: 10.1145/130385.130401 pp. 144–152. [Online]. Available: {http://dx.doi.org/10.1145/130385.130401}

[11] K. Dembczynski, W. Kotłowski, and R. Słowiński, "Ender: a statistical framework for boosting decision rules," *Data Mining and Knowledge Discovery*, vol. 21, no. 1, pp. 52–90, 2010. doi: 10.1007/s10618-010-0177-7. [Online]. Available: {http://dx.doi.org/10.1007/s10618-010-0177-7}

[12] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. of the 20th VLDB Conference*, Santiago, Chile, 2004, pp. 487–499.

[13] B. Kavsek and N. Lavrac, "Apriori-sd: Adapting association rule learning to subgroup discovery," *Applied Artificial Intelligence*, vol. 20, no. 7, pp. 543–583, 2006. doi: 10.1007/978-3-540-45231-7_22. [Online]. Available: {http://dx.doi.org/10.1007/978-3-540-45231-7_22}

[14] J. Stefanowski and D. Vanderpooten, "Induction of decision rules in classification and discovery-oriented perspectives," *International Journal of Intelligent Systems*, vol. 16, no. 1, pp. 13–27, 2001.

[15] L. Geng and H. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys*, vol. 38, no. 6, p. art. no. 9, 2006. doi: 10.1145/1132960.1132963. [Online]. Available: {http://dx.doi.org/10.1145/1132960.1132963}

[16] K. McGarry, "A survey of interestingness measures for knowledge discovery," *The Knowledge Engineering Review*, vol. 20, no. 1, pp. 39–61, 2005. doi: 10.1017/S0269888905000408. [Online]. Available: {http://dx.doi.org/10.1017/S0269888905000408}

[17] S. Sahar, *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag, 2010, ch. Interestingness measures - On determining what is interesting, pp. 603–612. [Online]. Available: {http://dx.doi.org/10.1007/978-0-387-09823-4_30}

[18] D. Gamberger and N. Lavrac, "Expert-guided subgroup discovery: methodology and application," *Journal of Artificial Intelligence Research*, vol. 17, no. 1, pp. 501–527, 2002. doi: 10.1613/jair.1089. [Online]. Available: {http://dx.doi.org/10.1613/jair.1089}

[19] N. Lavrac, B. Kavsek, and P. Flach, "Subgroup discovery with cn2-sd," *Journal of Machine Learning Research*, vol. 5, pp. 153–188, 2004.

[20] A. An and N. Cercone, "Rule quality measures for rule induction systems: description and evaluation," *Computational Intelligence*, vol. 17, no. 3, pp. 409–424, 2001. doi: 10.1111/0824-7935.00154. [Online]. Available: http://dx.doi.org/10.1111/0824-7935.00154

[21] I. Bruha and J. Tkadlec, "Rule quality for multiple-rules classifier: Empirical expertise and theoretical methodology," *Intelligent Data Analysis*, vol. 7, no. 2, pp. 99–124, 2003.

[22] J. Fürnkranz and P. Flach, "Roc 'n' rule learning - towards a better understanding of covering algorithms," *Machine Learning*, vol. 39–77, 2005. doi: 10.1007/s10994-005-5011-x. [Online]. Available: {http://dx.doi.org/10.1007/s10994-005-5011-x}

[23] F. Janssen and J. Fürnkranz, "On the quest for optimal rule learning heuristics," *Machine Learning*, vol. 78, pp. 343–379, 2010. doi: 10.1007/s10994-009-5162-2. [Online]. Available: {http://dx.doi.org/10.1007/s10994-009-5162-2}

[24] M. Sikora, "Rule quality measures in creation and reduction of data rule models," *Lecture Notes in Artificial Intelligence*, vol. 4259, pp. 716–725, 2006. doi: 10.1007/11908029_74. [Online]. Available: {http://dx.doi.org/10.1007/11908029_74}

[25] ——, "Decision rule-based data models using trs and nettrs - methods and algorithms," *Transaction on*

*Rough Sets - Lecture Notes on Computer Science*, vol. 5946, pp. 130–160, 2010. doi: 10.1007/978-3-642-11479-3_8. [Online]. Available: {http://dx.doi.org/10.1007/978-3-642-11479-3_8}

[26] M. Sikora and Ł.Wróbel, "Data-driven adaptive selection of rule quality measures for improving the rule induction algorithm," *Lecture Notes in Computer Science*, vol. 6743, pp. 279–287, 2011. doi: 10.1007/978-3-642-21881-1_44. [Online]. Available: {http://dx.doi.org/10.1007/978-3-642-21881-1_44}

[27] I. Bruha, *Machine Learning and Statistics: The Interface*. John Wiley, 1997, ch. Quality of decision rules: Definitions and classification schemes for multiple rules, pp. 107–131.

[28] N. Lavrac, P. Flach, and B. Zupan, "Rule evaluation measures: A unifying view," *Lecture Notes in Artificial Intelligence*, vol. 1634, pp. 174–185, 1999. doi: 10.1007/3-540-48751-4_17. [Online]. Available: {http://dx.doi.org/10.1007/3-540-48751-4_17}

[29] Y. Yao and N. Zhong, "An analysis of quantitative measures associated with rules," *Lecture Notes in Computer Science*, vol. 1574, pp. 479–488, 1999. doi: 10.1007/3-540-48912-6_64. [Online]. Available: {http://dx.doi.org/10.1007/3-540-48912-6_64}

[30] S. Greco, Z. Pawlak, and R. Słowiński, "Can bayesian confirmation measures be useful for rough set decision rules?" *Engineering Applications of Artificial Intelligence*, vol. 17, no. 4, pp. 345–361, 2004. doi: 10.1016/j.engappai.2004.04.008. [Online]. Available: {http://dx.doi.org/10.1016/j.engappai.2004.04.008}

[31] P. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in *Proc. of the 8th International Conference on Knowledge Discovery and Data Mining*, 2002. doi: 10.1145/775047.775053 pp. 32–41. [Online]. Available: {http://dx.doi.org/10.1145/775047.775053}

[32] T. Agotnes, J. Komorowski, and T. Loken, "Taming large rule models in rough set approaches," *Lecture Notes in Artificial Intelligence*, vol. 1704, pp. 193–203, 1999. doi: 10.1007/978-3-540-48247-5_21. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-48247-5_21

[33] M. Sikora and Ł.Wróbel, "Data-driven adaptive selection of rule quality measures for improving rule induction and filtration algorithm," *International Journal of General Systems*, vol. 42, no. 6, pp. 594–613, 2013. doi: 10.1080/03081079.2013.798901. [Online]. Available: {http://dx.doi.org/10.1080/03081079.2013.798901}

[34] T. Amin, I. Chikalov, M. Moshkov, and B. Zielosko, "Relationships between length and coverage of decision rules," *Fundamenta Informaticae*, vol. 129, no. 1–2, pp. 1–13, 2014. doi: 10.3233/FI-2014-956. [Online]. Available: {http://dx.doi.org/10.3233/FI-2014-956}

[35] U. Stańczyk, "Decision rule length as a basis for evaluation of attribute relevance," *Journal of Intelligent and Fuzzy Systems*, vol. 24, no. 3, pp. 429–445, 2013. doi: 10.3233/IFS-2012-0564. [Online]. Available: {http://dx.doi.org/10.3233/IFS-2012-0564}

[36] A. Ławrynowicz and J. Potoniec, "Pattern based feature construction in semantic data mining," *International Journal on Semantic Web & Information Systems*, vol. 10, no. 1, pp. 27–65, 2014. doi: 10.4018/ijswis.2014010102. [Online]. Available: {http://dx.doi.org/10.4018/ijswis.2014010102}

[37] L. Wohlrab and J. Fürnkranz, "A review and comparison of strategies for handling missing values in separate-and-conquer rule learning," *Journal of Intelligent Information Systems*, vol. 36, no. 1, pp. 73–98, 2011. doi: 10.1007/s10844-010-0121-8. [Online]. Available: {http://dx.doi.org/10.1007/s10844-010-0121-8}

[38] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006. doi: 10.1016/j.patrec.2005.10.010. [Online]. Available: {http://dx.doi.org/10.1016/j.patrec.2005.10.010}

[39] G. Webb and D. Brain, "Generality is predictive of prediction accuracy," *Lecture Notes in Computer Science*, vol. 3755, pp. 1–13, 2006. doi: 10.1007/11677437_1. [Online]. Available: {http://dx.doi.org/10.1007/11677437_1}

[40] H. Xiong, S. Shekhar, P. N. Tan, and V. Kumar, "Exploiting a support-based upper bound of pearson's correlation coefficient for efficiently identifying strongly correlated pairs," in *In Proceedings of the 10th ACM SIGKDD*, 2004. doi: 10.1145/1014052.1014090 pp. 334–343. [Online]. Available: {http://dx.doi.org/10.1145/1014052.1014090}

[41] D. Christensen, "Measuring confirmation," *Journal of Philosophy*, vol. 96, no. 9, pp. 437–461, 1999. doi: 10.2307/2564707. [Online]. Available: {http://dx.doi.org/10.2307/2564707}

[42] J. Joyce, *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

[43] I. Brzezińska, R. Słowiński, and S. Greco, "Mining pareto-optimal rules with respect to support and confirmation or support and anti-support," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 5, pp. 587–600, 2007. doi: 10.1016/j.engappai.2006.11.015. [Online]. Available: {http://dx.doi.org/10.1016/j.engappai.2006.11.015}

[44] M. Wojnarski, A. Janusz, H. S. Nguyen, J. Bazan, and C. J. Luo, "Rsctc2010 discovery challenge: Mining dna microarrays data for medical diagnosis and treatment," *Lecture Notes in Artificial Intelligence*, vol. 6086, pp. 4–19, 2010. doi: 10.1007/978-3-642-13529-3_3. [Online]. Available: {http://dx.doi.org/10.1007/978-3-642-13529-3_3}

[45] A. Frank and A. Asuncion, "Uci machine learning repository [http://archive.ics.uci.edu/ml]."

[46] W. Malara, M. Sikora, and L. Wróbel, "An r package for induction and evaluation of classification rules," *Studia Informatica*, vol. 34, no. 2B, pp. 339–352, 2013.

DOI: 10.15439/2015F389
ACSIS, Vol. 5

# Constructive heuristics for technology-driven Resource Constrained Scheduling Problem

Paweł B. Myszkowski
Wrocław University of Technology, Wrocław, Poland
Email: pawel.myszkowski@pwr.edu.pl

Michał Przewoźniczek
Wrocław University of Technology MP2 company, Wrocław, Poland
Email: michal.przewozniczek@pwr.edu.pl

Marek Skowroński
Wrocław University of Technology, Wrocław, Poland
Email: m.e.skowronski@pwr.edu.pl

*Abstract*—**In this paper, we define a new practical technology-driven Resource Constrained Scheduling Problem (t-RCPSP). We propose three approaches, applying constructive heuristics to tackle effectively the practical application of RCPSP. In the RCPSP formulation, the constraints are defined to design the tasks in the spaces constructed by non- and renewable resources, without violating the precedence relationships and technologies in real world problem that exists in Plastic and Rubber Processing company. The difficulty of t-RCPSP is NP-hard and we proposed three constructive specialized methods: duration based heuristics (DBH), locally optimal resource usage PEC and NEH heuristic adaptation. The paper presents results of computational experiments that show the effectiveness of the proposed approaches.**

## I. Introduction

The automated computer-aided scheduling in real world application has a tremendous impact on the enterprise. Production schedule building process by human needs a lot of time (long hours), what increases costs and strongly depends on the human condition (costly mistakes). Moreover, the automated scheduling process requires less time (only seconds), is faultless and can be run anytime, e.g. to reschedule in the case of break-down production. In most cases, schedule generated by computer is more efficient than schedule built by the human.

In this paper, an automated scheduling problem practical application in Plastic and Rubber Processing Industry is investigated. Mainly, there is a set of injection molding machines, specialized devices, set of (sub)products and ingredients. Such renewable (e.g. machines, devices) and non-renewable (product's ingredients) resources should be assigned to client requests (tasks) to get near optimal usage in the production process. The major element of automated scheduling system is schedule builder. If solution is to be useful in practice, schedule builder should give the (sub)optimal production schedule in reasonable time: less than 1 minute is acceptable.

It is widely known that the automated computer-aided scheduling in real world application may reduce human work. However, our specific domain requirements make complicated application of classical algorithms. We propose three types of RCPSP solving methods: duration based heuristics (DBH) based on the greedy algorithm, classical NEH adaptation and method (PEC locally optimal resources usage driven.

The proposed technology-driven t-RCPSP can be generalized to RCPSP, which in literature is presented as NP-hard [1] and there are no exact algorithms to solve it in reasonable computation time. Some researches recommend heuristics [7][8]Error: Reference source not foundError: Reference source not found as fast and quite effective RCPSP solving tools.

To get schedule near optimal some metaheuristic approaches are recommended, e.g. Simulated Annealing [2], Tabu Search [12][15], Genetic Algorithms [18][22], Evolutionary Algorithms [5]Error: Reference source not found (hybrids EA [21]). Also, some swarm intelligence methods can be successfully applied to RCPSP, like Ant Colony Optimisation [4][9][10][11], hybrid ACO [13], Particle Swarm Optimization [23] or Bee Colony Algorithms [25].

The rest of the paper is organized as follows. Section 2 presents general RCPSP problem statement and specific domain requirements; technology-driven t-RCPSP model is proposed. Section 3 describes details of three proposed heuristics. Experiments of developed methods in a given dataset are presented in section 4. Finally, section 5 presents summary of gained results and gives some possible further research directions.

## II. Problem statement

In this section, the main elements of classical RCPSP are presented. In a real world problem, such RCPSP model can be useless. The main reason is that, in practical application, to realize the client's request machines and devices can use several configurations of ingredients that may cause entirely different task duration. Thus, problem that we met in MP2 company enforced us to extend RCPSP by several elements. Proposed technology-driven RCPSP (*t-RCPSP*) model in details is presented below.

### *A. Short description of technology-driven RCPSP*

In classical RCPSP Error: Reference source not found each task is described by duration, start and finish time. Tasks are non-preemptive, which means that preemption is not allowed. Each task can be lined to other one in timeline. We use discrete time measure - timeslots.

978-83-60810-66-8/$25.00©2015, IEEE

In the presented t-RCPSP application (schema is presented on Figure 1), we need to run several tasks, which are non-preemptive. Each task has its execution deadline and duration time that depends on used resources. To apply task and produce required product(s) some resources are used: machines, devices, materials and subproducts.

Resources are renewable (machines and devices) and non-renewable: materials (such plastic, paints and other ingredients) and subproducts. As some products are composed of other products, there is relation start-finish between tasks that produce needed subproducts of the given task. Some resources (machines and devices) are dedicated, what means that can be assigned only one activity at a given time [6].

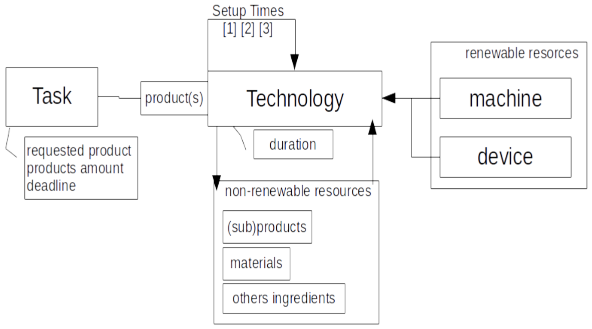


Fig 1. Schema of t-RCPSP

In t-RCPSP specific set of constraints that should be satisfied is defined. The feasible schedule satisfies all constraints defined as follows:

C1. Each task is applied only on one proper machine using specialized device,
C2. Each machine and device can be used only once in selected timeslot,
C3. Each device can cooperate with machine in various way, using other configuration of ingredients,
C4. Each task requires a given amount of ingredients: materials and subproducts,
C5. The task that produces subproducts must be finished before task requiring it,
C6. Each task has defined deadline and number of products,
C7. Each task has assigned duration time that depends on number of required products and used machine and device,
C8. There are 4 types of machine setup times that depend on two adjacent tasks:

- no operation – if two tasks produce the same products,
- start (15 timeslots duration) two tasks produce the same products and machine has been stopped,
- rinse (30 timeslots) two tasks use the same device and machine but provide other products,
- full refitting (120 timeslots) to clean machine and change device.

All devices are specialized to provide a given type of products using the machine, materials or/and subproducts. The device can be applied only to selected machines, and its effectiveness is connected with machine and configuration of ingredients. The main goal of t-RCPSP is to generate feasible schedule (according to C1-C8 constraints) to minimize its duration – makespan, calculated as the difference between first task start and end of the last task in the final schedule. The minor criteria is to reduce the average latency of schedule execution given as the difference between each task end time and its defined deadline.

This problem is NP-hard [1][3][6] and overconstrained. There are no effective algorithms therefore we propose use some heuristics to solve it in acceptable time. The extra constraint required by MP2 company is time limit, i.e. solving method execution cannot exceed 1 minute of CPU computational time of reference machine.

### *B. t-RSPSP - formulation*

The feasible schedule ($S$) consists $j=1,..,J$ tasks and each task is defined as a tuple:

$$J{:}=<\{request\_products\ [amounts]\ldots\},\ sj,\ dtj,\ ddj> \tag{1}$$

where $ddj$ defines task execution deadline, $sj$ means timeslot to start task in the discrete time period; However $dtj$ value strongly depends on used resources: machine, device and materials. To link such aspects model we defined technology $t=1,..,T$ as follows:

$$T{:}=<M,\ D,\ \{resources\ I\ [amount]\},\ \{products\ P\ amount\},\ dtt> \tag{2}$$

where $dtt$ value determines the task execution time using given set of resources. To apply technology to produce products ($P$) it uses renewable resource ($m=1,..,M$ machines and $d=1,..,D$ devices) and some product ingredients ($I$) as non-renewable resources: $MA=1,..,i$ materials and other resources $R=1,..,r$, including subproducts.

Various technologies can produce the same product using other resources and give other execution time. Such technology definition as an abstract layer makes possible to link the same resources in another way. Set of technologies describes the effectiveness of model and makes optimization simpler. The primary optimization goals are defined as follows:

$$min\ MAKESPAN(S) = max\ (sj+ddj) - min(sj) \tag{3}$$

Such formula gives information about the total schedule $S$ execution time, calculated as the differences between the last task's finish and start of the first task. It should be minimized to make schedule execution possible shorter. Another measure that gives quality of given schedule $S$ is the average latency defined as follows:

$$min\ AVG\_LATENCY(S) = \tag{4}$$

$$\frac{1}{k}\sum_{1..k}\left\{\begin{matrix} 0\, if\; sj+dtj<ddj \\ else\; ddj-sj+dtj \end{matrix}\right\}$$

Such measure gives the averaged value of how late each task is due to its defined deadline. It should be minimized to finish each task before its deadline and possible to avoid the delay (and potential financial penalties).

## III. Proposed methods

Each of three proposed methods: NEH adaptation, duration based heuristic (DBH) and resource optimal usage PEC are based on some observations and motivations. Moreover, methods differ not only in implementation but they also use various parameters. Proposed heuristics use sorting deadline criteria of tasks. We defined three basic criteria: ascending (tasks with earlier deadline have priority), descending (the opposite situation) and random order. We decided to implement the random task order to get reference to the other two. In this section, details of proposed methods are presented.

### *A. Duration based heuristic – DBH*

The main motivation of DBH is to build the shortest schedule using adaptation of classical greedy algorithm based on rule heuristics [19]. The DBH pseudocode is presented on Pseudocode 1. DBH heuristic works on all unassigned tasks and proposes first possible timeslot and uses the shortest technology to execute it.

The DBH heuristic asks model for set of tasks that can be preformed in selected timeslot (line 6). List of tasks is sorted by criteria (randomly, ascending or descending deadlines) to get one task (line 9). Then the technology with the shortest execution time is given to apply in given timeslot (line 10). If all model constraints are satisfied task is assigned in the schedule (line 11) and removed from list of unassigned tasks (line 12). If there no tasks that can run, the model takes next timeslot (line 13).

### *B. Local optimal resource usage heuristic – PEC*

In DBH heuristic technology is selected that gives the shortest time of task realization. Such strategy is optimal locally because doesn't take into consideration optimal renewable resource usage. In PEC heuristic (see Pseudocode 2) such aspect is included as some local search method. As DBH only assigns the first task, PEC heuristic tries to assign the larger number of tasks in given timeslot (line 11-17). All analyzed tasks are unassigned for schedule (line 19-23). Only the best task sequence for given timeslot is selected and all included tasks are assigned to final schedule (line 26-29).

To reduce the PEC computation complexity some limits are introduced – the *size* PEC parameter defines number of tasks that are analyzed in one sequence. As *size* parameter equals to 1 PEC heuristic works as DBH, the greater value needs much more CPU working time but returns production schedule more efficient.

The PEC heuristic is a type of compromise between semi-blind greedy DBH and brute force method that analyzes all possible permutations to get the (local) optimal schedule. The *size* parameter gives a range of above compromise to get possible better schedule than build by DBH.

### *C. NEH2 as NEH heuristic adaptation*

Results of experiments with PEC and DBH heuristic showed that tasks sequence for effectiveness of algorithms have big impact on the final schedule. Such observation encourage us to find algorithm which can optimize this aspect. The classical NEH (Nawaz, Enscore, Ham) [14] algorithm is considered as one of the most effective method of minimizing the makespan for Permutation Flowshop Scheduling Problem. The main goal in original NEH is to find the optimal sequence of operations to get more optimal

Pseudocode 1. DBH pseudocode

```
    procedure DurationBasedHeuristic ( SORT_CRITERIA )

1    UT = Tasks               // unsigned tasks
2    TS = 0;                  // timestamp
3    RUT_TS = {}              // sequence of ready to run unassigned tasks
4    do
5     assigned=false
6     RUT_TS = getApplicableTasks (UT, TS)
7     if ( |RUT_TS| > 0 )
8       RUT_TS:= SORT( RUT_TS, SORT_CRITERIA );
9       Task = RUT.getFirstTask();
10      Tech = getShortestDurationApplicableTechnology( Task, TS )
11      assigned = schedule.assign (Task, Tech, TS)
12    if (assigned==true) UT = UT / Task
13    else TS++
14   while ( |UT|>1 )
```

PSEUDOCODE 2. PEC PSEUDOCODE

```
     procedure PEC ( SIZE, SORT_CRITERIA )

1    UT = Tasks           // unsigned tasks
2    TS = 0;             // timestamp
3    RUT_TS = {}          // sequence of possible to run unassigned Tasks
4    do
5     RUT_TS = getApplicableTasks (UT, TS)
6     if ( |RUT_TS| > 0 )
7       RUT_TS:= SORT( RUT_TS, SORT_CRITERIA )
8       RUT_TS:= getFirstNElements( RUT_TS, SIZE)
9       AssignedTasksMax = 0;
10      for all P permutation RUT_TS
11        numberOfAssignedTasks = 0
12        for each T_j task RUT_TS
13           Task = RUT.getFirstTask;
14           Tech = getShortDurAppTechn( Task, TS )
15           assigned = schedule.assign (Task, Tech, TS)
16           if (assigned)  numberOfAssignedTasks++
17        for end
18
19        if ( numberOfAssignedTasks > AssignedTasksMax )
20          AssignedTasksMax = numberOfAssignedTasks
21          BestTaskSequence = P
22
23        schedule.unassignTasks( RUT_TS )
24      end for
25
26      for each Task from BestTaskSequence
27        Tech = getShortDurnApplTech( Task, TS )
28        assigned = schedule.assign (Task, Tech, TS)
29        if (assigned) UT = UT / Tech
30      end for
31    end if
32    TS++;
33   while ( |UT| > 0)
```

schedule. As evaluation can be applied makespan or other schedule measure. In this paper model RCPSP some NEH modification must be implemented.

The basic version of NEH heuristic builds schedule partially to find optimal sequence of tasks adding next task to partial schedule, finally composing the whole schedule. In our approach NEH is considered rather as metaheuristics that proposes sequence of tasks that make schedule optimal (see Pseudocode 3). The other algorithm schedules task to build partial schedule – we decided to use the classical greedy algorithm. The best task sequence is marked as base task sequence (line 16), that is extended by next tasks probing all positions in the task sequence. Let's analyze the NEH working illustration. Having task A and task B, NEH executes *greedyAlgorithm* to find optimal tasks sequence (AB or BA). Let's assume that BA is optimal, to extend sequence BA adding new task C the *greedyAlgorithm* probes sequences: CBA, BCA and BAC and so on.

The basic NEH procedure is too time-consuming to apply in real world application. The next step of NEH implementation was to optimize its computational complexity. The most expensive operation is *GreedyAlgorithm* and this factor should be reduced. We observed that *GreedyAlgorithm* builds each time the whole schedule which is a huge extravagance. The next step was to use partially build schedule and reverse task sequence build strategy.

PSEUDOCODE 3. NEH2 PSEUDOCODE

```
     procedure NEH2 (SORT_CRITERIA)

1    UT = Tasks          // unsigned tasks
2    BestTS = <>         // best sequence according to MAKESPAN
3    CurrentTS = <>      // current (candidate) tasks sequence
4    BaseTS = <>         // base task sequence
5
6    UT = SORT( UT, SORT_CRITERIA )
7    BestTS = CurrentTS = UT.getFirstTask()
8    for each Task from UT
9     CurrentTS = BaseTS
10    for each position i=|CurrentTS| insertion Task into CurrentTS
11              for each position CurrentTS to i
12                    rTask = CurrentTS.removeTask(i)
13                    ReTasks.addTask(rTask)
14             end for
15             value = GreedyAlgorithm(CurrentTS, ReTasks, Task)
16      if (value < bestValue or i==|CurrentTS|)
17        bestValue = value
18        BestTS = CurrentTS + Task + ReTasks
19      end if
20    BaseTS = BestTS
21   end for
```

For example, in basic NEH for three task (A, B and C, let assume that BA sequence is an optimal) the analyzed sequences are: CBA, BCA and BAC. In reverse order in NEH2 we build BA schedule as base, then BAC. In next sequence BCA from schedule is removed task A, then inserted C and A. In basic NEH to examine three task sequence 9 task is scheduled, in reduced version (NEH2, see Pseudocode 4) only 6 tasks is (re)scheduled.

The main modification of NEH2 heuristic is to remove from the initial sequence and reschedule only tasks that are next inserted (see line 11-14). To evaluate the partial schedule is build by *GreedyAlgorithm* (line 15) to examine the sequence of tasks.

The NEH2 computation complexity reduction makes possible practical application of heuristic in simpler cases. Such NEH2 heuristic has been examined. Results of test are presented in the next section.

## IV. EXPERIMENTS AND RESULTS

The t-RCPSP model is specialized to MP2 company requirements. All proposed methods in verification procedure need an empirical data. We analyzed real data and prepared dataset that is complete for the domain: various number of tasks, machines, devices and technologies. Such dataset allows us to do research and compare results of proposed methods.

All experiments are implemented in standard C/C++. Machine for test was equipped with Intel Core2 Duo 2.53 GHz, 4GB RAM and Windows7 OS. For each experiment, only one Core was used.

### A. *Experiments' set-up and dataset*

Prepared dataset MP2dataset[1] consists of seven various types of configurations – summary of dataset is presented in Table1. There are three types of simple settings (10_3x3_10, 50_10x20_40 and 75_10x20_40) that involve small number

TABLE I.
SUMMARY OF TESTING DATASET MP2DATASET

| | tasks | machines | devices | technologies |
|---|---|---|---|---|
| **10_3x3_10** | 10 | 3 | 3 | 10 |
| **50_10x20_40** | 50 | 10 | 20 | 40 |
| **75_10x20_40** | 75 | 10 | 20 | 40 |
| **100_30x30_100** | 100 | 30 | 30 | 100 |
| **200_30x30_100** | 200 | 30 | 30 | 100 |
| **300_30x100_500** | 300 | 30 | 100 | 500 |
| **500_30x45_100** | 500 | 30 | 45 | 100 |
| Legend: tasks _ machines x devices _ technologies | | | | |

[1] MP2dataset is published in http://imopse.ii.pwr.edu.pl/

of tasks (respectively 10 or 50) and small number of devices, less than 20. Two configurations have medium difficulty (100_30x30_100 and 200_30x30_100) where number of tasks is larger (100 or 200) and there is increased number of possible technologies to 100. Additionally, the configuration 300_30x100_500 is difficult because of large number of tasks and extremely a lot of technologies (500) and devices (100). The last configuration consists of 500 tasks, which is the most difficult for methods testing.

To get dataset more general, for each configuration 100 instances were generated. The data generator constructs instances randomly according to the specific domain requirements and configuration. Analyzing the real data we assumed some additional dataset parameters: task deadline *ddt* in <10,50> defined in discrete timeslots, the maximal number of generated products by technology is 5. Each technology can produce no more that 2 types of products and use no more than 4 types of materials. Each product can be generated by 2 or more technologies. The longest technology duration not exceeds 10 timeslots.

### B. *Experiments results*

The main goal of provided experiments was to investigate how presented methods are effective in solving t-RCPSP. The method's results are described by makespan and averaged latency of all tasks in given schedule. The other comparative aspect was computational CPU time needed by methods to obtain results. All methods were investigated using MP2dataset and results were averaged to compare to others (see Table II). Research consists all examined methods DBH, NEH2 and PEC using one main parameter, sorting criteria: by task deadline ascending, descending and random. The PEC uses extra parameter: *size* of task list.

Experiments results presented in Table II give information that all developed heuristics are useful in solving t-RCPSP. In 4/7 cases the minimal makespan provided DBH, however NEH2 in such cases gives slightly worse results and in 3/5

TABLE II.
AVERAGED SCHEDULE MAKESPAN (AND STANDARD DEVIATION) FOR MP2DATASET

| Tasks_mach.<br>device_techn | **10_3**<br>**x3_10** | **50_10**<br>**x20_40** | **75_10**<br>**x20_40** | **100_30**<br>**x30_100** | **200_30**<br>**x30_100** | **300_30**<br>**x100_500** | **500_30**<br>**x45_100** |
|---|---|---|---|---|---|---|---|
| **DBH asc** | 185,53<br>41,93 | 189,43<br>45,52 | 252,65<br>55,36 | 190,21<br>45,13 | 336,88<br>73,93 | **214,72**<br>**22,16** | 899,0<br>231,76 |
| **DBH dsc** | **184,9**<br>**41,7** | **183,9**<br>**48,3** | 252,5<br>61,2 | 190,4<br>41,4 | 333,4<br>69,9 | **214,7**<br>**18,5** | **880,0**<br>**236,4** |
| **DBH rand** | 191,31<br>42 | 192,84<br>48 | 253,57<br>58 | 194,12<br>45,9 | 338,48<br>72,95 | 217,07<br>20,04 | 920,0<br>250 |
| **PEC(3) asc** | 190,6<br>43,0 | 189,4<br>43,9 | **250,1**<br>**53,5** | 196,1<br>48,1 | 345,6<br>68,0 | 271,5<br>18,9 | 915,9<br>251,5 |
| **PEC(3) dsc** | 189,6<br>41,9 | 185,0<br>47,1 | 254,8<br>60,1 | 192,2<br>41,9 | 342,1<br>73,3 | 273,3<br>17,9 | 906,4<br>229,1 |
| **PEC(3) rand** | 193,0<br>43 | 189,44<br>46 | 256,87<br>54 | 196,33<br>44 | 339,9<br>70 | 257,63<br>19,6 | 927,51<br>242 |
| **PEC(4) asc** | 187,28<br>40 | 187,65<br>44 | **250,56**<br>**58** | 192,43<br>45 | 342,08<br>73 | 259,6<br>19,87 | 926,0<br>245 |
| **PEC(4) dsc** | 187,1<br>43,9 | 185,55<br>49,14 | 251,21<br>59,04 | 191,45<br>39,21 | 340,80<br>70,05 | 260,83<br>17,74 | 894,6<br>232,42 |
| **PEC(5) asc** | 187,04<br>39,04 | 190,2<br>41 | 251,78<br>59 | 192,14<br>49,9 | 344,57<br>71,93 | 249,11<br>19,29 | 920,65<br>242 |
| **PEC(5) dsc** | 190,02<br>41,85 | 185,78<br>48,36 | 254,4<br>61 | 191,0<br>40,8 | 339,09<br>71,67 | 245,78<br>17,53 | 896,9<br>226 |
| **NEH2 asc** | 185,5<br>41,9 | 189,3<br>45,7 | 252,7<br>55,2 | **189,7**<br>**46,1** | 336,8<br>73,0 | *time limit exceeded* | *time limit exceeded* |
| **NEH2 dsc** | 185,17<br>41,29 | **183,89**<br>**48,33** | 252,72<br>61,12 | **189,38**<br>**41,28** | **331,84**<br>**69,3** | *time limit exceeded* | *time limit exceeded* |

TABLE III.
AVERAGED SCHEDULE LATENCY (AND STANDARD DEVIATION) FOR MP2DATASET

| Tasks_mach.<br>device_techn | **10_3**<br>**x3_10** | **50_10**<br>**x20_40** | **75_10**<br>**x20_40** | **100_30**<br>**x30_100** | **200_30**<br>**x30_100** | **300_30**<br>**x100_500** | **500_30**<br>**x45_100** |
|---|---|---|---|---|---|---|---|
| **DBH asc** | 0,30<br>0,52 | 0,53<br>0,89 | 2,03<br>2,32 | 0,43<br>0,79 | **3,45**<br>**2,57** | 0,12<br>0,15 | **44,125**<br>**19,70** |
| **DBH dsc** | 2,39<br>1,60 | 3,65<br>3,19 | 8,66<br>4,37 | 2,57<br>2,05 | 10,48<br>4,06 | 7,76<br>2,00 | 74,31<br>23,53 |
| **DBH rand** | 1,6<br>1,37 | 2,2<br>2,25 | 5,83<br>3,83 | 1,62<br>1,42 | 7,63<br>3,59 | 4,70<br>1,57 | 63,32<br>19,9 |
| **PEC(3) asc** | 0,29<br>0,5 | 0,42<br>0,9 | 2,05<br>2,27 | 0,52<br>1,03 | 3,88<br>2,59 | **0,10**<br>**0,15** | 49,64<br>21,10 |
| **PEC(3) dsc** | 2,72<br>1,59 | 3,80<br>3,26 | 9,13<br>4,56 | 2,82<br>2,07 | 12,13<br>4,2 | 21,97<br>3,45 | 96,07<br>20,85 |
| **PEC(3) rand** | 1,73<br>1,4 | 2,58<br>2,82 | 5,93<br>3,47 | 1,82<br>1,62 | 8,1<br>3,5 | 9,0<br>2,2 | 68,12<br>19 |
| **PEC(4) asc** | **0,26**<br>**0,44** | 0,47<br>0,92 | 2,02<br>2,75 | 0,5<br>1,0 | 3,72<br>2,72 | 0,13<br>0,19 | 48,33<br>20,65 |
| **PEC(4) dsc** | 2,53<br>1,63 | 3,94<br>4,53 | 8,94<br>4,53 | 2,64<br>2,00 | 11,64<br>4,44 | 18,65<br>3,2 | 90,27<br>21,81 |
| **PEC(5) asc** | 0,28<br>0,53 | **0,40**<br>**0,74** | **2,00**<br>**2,57** | 0,47<br>0,9 | 3,52<br>2,6 | 0,14<br>0,17 | 47,34<br>19,57 |
| **PEC(5) dsc** | 2,6<br>1,6 | 3,85<br>3,22 | 9,05<br>4,43 | 2,63<br>2,03 | 11,13<br>4,22 | 15,34<br>2,94 | 87,10<br>21,0 |
| **NEH2 asc** | 0,30<br>0,52 | 0,52<br>0,89 | 2,05<br>2,33 | **0,42**<br>**0,76** | 3,47<br>2,6 | *time limit exceeded* | *time limit exceeded* |
| **NEH2 dsc** | 2,39<br>1,59 | 3,65<br>3,19 | 8,66<br>4,35 | 2,55<br>2,04 | 10,5<br>4,11 | *time limit exceeded* | *time limit exceeded* |

TABLE IV.
AVERAGED COMPUTATIONAL TIME [S] FOR MP2DATASET

| Tasks_mach.<br>device_techn | **10_3**<br>**x3_10** | **50_10**<br>**x20_40** | **75_10**<br>**x20_40** | **100_30**<br>**x30_100** | **200_30**<br>**x30_100** | **300_30**<br>**x100_500** | **500_30**<br>**x45_100** |
|---|---|---|---|---|---|---|---|
| **DBH** | 0,14 | 0,05 | 0,12 | 0,13 | 0,57 | 4,8 | 7,13 |
| **PEC(3)** | 0,11 | 0,05 | 0,12 | 0,08 | 0,39 | 2,58 | 4,11 |
| **PEC(4)** | 0,10 | 0,05 | 0,12 | 0,1 | 0,41 | 2,95 | 4,71 |
| **PEC(5)** | 0,24 | 0,11 | 0,27 | 0,20 | 0,87 | 6,2 | 9,3 |
| **NEH2** | 5,6 | 1,2 | 3,9 | 5,6 | 49,66 | *time limit exceeded* | *time limit exceeded* |

cases returns solutions that compete with others. The NEH2 in more complicated cases (300 and 500 tasks) execution time exceeded 1 minutes CPU time and results are not taken into consideration. PEC only in one case returns the best solution (75 tasks, PEC(3) asc). Increasing the PEC *size* parameter value in most cases reduces makespan, e.g. comparing results of PEC(3) and PEC(4) using descending task deadline order. To summary results of all methods for all instances: DBH dsc needs average 319,97 timeslots to execute all 700 instances, DBH asc needs 324,06 timeslots and PEC(5) dsc has the third place: 328,9 timeslots. The longest averaged makespan schedule (equals to 337 timeslots) achieved PEC(3) heuristic with ascending task order. Analysis of the results presented into Table II can draw the conclusion that descending sorting criteria of tasks gives better results in the minimization of schedule makespan. Random tasks order makes solution the worst in all investigated cases. All methods are deterministic, the averaged results are computed on 100 instances of each configuration. In case of random tasks order, results for each instances are repeated 10 times and then averaged.

Results presented in Table III describe how generated schedules are late using as measure the average latency of all tasks in the schedule. The gained results proved our intuition that the best results give ascending sorting criteria of tasks – task with the shorter deadline is taken into consideration earlier. All methods showed that are effective, but it is rather impossible to point the best one. In 2/7 cases DBH gives the best solution, PEC in 4/7 cases (using *size* parameter equals to 3, 4 or 5). Results gained by NEH2 are not qualitative. Moreover, NEH2 in one case returns the best solution (100 tasks configuration). Comparing the averaged latency for 700 instances (whole MP2dataset) the best method gives 7,28 latency (DBH asc), the second one 7,73 (PEC(5) asc) and third one 7,91 (PEC(4) asc). The worst averaged latency achieved PEC(3) dsc: 21,23.

Comparing methods' working time (see Table IV) it is worth mentioning that methods are fast and effective. The provided MP2dataset of 700 instances gives an opportunity to compare methods results and recommend them to real-world applications. Increasing size of the problem, methods are practical as computational time not exceeds 10 seconds. Such short computational time makes possible to run several methods to get set of schedules and give a human operator a real choice.

The computation complexity of presented heuristics is $O(k^2)$ for DBH and $O(size!k^2)$ for PEC. The NEH2 complexity is larger because core of NEH2 is $O(k^2)$ but in each step uses *greedyAlgorithm* that is $O(k)$, what gives finally NEH2 $O(k^3)$. The result is that the computational time of NEH2 increases so dramatically that needed time is unacceptable in construction process of schedules that consist of more that 200 tasks. In such cases other methods are more effective and less demanding for CPU working time.

## V. Summary

Standard RCPSP, in presented work, was extended by technologies to solve practical problem in Plastic and Rubber Processing Industry – we defined t-RCPSP. Such model makes possible, in simply and intuitive way, a formalization of real-world problem. Each technology links non-renewable and renewable resources, uses various types of ingredients in production. Technologies that produce the same products may use resources in different way, more or less effectively. It can case be other production time consumption, too. The technology-driven RCPSP model gives a lot of possibilities to build effective heuristics. We proposed three of them: NEH2 adaptation, locally optimal resource usage PEC and duration driven heuristic DBH. Analyzing a real production schedule instances we implemented a data generator to get the MP2dataset (published in Internet) that includes 700 instances in 7 basic problem configurations to empirically prove efficiency of proposed methods. Results of experiments showed that proposed methods can be used as effective tool for scheduling in production company. Moreover, the next step was done: all presented methods were developed as automated scheduling module in MP2 company computer system.

### A. Further research

As the practical aspect of further research is the definition of more domain-based measures of final schedule and used technologies. In presented paper, only the makespan and latency are analyzed as the measure. However, in practice such technology can be more energy consuming, may need more labor (including human work) or can be less effective as a scrap measure is considered. The existence of several measures of schedule leads to the situation when multiobjective optimisation should be considered.

Investigating the comparison of presented methods we can see some possible directions of further work. The most promising is using metaheuristics (such as Evolutionary Algorithms or Tabu Search) to build schedule near optimal in cost/time criterion. Metaheuristics usage needs more computational time than simple heuristic – but our experience (e.g. Error: Reference source not foundError: Reference source not found) shows that results are (sub)optimal. Additionally, metaheuristics are using evaluation function (as the superposition of several schedule measures) can provide schedule dedicated to given user. Especially Tabu Search [12][15][16][20] application to RCPSP is very strong trend in literature.

## Acknowledgment

The research has been partially supported as a part of the Support Programme of the Partnership between Higher Education and Science and Business Activity Sector finances by City of Wroclaw.

## References

[1] Blazewicz J., Lenstra J.K., Rinnooy Kan A.H.G.; Scheduling subject to resource constraints: Classification and complexity, Discrete Applied Mathematics (5), pp. 11-24, 1983.

[2] Bouleimen K., Lecocq H.; A new efficient simulated annealing algorithm for the resource-constrained project scheduling problem and its multiple mode version, Eur. J of Operational Research (149), pp. 268-281, 2003.

[3] Brucker P., Drexl A., Mohring R., Neumann K., Pesch E.; Resource–constrained project scheduling: Notation, classification, models, and methods, European Journal of Oper. Research (112), pp. 3–41, 1998.

[4] Dorigo M.; Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem, IEEE Transactions of Evolutionary Computation (1/1), pp. 53-66, 1997.

[5] Hartmann S.; A competitive genetic algorithm for resource–constrained project scheduling, Naval Research Logistics (45), pp. 733–750, 1998.

[6] Hartmann S., Briskorn D., A survey of variants and extensions of the resource-constrained project scheduling problem, European Journal of Operational Research 207(2010), pp.1-14. 2010.

[7] Kolisch R., Hartmann S., Experimental evaluation of state-of-the-art heuristics for the resource- constrained project scheduling problem, European Journal of Oper. Research (127), pp. 394–407, 2000.

[8] Kolisch R., Hartmann S., Experimental investigation of heuristics for resource-constrained project scheduling: An update, Euro. Journal of Oper. Research (174), pp. 23-37, 2006.

[9] Liang Y., Chen A., Kao W., Chyu C.; An Ant Colony approach to Resource–Constrained Project Scheduling Problems, Proc of the 5th Asia Pacific Indust. Eng. and Manag Systems Conf 2004, pp. 31.5.1-31.5.10, 2004.

[10] Luo S., Wang C., Wang J.; Ant Colony Optimization for Resource-Constrained Project Scheduling with Generalized Precedence Relations, Proc of the 15th IEEE International Conference on Tools with A(ICTAI03), pp. 284–289, 2003.

[11] Merkle D., Mittendorf M., Schmeck H.; Ant Colony Optimization for Resource–Constrained Project Scheduling, IEEE Transactions on Evolutionary Computation (6/4), pp. 333–346, 2002.

[12] Myszkowski P.B., Skowroński M. E., Myszkowski P. B., Kwiatek P., Adamski M., Tabu Search approach for Multi-Skill Resource-Constrained Project Scheduling Problem, Annals of Computer Science and Information Systems Volume 1, Proc. of the 2013 FeDCSIS Confeences, pp. 153-158, 2013.

[13] Myszkowski P.B., Skowronski M.E., Olech Ł.P. and Oślizło K., Hybrid ant colony optimization in solving multi-skill resource-constrained project scheduling problem, Soft Computing Journal, Sep 2014.

[14] Nawaz, M., Enscore, J., Ham, I.: A Heuristic Algorithm for the M-machine, N-task Flow-shop Sequencing Problem. Omega-Int. J. Manage. S. 11(1), 91–95 (1983)

[15] Pan H.I., Hsaio P.W., Chen K.Y.; A study of project scheduling optimization using Tabu Search algorithm, Engineering Applications of Artificial Intelligence (21), pp. 1101-1112, 2008.

[16] Pan N.H., Lee M.L., Chen K.Y.; Improved Tabu Search Algorithm Application in RCPSP, Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol I), 2009.

[17] Santos M., Tereso A. P.; On the multi-mode, multi-skill resource constrained project scheduling problem - computational results, Soft Computing in Industrial Applications, Advances in Intelligent and Soft Computing (96), pp. 239–248, 2011.

[18] Skowroński M. E., Myszkowski P. B., Specialized genetic operators for Multi-Skill Resource-Constrained Project Scheduling Problem, 19th Inter. Conference on Soft Computing Mendel 2013, pp. 57-62, 2013.

[19] Skowroński M. E., Myszkowski P. B., Podlodowski L., Novel heuristic solutions for Multi-Skill Resource- Constrained Project Scheduling Problem, Annals of Computer Science and Information Systems Volume 1, Proc. of the 2013 Federated Conference on Computer Science and Information Systems, pp. 159-166, 2013.

[20] Thomas P. R., Salhi S.; A Tabu Search Approach for the Resource Constrained Project Scheduling Problem, Journal of Heuristics (4), pp. 123-139, 1998.

[21] Valls V., Ballestin F., Quintanilla S.; A hybrid genetic algorithm for the resource–constrained project scheduling problem, European Journal of Operational Research (185), pp. 495-508, 2008.

[22] Zhang H., Xu H., Peng W., A Genetic Algorithm for Solving RCPSP, 2008 International Symposium on Computer Science and Computational Technology, pp. 246–249, 2008.

[23] Zhang H., Li H., Tam C.; Particle swarm optimization for resource–constrained project scheduling, In- ter. Jour. of Project Management (24), pp. 83-92, 2006.

[24] Zhang K., Zhao G., Jiang J.; Particle Swarm Optimization Method for Resource-Constrained Project Scheduling Problem, The Ninth International Conference on Electronic Measurement & Instruments ICEMI2009, pp. 792–796, 2009.

[25] Ziarati K., Akbari R., Zeighami V.; On the performance of bee algorithms for resource–constrained project scheduling problem, Applied Soft Computing (11), pp. 3720–3733, 2011

DOI: 10.15439/2015F273
ACSIS, Vol. 5

# A new benchmark dataset for Multi-Skill Resource-Constrained Project Scheduling Problem

Paweł B. Myszkowski, Marek E. Skowroński, Krzysztof Sikora
Institute of Informatics, Department of Artificial Intelligence
Faculty of Computer Science & Management, Wrocław University of Technology, Poland
Email: {pawel.myszkowski, m.e.skowronski, krzysztof.sikora}@pwr.edu.pl

***Abstract*—In this paper novel project scheduling difficulty estimations are proposed for Multi-Skill Resource-Constrained Project Scheduling Problem (MS–RCPSP). The main goal of introducing the complexity estimations is an attempt of estimation the project complexity before launching the optimization process. What is more, the dataset instance generator is also presented as a tool to create new instances for extending the research area. Furthermore, the dataset proposed in previous works is extended by new instances, described thoroughly and released as a benchmark dataset. The dataset instances are also scheduled using simple heuristic and greedy algorithm in duration- and cost- oriented optimization modes. Finally, a brief summary of investigated methods and potential further research directions is presented.**

***Index Terms*—scheduling, RCPSP, dataset, benchmark, heuristics, indicator**

## I. Introduction

Resource-Constrained Project Scheduling Problem stands as one of the most important and the most investigated [9], [10] kind of known types of scheduling problems. It is because of its practical nature and the need to find good ways for resolving it not only for scientific, but also industry purposes. Its goal is to find the best schedule for the project, by assigning scarce resource to defined tasks. The quality of the schedule is mostly defined as its duration, cost or some combinations of those indicators.

As MS–RCPSP is the extension of classical RCPSP, it makes the problem NP–hard [2]. Hence, there is no way to find a method that would be able to find the optimal solution in polynomial, reasonable time. Therefore, one of the main approach to solve RCPSP and its potential extensions is to use soft computing methods, especially metaheuristics [19].

To make the problem definition more practical in industrial point of view, we introduced the skill domain. Tasks require some specified skill to be performed by resources owning some subset of skills defined in the project. Therefore, not every resource is able to perform every task in the project. It makes the problem more constrained but on the other hand – more realistic. RCPSP extended by skills domain is called Multi–Skill RCPSP (MS–RCPSP).

The goal of the paper is to present several indicators for the difficulty of the project to be scheduled. The difficulty could be understood as a measure how much the solution space is constrained – how hard is to build feasible and good enough schedule. The secondary objective is to share the dataset and propose it as a benchmark for other researchers, to build a common platform for evaluating methods solving MS–RCPSP.

The rest of the paper is organised as follows. Section II presents some approaches in solving MS–RCPSP using some of the mentioned metaheuristics. Then section III describes the problem statement. Then Section IV presents complexity estimations we proposed for MS–RCPSP. Section V describes the way how new instances are generated. Furthermore in the Section VI the dataset has been presented and then its instances have been used in experiments in Sec. VII while the last Section VIII describes approaches we have recently investigated and proposes ways for further research.

## II. Related work

NP–hard [2], combinatorial nature of MS–RCPSP is one of the reasons of common use of metaheuristic–based approaches in solving the problem. Nevertheless, some constraint programming methods or simpler heuristics are also used to solve this kind of problems [20].

However, there is still lack of papers regarding multi–objective Multi–Skill extension of RCPSP. Some approaches solving MS–RCPSP in project duration domain [1], [16] or project cost domain [12] could be found. On the other hand, there are methods solving classical RCPSP extended by cost domain, but without skills considerations. Such research has been presented in [15], [6], [4], [13] and [24]. Hence, we have decided to combine those two elements: multi–objective optimization and multi–skill domain for project scheduling problem.

Although classical RCPSP is deeply investigated and numerous approaches could be easily compared using PSPLIB instances, it is very hard to find multi–objective MS–RCPSP methods working on datasets that could be regarded as a benchmark. Some papers describe instances artificially generated ([5], [16]), while some others propose methods of PSPLIB dataset adaptation ([1], [3], [7], [12]). We analysed some published benchmark datasets, but they were usually unsuitable for our approach as they do not cover multi-objective nature of the problem, even multi–skill domain has been developed ([23]). The other unsuitable example is a benchmark for Multi–Mode RCPSP (MM-RCPSP) published in [21], but it does not involve skills constraints and make the main focus on multi–mode characteristics. Hence, the need of definition new dataset has arisen.

978-83-60810-66-8/$25.00©2015, IEEE

Some difficulty estimations for any project scheduling problems could be found in [11] or [13]. However those proposed difficulty estimations based mostly on tasks, precedence relations between them and resource properties. There is a lack of difficulty estimations that would be dedicated for MS–RCPSP, involving skills domain.

Among many papers regarding the resource – constrained project scheduling problems and its extensions, we found that we had something in common with the approach presented in [13]. Despite some similarities, there are some crucial elements that make our approach, defined in detail in [14], different. We regard both of them as worth of investigating. Tab. I presents the comparison – similarities and differences in four main areas we decided to point out.

Based on the information in the Tab. I, some common elements between our approach and the one presented in [13] in all of the mentioned areas can be found. Firstly, investigated problems are similar in a way that both of them regard multi–objective optimization in Multi–Skill Resource–Constrained Project Scheduling Problem (MS–RCPSP). Both problems are additionally defined by some resource, precedence and skills constraints. However those constraints differ in details (i.e. distinguishing skill types).

There are also some similarities regarding the dataset published. First of all, both of the datasets are published in the Internet, so anyone has an access to dataset instances and can use them to investigate own optimization methods in MS–RCPSP. What is more, the number of dataset instances is the same. However, the strategy of building the dataset was different. We used information about number of different skill types and number of precedence relations. Not only the most common indicators, like number of tasks or number of resources, what can be found in [13]. Based on those additional indicators we tried to build as most balanced dataset as possible. Therefore, we tried to adjust the number of resources, skills and precedence relations in a way to make our complexity indicators similar for 100 and 200-tasks project instances as well.

## III. Problem statement

The goal of the MS–RCPSP is to order given set of tasks and assign resources to them in a way to provide feasible and as good solution as possible. The quality of the solution could be measured in its duration, cost or any other measure defined according to business requirements.

In MS–RCPSP the set of tasks ($J$) is given, while every task has to be performed during the project execution. Each task is described by its start ($S_j$) and finish dates ($F_j$), duration ($d_j$) and skill required by it to be performed. What is more, tasks can be related between themselves by precedence relations. It means some tasks (successors) cannot start before some other would be finished (predecessors). It makes the solution space for given instance more constrained, as there are fewer possibilities to put the tasks in given period. Predecessors of given task $j$ are defined as $P_j$ while overall number of predecessors in a project is $p$.

Furthermore, the set of resources ($K$) is also given. Every resource $k$ is described by its salary ($s_k$) and skills covered ($Q^k$). Therefore, a subset of tasks than can be performed by $k$ resource could be obtained and is denoted as $J^k$.

Skill required by task to be performed determines which resource can be assigned to it. Every skill type could appear in a project in various familiarity levels, denoted as an integer value from 0 to 4. Resources with skill type required by given task but on the lower than required level cannot be assigned to such task. The number of all skills (including different familiarity levels) is denoted as $q$, while the number of skill types in the project is denoted as $\bar{q}$

What is more, any resource cannot be assigned to more than one task in the overlapping period. If such a situation occurs, **conflict** is detected and has to be resolved, to get valid, feasible solution. It is made by shifting some of conflicting tasks in time–line in a way to make it start just after another conflicted task would be finished. The decision which of conflicted tasks should be taken to be shifted is made by checking which has been previously added to the project definition, because we do not distinguish various levels of task priority. Each task is equally important to be scheduled in the project.

Any project schedule ($PS$) has to be conflicts–free and has to satisfy the precedence constraints between tasks. If it satisfied those both kind of constraints, we would call it as a **feasible** schedule. Only feasible ones can be regarded as finite solutions. What is more, every infeasible solution could be made feasible. However, making schedule feasible could make its duration longer.

### A. Calendar restrictions

Due to use Microsoft Project as a base for our dataset, we needed to obtain some calendar restrictions that are strictly related to used software.

First of all, standard calendar in Microsoft Project is designed to handle projects in real–life, where classical five–days week of work is used. It was also a requirement asked by the VolvoIT enterprise that we cooperate in the research field of project scheduling. However, weekends are taken into account. If resource is assigned to task that cannot be finished before weekend, it will be finished after the weekend. The task duration is bigger but number of man–hours (or man–days) required for given task does not change. Therefore various project duration measures can be obtained. One can be made based on the overall duration of project – between its start and finish dates, including weekends where tasks are not performed but influence on other tasks' start dates and the project finish. Other approach could be to ignore any festivals and weekends and regard seven days week of work. Until now we prefer the first approach as it is more practical.

Furthermore, the localization issue has to be taken into account when considering calendar restrictions. Depending on the localization settings, some changes in the calendar could appear, regarding some national or cultural–related festivals. In example, there would be other free days in China than in Poland, where different festivals are taken place.

TABLE I
SIMILARITIES AND DIFFERENCES BETWEEN IMOPSE [14] APPROACH AND THE APPROACH PRESENTED IN [13]

| Area | Similarities | Differences |
|---|---|---|
| Problem definition | Multi-skill<br>Multi-objective<br>Resource-constrained<br>Precedence relations<br>Skills<br>Minimal one resource required by given task to be performed<br>Repair operator → enlarging project duration | Resource load - 'dedication' measure<br>No skill levels<br>Task can be assigned to more than one resource<br>No conflict, different rule of repair operator<br>Approach more academic than practical |
| Dataset | Same number of instances: 36<br>Published in the Internet | Instances distinguished only by number of tasks and number of resources<br>Constant vs. varied number of skill types in a dataset instance |
| Generator | Published in the Internet as a benchmark<br>Developed in the JAVA programming language | No graphical user interface<br>Output format not connected with MS Project |
| Methodology | Time vs. cost tradeoff<br>Focus on multiobjective, Pareto–based optimization | Complexity estimators vs. hypervolume, attainment |

Linking above constraints with potential dynamic date of project start – the date, when the first task is assigned to resource in the timeline – there is a risk that the same project with the same task–to–resource assignments (schedule) can be finished in various dates, depending on the day of start. Project instances start at various dates, except the ones with D* suffix that have been prepared strictly for given enterprise and were required to be start all at the same day. It has been set to 12th April of 2012.

To avoid those calendar restrictions .def format has been introduced. It is described in detail in Subsec. VI-A.

### B. Evaluation function

The goal of MS–RCPSP is to find the best (as quick or / and as cheap as possible) final project schedule. Hence, we could present it as bi–objective optimization problem. Because of totally different domains of duration and cost, we cannot simply aggregate those two objectives. Therefore, the normalization process is performed, to get the value scope between 0 and 1. It allows us to aggregate those two objectives and combine them into one evaluation function.

We have also preserved the possibility to choose which objective is more important in given optimization process. It is made by setting weights both for the duration ($\omega_\tau$) and cost aspect. The sum of both weights sum to 1 and the scope of values is from 0 to 1. It means that setting the weight of duration aspect to one automatically sets the weight of cost to 0 and vice versa. Naturally that weight can be set by float value. Specifically, both weights could be set to 0.5. In that case, both objectives would be equally important in the optimization process. We proposed three baseline weight configurations: duration optimization (DO, $\omega_\tau = 1$), balanced optimization (BO, $\omega_\tau = 0.5$) and cost optimization (CO, $\omega_\tau = 0$) [14].

An important remark is that those objectives are in opposition to each other. It means that setting weights to make the optimization process more cost–oriented could cause getting cheaper project schedule, but with the risk that final schedule would be longer. Analogously, shorter project schedule could be obtained with spending more money on it.

Evaluation function is formulated as follows:

$$\min f(PS) = \omega_\tau f_\tau(PS) + (1 - \omega_\tau) f_c(PS) \tag{1}$$

where: $w_\tau$ – weight of duration component, $f_\tau(PS)$ – duration evaluation component, $f_c(PS)$ – cost evaluation component. Both components are non–negative values, while $w_\tau \in [0; 1]$.

The time component $f_\tau(PS)$ is calculated as follows:

$$f_\tau(PS) = \frac{\tau}{\tau_{max}} \tag{2}$$

Where: $\tau_{max}$ – maximal (pessimistic) possible duration of the schedule $PS$, computed as the sum of all tasks' duration [14]. It occurs when all tasks are performed serially in project: one–by–one. No matter, how many and how flexible resources are.

The cost component $f_c(PS)$ is defined as follows:

$$f_c(PS) = \frac{\sum_{i=1}^{J} c_j - c_{min}}{c_{max} - c_{min}} \tag{3}$$

where: $c_{min}$ – minimal schedule cost – a total cost of all tasks assigned to the cheapest resource, $c_{max}$ – maximal schedule cost – a total cost of all tasks assigned to the most expensive resource [14]. Note: $c_{max}$ and $c_{min}$ do not involve skill constraints. It means that $c_{min}$ value could be reached also for non–feasible solution. Analogously to $c_{max}$.

### C. Solution space size

Given number of tasks and number of resources, we can estimate the solution space size (SS), as:

$$SS(n, m) = n! * m^n \tag{4}$$

Where $n$ is a number of tasks and $m$ is a number of resources [14]. However, that estimation also takes into account non–feasible solutions, because skill–constraints are not satisfied. To give an example, let's assume $n = 10$ and $m = 5$ – without any precedence relations we get $SS(10, 5) = 3.54 * 10^{13}$ combinations. It is worth mentioning that each task can be placed only once in the schedule, but resources could be assigned more often. An extreme situation occurs if the same one resource would be assigned to perform each task.

Large solution space size makes impossible checking each of the combinations manually. However, space includes also non–feasible solutions that do not satisfy defined conditions. Moreover, above example is a simplification and in real world problems we meet a higher number tasks (about $n = 100$) and resources ($m = 20$) – it gives $SS(100, 20) = 1.19 * 10^{288}$ of all solutions.

## IV. Complexity estimations

As a result of cooperation with VolvoIT Department in Wroclaw [18], [20], [19], we defined following elements [14]:

- requirements and constraints dedicated to the industry,
- project scheduling difficulty indicators.

Project difficulty indicators have been verified and approved by experienced project manager in the enterprise.

The main goal of investigating such estimations was to compare how the project elements (tasks, resources, precedence relations, skills) characteristics could influence on the optimization process based on the quality of obtained result (project schedule duration or performance cost) or optimization processing time.

Proposed difficulty estimations are described below. All estimations are normalised before being taken to compute the overall complexity measure.

### A. affiliation ($\lambda$)

States, how much tasks are related between them. The bigger value means the tasks are more related. The project complexity is bigger, because the scheduling flexibility is restricted (more tasks are related to others, so they cannot be scheduled flexibly). It is computed as follows:

$$\lambda = \frac{p}{n} \tag{5}$$

Where $p$ – number of precedence relations, $n$ – number of tasks.

### B. load ($\nu$)

Reflects, how much resources are loaded by tasks. The bigger value means, the more tasks are assigned to one resource (the project complexity is bigger, because the solution space is bigger). It is computed as follows:

$$\nu = \frac{n}{m} \tag{6}$$

Where $m$ – number of resources.

### C. time difference ($\Phi_T$)

Describes how tasks are varied by their duration. The bigger value means tasks are more varied. That makes scheduling more difficult, because tasks' order influences on overall duration time. It is computed as follows:

$$\Phi_T = \frac{\sigma_d}{d_{max} - d_{min}} \tag{7}$$

Where: $\sigma_d$ – standard deviation of tasks' duration in schedule, $d_{max}$ – maximal task duration in schedule, $d_{min}$ – minimal task duration in schedule.

### D. cost difference ($\Phi_C$)

Indicates how tasks are varied by their performance cost. The interpretation is similar to the time difference ($\Phi_T$). It is computed as follows:

$$\Phi_C = \frac{\sigma_C}{c_{max} - c_{min}} \tag{8}$$

Where: $\sigma_C$ – standard deviation of tasks' cost in schedule, $c_{max}$ – maximal task cost in schedule, $c_{min}$ – minimal task cost in schedule.

### E. variety ($\mu$)

Reflects how resources are varied by their skills. The bigger value means the project is more difficult to be scheduled because tasks are more dedicated to resources (no other can be assigned to the specified task). It is computed as follows:

$$\mu = \frac{q}{m} \tag{9}$$

Where: $q$ – number of different skills existing in the project. Important: each level of the same skill name is regarded as a new skill.

### F. universality ($\beta$)

States the average number of resource skills. The bigger value means it is easier to schedule a project because resources are more universal. It is computed as follows:

$$\beta = \frac{\sum_{i=1}^{m} Q^i}{m} \tag{10}$$

Where: $Q^i$ – number of skills owned by $i$ resource.

### G. adjustment ($\pi$)

Shows how many resources available to be assigned in the project are capable of performing tasks that are needed to be performed. Ergo – how many resources can deal with each task. The bigger value means it is more difficult to schedule a project because resources are strictly adjusted to the tasks by their skills covered, and skills needed. It is computed as follows:

$$\pi = \frac{\sum_{i=0}^{Q} \Delta(q_i) * \sigma(\Delta(q))}{max(\Delta(q_1), \Delta(q_2), ..., \Delta(q_q)) * q} \tag{11}$$

Where:

$$\Delta(q_i) = \frac{|Q^i - TQ(i)|}{min(RQ(i), TQ(i))} \tag{12}$$

Where: $RQ(i)$ – number of resources covering skill $i$ (normalized by number of all resources ($m$) in the project). $TQ(i)$ – number of tasks, that require skill $i$ to be performed (normalized by number of all tasks ($n$) in the project).

### *H. Flexibility ($\theta$)*

The flexibility $\theta$ of the instance $PS$ has been estimated as the sum of a number of potential assignments of tasks to a given resource divided by number of resources ($n$). It can be stated as follows:

$$\theta = \frac{\sum_{k=1}^{n} \bar{J}^k}{m} \tag{13}$$

Where $\bar{J}^k$ is the number of tasks that can be performed by resource $k$, while $m$ is the number of resources in a project.

Having discussed the usage of those estimations' legitimacy, each measure has been subjectively weighted and confirmed by an experienced project manager (to determine their priority in overall project's difficulty measure). Having those weights set up, the project's (*PF*) difficulty measure function could be defined as follows:

$$diff(PF) = 8\lambda(PF) + 9\nu(PF) + 3\Phi_\tau(PF) + \\ +3\Phi_C(PF) + 6\mu(PF) - 4\beta(PF) + 7\pi(PF) + 5\theta(PF) \tag{14}$$

The bigger the value $diff(PF)$ is, the more difficult to schedule the project is. Universality measure has been taken with a negative value. It is because the bigger the universality value is, the project is easier to schedule as resources are more skill–flexible and can be assigned to more different tasks, relaxing more skill constraints.

Depending on project manager preferences, weights assigned to given estimations could be changed, what would influence on the overall $diff(PF)$ measure.

## V. Instances generator

The main goal of implementing the dataset instance generator is to provide other researchers the possibility to investigate their methods not only on proposed dataset instances, but also on some other that would be created individually by given researcher. Dataset instance generator has been prepared for MS–RCPSP but it can be easily adjusted to handle classical RCPSP instances like PSPLIB [8]. It has been implemented in JAVA programming language, using MPXJ[1] library for processing project files from MS Project. It can create project definition not only in .mpp (XML) format, but also the simpler (.def) one. The more detailed description of .def format is available in Subsec. VI-A. Instances have been created based on the real–life project instances got from international entreprise (Volvo IT).

Instances generator is an element of resources developed in our **Intelligent Multi–Objective Project Scheduling Environment**[2] platform. Besides instances generator, the platform contains solution validator (see Subsec. VI-B), instances we generated and used to verify our approaches and the best found solutions for those instances in three above–mentioned optimization modes: DO, BO, CO. Every solution is saved in ready–to–use in MS Project .xml format, containing all tasks, resources, skills, precedence relations and obtained schedule.

The general process of generation new instances could be split into main steps:

1) Read and validate parameter values provided by the end user
2) Define resources,
3) Define skills and assign them to resources,
4) Define tasks and precedence relations,
5) Assign resources if necessary,
6) Save project.

In the further parts of this section, following steps would be described in detail. The pseudocode of the generator has been presented in Alg. 1

**Algorithm 1** Generator pseudocode

```
 1: $pool \leftarrow \emptyset$
 2: #generate_resources
 3: for $r \in K$ do
 4:     #generate_resource
 5:     $set\_standard\_salary(minSt, maxSt)$
 6:     #set_skills($r_i$)
 7:     $q \leftarrow setNumSkills(min, max)$
 8:     for $j = 0; j < q$ do
 9:         $set\_skill\_type\_from\_range(minST, maxST, q_j)$
10:         $set\_skill\_level\_from\_range(minSL, maxSL, q_j)$
11:         if $skill\_not\_exists(q_j, pool)$ then
12:             $pool \leftarrow pool.add(q_j)$
13:         $r \leftarrow addSkill(q_j)$
14: #generate_tasks
15: for $t \in J$ do
16:     $t \leftarrow set\_duration(minDuration, maxDuration)$
17:     $t \leftarrow set\_skill(pool)$
18: #generate_relations
19: for $i \in P$ do
20:     $relSource \leftarrow rand(T)$
21:     $relDest \leftarrow rand(relSource, T, min, max)$
22:     $relSource \leftarrow addPredecessor(relSource)$
23: if $make\_assignments$ then
24:     #assign_resources [initial schedule builder]
25:     for $n \in J$ do
26:         $R \leftarrow capable\_resources(n)$
27:         $r' \leftarrow rand(R)$
28:         $assign(n, r')$
29: $save\_result$
```

### *A. Resources*

The number of resources that would be generated is provided as a parameter for the proposed tool. For every generated resource its standard rate salary is set as a random between the minimal and maximal value (see Alg. 1, line: 5) set by the end–user in the configuration of the generator.

### *B. Skills*

Analogously to resource definition, the number of different skill types is set by the end–user during the configuration. We

[1] http://mpxj.sourceforge.net

[2] http://imopse.ii.pwr.edu.pl

declared 4 levels of the skill familiarity for given resource. However, the end–user is also obliged to define how many types of skills could be covered with given resource. It is desired that number of skill types owned by resource would be no greater than the number of skill types existing in the project. Number of skill types is set randomly from the minimal and maximal value (set by end–user). However, during recent dataset instances generation, we decided to make the number of skill types for every resource as a constant – minimal ($min$) and maximal ($max$) number of skill types have been set as the same value – line 7.

During skill generation process for given resource, a skill type is selected from given range of types (line: 9) while skill level is also selected from given scope (line: 10). We decided to make four levels of skills as it covered the requirements presented by project manager from the enterprise. If selected skill is not available in skills pool, it is both assigned to the resource and added to the skills pool (line 13). It provides that skills required by any tasks to be performed would be selected from the pool of skills that are owned by at least one resource (line 17).

### C. Tasks

Having resources, and skills covered by them defined, tasks could be obtained. The number of tasks is set by the end–user. What it more user also sets the duration scope of the task (line: 16). Those are the bounds within the task duration would be randomly set (line: 16). For the sake of generation project instances for our research, we made an assumption that task duration would be the number between 8 and 40 hours. It reflects to the range between 1 and 5 days of any task's duration. The skill required by any task is selected from the pool of available skills in given project instance (line: 17).

### D. Precedence relations

One of the last steps during generation process is to define the precedence relations. End–user defines the number of those relations. S/he is also responsible for defining the general scope of relations. It means, the bigger relations scope set, the bigger distance between tasks is allowed in building the precedence relations diagram (line: 21). In other words, setting small relation scope could cause that resulted schedule would contain precedence relations between tasks that have been defined one by one or with slight distance (like task first and third). However, if the relation scope would be set to a bigger value, there could be relations in the final schedule between some tasks defined in the beginning and the end of the generation process (e.g. precedence relation between the first and the last task defined).

The bigger the distance between source and destination task is, the more complex the project instance critical path is. As a consequence, duration–based optimization would potentially be more difficult for such project instance.

### E. Assign resource

Finally, the initial schedule could be built by assigning resources to given tasks, preserving precedence and skill constraints (lines: 26–28). Produced schedule would always be feasible. The way how resources are assigned to tasks is set randomly. Hence, if there is more than one resource that can be assigned to given task, then generator could assign this task in different ways in different executions of generation process. Schedule is generated using the Serial Generation Scheme [9], what provides that generated schedule would be always feasible.

### F. Save project to file

The last step in the process of generation an instance is to save (line: 29) the resulted project. If user sets the output file type to *xml* (*mpp*), then generator produces the result in the format that could be easily loaded in Microsoft Project tool. If user selects *def* output format or does not select any, then the result would be saved in more compact format that could be read by any text editor. If *assign resources* option has been ticked, then tasks can have resources assigned. However it regards only generating output only in xml (mpp) format. The name of produced file relates to the name proposed by the end–user in given text field in the configuration screen.

## VI. Dataset summary

Due to evaluate not only the project schedule duration, but also the cost of the schedule including skills domain, we cannot use the standard PSPLIB benchmark dataset [8] in our research; that does not contain any information about the task performance cost. What is more, PSPLIB dataset instances do not reflect the MS–RCPSP. Hence, we prepared the dataset, containing 36 project instances, which have been artificially created, in a base of real–world instances, got from the Volvo IT Department in Wroclaw.

The dataset summary has been presented in the Table II. There are two groups of created project instances: one contains 100 tasks and the second – 200 tasks as typical ones performed in given international enterprise. Within each group, project instances are varied by number of available resources and the precedence relationship complexity. Number of resources for instances from both groups were chosen in a way to preserve constant average resource load and average task relations ratio for given instances. The skill variety has been set up to 9 or 15 different skill types for each project instance while any resource can dispose of exactly six different skill types. Because of the different resources and relations number, the scheduling complexity for each project is varied.

This dataset stands as an extension of dataset presented in [18], [19], [20], and that is the reason some instances are named with suffix *Dx*. This suffix refers to dataset instances that have been previously created and presented in those papers. Because of the extension the dataset, the need of introducing more clear namesystem has arisen. Suffix has been added to refer previously created files, keeping the naming convention applied after dataset extension.

### A. Project definition format (.def)

Because of changing the research's approach to be more generic, we decided to focus more on the dataset instances

TABLE II
COMPLEXITY INDICATORS AND DIFFICULTY MEASURE FOR IMOPSE DATASET INSTANCES. PROJECT INSTANCES REGARDED AS THE MOST DIFFICULT TO BE SCHEDULED ARE WRITTEN **BOLD**, WHILE THOSE ONES, WHO ARE INDICATED AS THE EASIEST TO SCHEDULE ARE WRITTEN *ITALIC*.

| Dataset instance | Features | | | | Indicators | | | | | | | | Difficulty |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $m$ | $n$ | $p$ | $\bar{q}$ | $\lambda$ | $\nu$ | $\Phi_d$ | $\Phi_c$ | $\mu$ | $\beta$ | $\pi$ | $\theta$ | |
| 100_10_26_15 | 100 | 10 | 26 | 15 | 0.100 | 0.053 | 0.316 | 0.123 | 1.000 | 0.117 | 0.200 | 0.278 | 0.243 |
| 100_10_27_9_D2 | 100 | 10 | 27 | 9 | 0.100 | 0.062 | 0.316 | 0.412 | 1.000 | 0.087 | 0.102 | 0.417 | 0.267 |
| 100_10_47_9 | 100 | 10 | 47 | 9 | 0.100 | 0.095 | 0.314 | 0.182 | 1.000 | 0.087 | 0.094 | 0.437 | 0.259 |
| 100_10_48_15 | 100 | 10 | 48 | 15 | 0.100 | 0.097 | 0.313 | 0.125 | 1.000 | 0.113 | 0.263 | 0.292 | 0.263 |
| **100_10_64_9** | 100 | 10 | 64 | 9 | 0.100 | 0.129 | 0.321 | 0.129 | 1.000 | 0.083 | 0.176 | 0.453 | **0.277** |
| 100_10_65_15 | 100 | 10 | 65 | 15 | 0.100 | 0.131 | 0.321 | 0.137 | 1.000 | 0.120 | 0.179 | 0.281 | 0.256 |
| 100_20_22_15 | 100 | 20 | 22 | 15 | 0.050 | 0.044 | 0.317 | 0.119 | 1.000 | 0.075 | 0.116 | 0.263 | 0.221 |
| 100_20_23_9_D1 | 100 | 20 | 23 | 9 | 0.050 | 0.052 | 0.317 | 0.356 | 1.000 | 0.045 | 0.135 | 0.451 | 0.265 |
| 100_20_46_15 | 100 | 20 | 46 | 15 | 0.050 | 0.093 | 0.321 | 0.125 | 1.000 | 0.072 | 0.102 | 0.264 | 0.229 |
| 100_20_47_9 | 100 | 20 | 47 | 9 | 0.050 | 0.095 | 0.314 | 0.122 | 1.000 | 0.043 | 0.081 | 0.397 | 0.243 |
| 100_20_65_15 | 100 | 20 | 65 | 15 | 0.050 | 0.131 | 0.314 | 0.117 | 1.000 | 0.073 | 0.091 | 0.248 | 0.232 |
| 100_20_65_9 | 100 | 20 | 65 | 9 | 0.050 | 0.131 | 0.318 | 0.114 | 1.000 | 0.045 | 0.068 | 0.426 | 0.251 |
| **100_5_20_9_D3** | 100 | 5 | 20 | 9 | 0.200 | 0.043 | 0.315 | 0.503 | 1.000 | 0.133 | 0.147 | 0.480 | **0.296** |
| **100_5_22_15** | 100 | 5 | 22 | 15 | 0.200 | 0.044 | 0.317 | 0.207 | 1.000 | 0.140 | 0.250 | 0.325 | **0.275** |
| **100_5_46_15** | 100 | 5 | 46 | 15 | 0.200 | 0.093 | 0.320 | 0.243 | 1.000 | 0.153 | 0.345 | 0.281 | **0.296** |
| **100_5_48_9** | 100 | 5 | 48 | 9 | 0.200 | 0.097 | 0.315 | 0.294 | 1.000 | 0.140 | 0.213 | 0.376 | **0.291** |
| **100_5_64_15** | 100 | 5 | 64 | 15 | 0.200 | 0.129 | 0.322 | 0.197 | 1.000 | 0.147 | 0.176 | 0.294 | **0.276** |
| **100_5_64_9** | 100 | 5 | 64 | 9 | 0.200 | 0.129 | 0.315 | 0.149 | 1.000 | 0.133 | 0.176 | 0.391 | **0.285** |
| 200_10_128_15 | 200 | 10 | 128 | 15 | 0.100 | 0.064 | 0.314 | 0.115 | 1.000 | 0.117 | 0.136 | 0.130 | 0.218 |
| 200_10_135_9_D6 | 200 | 10 | 135 | 9 | 0.100 | 0.084 | 0.318 | 0.428 | 1.000 | 0.087 | 0.139 | 0.200 | 0.254 |
| 200_10_50_15 | 200 | 10 | 50 | 15 | 0.100 | 0.025 | 0.317 | 0.111 | 1.000 | 0.110 | 0.130 | 0.147 | 0.212 |
| 200_10_50_9 | 200 | 10 | 50 | 9 | 0.100 | 0.025 | 0.318 | 0.130 | 1.000 | 0.087 | 0.127 | 0.212 | 0.222 |
| 200_10_84_9 | 200 | 10 | 84 | 9 | 0.100 | 0.042 | 0.313 | 0.124 | 1.000 | 0.083 | 0.200 | 0.231 | 0.238 |
| 200_10_85_15 | 200 | 10 | 85 | 15 | 0.100 | 0.043 | 0.315 | 0.121 | 1.000 | 0.107 | 0.176 | 0.143 | 0.223 |
| *200_20_145_15* | 200 | 20 | 145 | 15 | 0.050 | 0.073 | 0.313 | 0.124 | 1.000 | 0.072 | 0.096 | 0.133 | *0.209* |
| 200_20_150_9_D5 | 200 | 20 | 150 | 9 | 0.050 | 0.093 | 0.315 | 0.386 | 1.000 | 0.043 | 0.055 | 0.228 | 0.238 |
| *200_20_54_15* | 200 | 20 | 54 | 15 | 0.050 | 0.027 | 0.314 | 0.119 | 1.000 | 0.070 | 0.102 | 0.124 | *0.200* |
| 200_20_55_9 | 200 | 20 | 55 | 9 | 0.050 | 0.028 | 0.315 | 0.115 | 1.000 | 0.045 | 0.226 | 0.230 | 0.233 |
| *200_20_97_15* | 200 | 20 | 97 | 15 | 0.050 | 0.049 | 0.315 | 0.118 | 1.000 | 0.073 | 0.116 | 0.123 | *0.206* |
| 200_20_97_9 | 200 | 20 | 97 | 9 | 0.050 | 0.049 | 0.314 | 0.116 | 1.000 | 0.045 | 0.078 | 0.206 | 0.212 |
| 200_40_130_9_D4 | 200 | 40 | 130 | 9 | 0.025 | 0.088 | 0.316 | 0.342 | 1.000 | 0.023 | 0.165 | 0.183 | 0.243 |
| *200_40_133_15* | 200 | 40 | 133 | 15 | 0.025 | 0.067 | 0.314 | 0.131 | 1.000 | 0.038 | 0.067 | 0.118 | *0.201* |
| *200_40_45_15* | 200 | 40 | 45 | 15 | 0.025 | 0.023 | 0.314 | 0.115 | 1.000 | 0.038 | 0.046 | 0.135 | *0.190* |
| 200_40_45_9 | 200 | 40 | 45 | 9 | 0.025 | 0.023 | 0.316 | 0.113 | 1.000 | 0.023 | 0.122 | 0.216 | 0.212 |
| 200_40_90_9 | 200 | 40 | 90 | 9 | 0.025 | 0.045 | 0.314 | 0.117 | 1.000 | 0.023 | 0.116 | 0.221 | 0.216 |
| *200_40_91_15* | 200 | 40 | 91 | 15 | 0.025 | 0.046 | 0.317 | 0.112 | 1.000 | 0.037 | 0.075 | 0.131 | *0.198* |

stored in *.def* format that is easier to maintain and use by researchers. Hence we adopted instances created for MS Project to more generic form.

It led to remove the summary tasks that are specific for MS Project *.mpp* format. Summary tasks are used to group atomic tasks into more complex (i.e. task called 'development' could be split to some atomic tasks: database structures creation, development of business logic and development of user interface). However, MS Project allows to use summary tasks as predecessors for others. Therefore we multiplied precedence relations by copying them from predecessor's summary task to all of tasks included by this summary one. As a result new *Dx* instances have been created. Furthermore, some tasks, represented as summary ones, have been removed from the project. To be consistent with previous works, we keep names of those files the same. Those files are provided with additional description explaining the difference in number of tasks and precedence relations between file name and file content.

Adjusted dataset instances with *Dx* suffix have smaller number of tasks. Number of precedence relations is significantly bigger in all *Dx* instances. Roughly describing, it is more than twice precedence relations as in former instances, while number of tasks has been decreased in all instances in about 20% (about 20 tasks for instances with 100 tasks and 40 for instances with 200 tasks).

We have also presented in Tab. II the values of proposed complexity estimations. Finally, the overall complexity measure, as an aggregation value of complexity estimations components has been presented. Based on the overall complexity value, the most complex projects in scheduling point of view has been highlighted by bold. The overall complexity measure has been computed according to the Eq. 14. Changing weights of complexity estimations components would affect the final complexity value for each dataset instance. Hence, the complexity of each project could be different depending on priorities set by project manager.

In our approach all universality estimations values are equal to 1. It is because we made an assumption that every resource

TABLE III
COMPARISON OF RESULTS OBTAINED FOR GREEDY ALGORITHM, SIMPLE HEURISTICS, ACO [14] AND HANTCO [14] FOR VARIOUS OPTIMIZATION MODES FOR CALENDAR–CONSTRAINED DATASET INSTANCES (*.mpp*).

| Dataset instance | DO | | | | | | | | CO | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Heuristic | | Greedy | | ACO | | HAntCO | | Heuristic | | ACO | | HAntCO | |
| | Days | Cost | Days | Cost | Days | Cost | Days | Cost | Days | Cost | Days | Cost | Days | Cost |
| 100_10_26_15 | 37 | 126361 | 38 | 119336 | 32 | 124687 | **31** | 126216 | 85 | **70326** | 85 | **70326** | 85 | **70326** |
| 100_10_27_9_D2 | 38 | 44309 | 38 | 43438 | 34 | 44999 | **33** | 42199 | 129 | **26323** | 129 | **26323** | 129 | **26323** |
| 100_10_47_9 | 41 | 142759 | 40 | 135161 | 36 | 143100 | **34** | 140865 | 145 | **90992** | 145 | **90992** | 145 | **90992** |
| 100_10_48_15 | 36 | 135534 | 44 | 120664 | 33 | 133062 | **33** | 133495 | 85 | **87187** | 85 | **87187** | 85 | **87187** |
| 100_10_64_9 | 39 | 113124 | 43 | 117993 | 35 | 110643 | **33** | 113774 | 121 | **62102** | 121 | **62102** | 121 | **62102** |
| 100_10_65_15 | 40 | 152955 | 43 | 140782 | 35 | 150294 | **32** | 149185 | 98 | **106296** | 98 | **106296** | 98 | **106296** |
| 100_20_22_15 | 25 | 117493 | 24 | 112135 | 20 | 120949 | **19** | 123642 | 86 | **55240** | 87 | 55240 | 87 | 55240 |
| 100_20_23_9_D1 | 32 | 53154 | 32 | 50279 | 32 | 52119 | **23** | 53358 | 119 | 30104 | 121 | 30107 | 117 | **30104** |
| 100_20_46_15 | 28 | 138270 | 29 | 133739 | 25 | 138565 | **24** | 138568 | 75 | **68899** | 75 | **68899** | 75 | **68899** |
| 100_20_47_9 | 21 | 129160 | 28 | 140626 | 21 | 124817 | **18** | 134312 | 131 | **55197** | 131 | **55197** | 131 | **55197** |
| 100_20_65_15 | 32 | 110503 | 34 | 118569 | **27** | 109831 | 27 | 108991 | 69 | **57085** | 69 | **57085** | 69 | **57085** |
| 100_20_65_9 | 25 | 127149 | 24 | 124291 | 23 | 130934 | **21** | 126659 | 114 | **59736** | 114 | **59736** | 114 | **59736** |
| 100_5_20_9_D3 | 57 | 40539 | 55 | 40958 | **50** | 41029 | 53 | 40811 | 167 | **30164** | 167 | **30164** | 167 | **30164** |
| 100_5_22_15 | 63 | 119266 | 77 | 128354 | **60** | 119434 | 60 | 119158 | 86 | **109111** | 86 | **109111** | 86 | **109111** |
| 100_5_46_15 | 75 | 202238 | 80 | 202607 | **67** | 204110 | 67 | 204730 | 125 | **184409** | 125 | **184409** | 125 | **184409** |
| 100_5_48_9 | 72 | 193383 | 78 | 196893 | **62** | 191712 | 62 | 191888 | 130 | **175225** | 130 | **175225** | 130 | **175225** |
| 100_5_64_15 | 71 | 141407 | 66 | 141882 | 62 | 144972 | **61** | 143956 | 141 | **109091** | 141 | **109091** | 141 | **109091** |
| 100_5_64_9 | 71 | 102439 | 67 | 107014 | 61 | 102777 | **61** | 101297 | 173 | **72848** | 173 | **72848** | 173 | **72848** |
| 200_10_128_15 | 71 | 180812 | 78 | 198378 | 62 | 178264 | **60** | 178375 | 159 | **134425** | 143 | 136551 | 143 | 136551 |
| 200_10_135_9_D6 | 216 | 105593 | 216 | 93426 | 216 | 99375 | **186** | 103561 | 256 | **71986** | 274 | 72036 | 270 | 71986 |
| 200_10_50_15 | 66 | 189660 | 75 | 183673 | 63 | 191856 | **62** | 190956 | 167 | **84308** | 167 | **84308** | 167 | **84308** |
| 200_10_50_9 | 66 | 251158 | 70 | 250732 | 65 | 250075 | **64** | 250850 | 318 | **105198** | 318 | 105232 | 318 | **105198** |
| 200_10_84_9 | 70 | 224121 | 66 | 222976 | 69 | 226666 | **66** | 222655 | 338 | **117543** | 316 | 117754 | 318 | **117543** |
| 200_10_85_15 | 65 | 304277 | 68 | 301357 | **61** | 306949 | 62 | 302064 | 215 | **195820** | 215 | 195820 | 215 | **195820** |
| 200_20_145_15 | 36 | 275983 | 46 | 277097 | 36 | 278199 | **35** | 272504 | 158 | **143497** | 152 | 143688 | 158 | **143497** |
| 200_20_150_9_D5 | 183 | 92821 | 183 | 95667 | 186 | 91461 | **177** | 92567 | 337 | **51496** | 296 | 51678 | 345 | 51496 |
| 200_20_54_15 | 37 | 295786 | 41 | 290656 | 39 | 299993 | **34** | 298822 | 125 | **161412** | 131 | 161614 | 125 | **161412** |
| 200_20_55_9 | 37 | 230150 | 37 | 232766 | 38 | 231094 | **36** | 223879 | 332 | **70057** | 250 | 72176 | 332 | **70057** |
| 200_20_97_15 | 49 | 290399 | 69 | 346527 | 42 | 280951 | **42** | 277860 | 171 | **156951** | 169 | 157202 | 171 | **156951** |
| 200_20_97_9 | **35** | 273378 | 43 | 282379 | 37 | 275819 | 35 | 278797 | 169 | **98480** | 150 | 99901 | 169 | **98480** |
| 200_40_130_9_D4 | 112 | 101879 | 112 | 90907 | 112 | 94488 | **108** | 104965 | 214 | **46133** | 205 | 48419 | 216 | 46275 |
| 200_40_133_15 | 24 | 276456 | **23** | 279170 | 27 | 281933 | 24 | 279073 | 155 | 97345 | 131 | 99329 | 144 | **97345** |
| 200_40_45_15 | 31 | 260738 | 32 | 269623 | 25 | 248717 | **23** | 256687 | 213 | **87955** | 161 | 91010 | 213 | **87955** |
| 200_40_45_9 | **22** | 270758 | 23 | 276416 | 26 | 273632 | 25 | 270428 | 334 | **77236** | 179 | 94142 | 315 | 82192 |
| 200_40_90_9 | 24 | 290028 | **20** | 294909 | 26 | 287694 | 24 | 298340 | 285 | **80732** | 142 | 96312 | 247 | 84038 |
| 200_40_91_15 | **19** | 249909 | 35 | 250843 | 25 | 257927 | 23 | 241492 | 184 | **86476** | 132 | 88616 | 184 | **86476** |

has the same number of different skills and this is also the maximal number of potential skills covered by resource, used in normalization. If we decided to make the number of skills covered by resource various, depending to given resource, then the universality estimation values would not be always equal to 1.0.

### B. iMOPSE Solution Validator

We released also an additional tool to validate generated solutions in case of preserving all constraints defined in MS–RCPSP. Such validator is available on the iMOPSE project website[3]. Validator checks whether all tasks have any resource assigned (assignments validation), final schedule is conflict–free (conflicts validation), any task having predecessors is set to be started after all its predecessors would be finished (precedence relations validation) and whether any task has resource assigned that is capable of performing it (skill validation).

Validator shows not only the validation results but also the quality the validated solution – its duration measured in hours and cost measured in some currency units. If some validation rules are broken, they are shown to end–user.

Validator is compatible with .def project definition format. For further information how to use the validator, please refer to documents related with the tool – User's Manual or Case study – available on iMOPSE Platform.

## VII. EXPERIMENTS AND RESULTS

The main goal of conducted experiments was to link and compare both (*.mpp* [14] and *.def* based) approaches, considering the impact of calendar restrictions.

We decided to use simple duration– and cost– oriented heuristic [20], greedy algorithm and compare them with ACO and HAntCO approaches described in [14]. Furthermore, greedy algorithm and simple heuristics have been used to schedule .def dataset instances.

However, proposed heuristic and greedy approaches for cost optimization turned out to become the same method. Therefore, presented results are divided into main two parts regarding optimization modes: duration optimization ($\omega_\tau = 1$) and cost optimization ($\omega_\tau = 0$). Each of those main part is also divided for three parts in cost optimization (heuristic, ACO, HAntCO) and four parts in duration optimization (heuristic, greedy, ACO, HAntCO).

Table III presents the obtained results for both optimization modes using both proposed methods (simple heuristic and greedy algorithm). It also contains results obtained by ACO and HAntCO approaches described in detail in [14]. This

[3] http://imopse.ii.pwr.edu.pl

table presents optimization results for dataset instances with calendar restrictions (*.mpp*).

Greedy algorithm is a method that works iteratively. In every step of greedy scheduling, one task is added to the schedule. The decision, which task to which resource should be assigned in given algorithm step is made based on the current partial schedule. In other words: in a given step, currently best task–to–resource assignment option is chosen and the next step is performed until all tasks would be scheduled. Classical greedy algorithm assumes possibility to analyse not only current state of the partial schedule, but also to investigate several further steps. In that approach combinations of several tasks are analysed and the best one, containing given number of tasks–to–resource assignments is selected and added to the partial schedule. However, in our approach we discuss only current schedule state, omitting the analysis of several assignments. Therefore, number of steps of proposed greedy algorithms would be always equal to the number of tasks in a project.

For duration oriented optimization mode, greedy algorithm analyses which task should be added to make the partial schedule the shortest. For cost-oriented optimization mode, the criteria of selecting tasks bases on the cost of the assignment given task to given resource. For every task, various resources are analysed to be assigned, and the cheaper one is chosen.

In cost-oriented optimization mode, both greedy algorithm and simple heuristic (Resource Salary based [20]) works according to the same schema, described above. However, for the duration–oriented optimization, heuristic and the greedy algorithm differs in details. In the greedy algorithm task is assigned to given resource and then added to the partial schedule then conflicts are fixed and finally the project duration is computed. In simple heuristic (Successors List Size based [20]) firstly the resource that would be the earliest free (not assigned to any task) is selected. Then task is assigned to this found task while its start time is set just after then end of the last of tasks previously assigned to given resource. It allows to build feasible schedule without the necessity of fixing conflicts as the method is the resource conflict–free.

Taking into account results gathered in the Tab. III we can conclude that for duration optimization method, the HAntCO outclassed other methods, provided the best results in 28 of 36 cases (78%). ACO turned out to be the best method for 6 cases (16%), while greedy gave best results in 3 of 36 cases (8%) and heuristic was the most suitable in 2 of 36 cases (6%).

For cost optimization method, simple heuristic gives the best result for almost all of dataset instances (34/36, 94%). However, for remaining two instances heuristic also provided solution with the smallest cost, but the schedule duration was bigger than for other method (HAntCO). For most of the instances (32/36, 89%) HAnt-CO provided the same, best result than obtained from heuristic. ACO–approach provided the same, best results in 18/36 (50%) cases. The most interesting fact for cost optimization is that ACO provided best results mostly for dataset instances containing 100 tasks - 17/18 cases (94%) and only once for dataset instances containing 200 tasks (6%).

In the Tab. IV we compiled the summary of obtained best results for classical optimization methods - heuristics and greedy algorithm for instances not regarding calendar restrictions (*.def*). It can be also found in the iMOPSE website. As we are oriented to use *.def* format in further research, obtained project schedules are measured by hours rather than days as it has been so far, in *.mpp* format. Obtained results stand as a benchmark for further research when using .def format. On the other hand, Tab. III is still regarded as a benchmark for methods working on *.mpp* format.

TABLE IV
SUMMARY OF BEST OBTAINED RESULTS FOR DATASET INSTANCES NOT REGARDING CALENDAR CONSTRAINTS (*.def*).

| Dataset instance | Heuristic CO | | Heuristic DO | | Greedy CO | |
|---|---|---|---|---|---|---|
| | Hours | Cost | Hours | Cost | Hours | Cost |
| 100_10_26_15 | 728 | 71616 | 316 | 125073 | 370 | 130315 |
| 100_10_27_9_D2 | 1184 | 26771 | 334 | 44319 | 646 | 42984 |
| 100_10_47_9 | 1224 | 92771 | 310 | 144840 | 549 | 162642 |
| 100_10_48_15 | 766 | 88794 | 325 | 138845 | 344 | 139761 |
| 100_10_64_9 | 1028 | 63279 | 324 | 117759 | 533 | 124897 |
| 100_10_65_15 | 831 | 108239 | 285 | 152669 | 426 | 173754 |
| 100_20_22_15 | 756 | 56151 | 162 | 121561 | 353 | 98621 |
| 100_20_23_9_D1 | 1219 | 30643 | 247 | 52436 | 617 | 63210 |
| 100_20_46_15 | 639 | 70061 | 231 | 142962 | 394 | 140994 |
| 100_20_47_9 | 1114 | 56190 | 179 | 130612 | 390 | 119462 |
| 100_20_65_15 | 582 | 58134 | 298 | 111130 | 310 | 125081 |
| 100_20_65_9 | 964 | 60954 | 174 | 127260 | 408 | 147952 |
| 100_5_20_9_D3 | 1408 | 30728 | 523 | 40976 | 625 | 38725 |
| 100_5_22_15 | 723 | 111189 | 537 | 120039 | 630 | 121369 |
| 100_5_46_15 | 1054 | 187623 | 658 | 207810 | 693 | 212261 |
| 100_5_48_9 | 1092 | 178346 | 580 | 196221 | 779 | 191888 |
| 100_5_64_15 | 1195 | 111388 | 574 | 146661 | 640 | 149635 |
| 100_5_64_9 | 1506 | 74199 | 567 | 109518 | 597 | 101062 |
| 200_10_128_15 | 1217 | 139149 | 537 | 179335 | 780 | 213091 |
| 200_10_135_9_D6 | 2581 | 73207 | 1079 | 105604 | 1426 | 105196 |
| 200_10_50_15 | 1414 | 86008 | 549 | 190555 | 763 | 190981 |
| 200_10_50_9 | 2681 | 106986 | 536 | 251903 | 817 | 239238 |
| 200_10_84_9 | 2702 | 119500 | 589 | 231457 | 999 | 232937 |
| 200_10_85_15 | 1813 | 199585 | 545 | 314599 | 706 | 346573 |
| 200_20_145_15 | 1331 | 146303 | 293 | 280623 | 480 | 280774 |
| 200_20_150_9_D5 | 3024 | 52542 | 1232 | 94355 | 1930 | 116179 |
| 200_20_54_15 | 1054 | 164142 | 306 | 299677 | 488 | 322627 |
| 200_20_55_9 | 2809 | 71262 | 280 | 233960 | 999 | 276513 |
| 200_20_97_15 | 1491 | 159680 | 347 | 294938 | 680 | 324041 |
| 200_20_97_9 | 1515 | 100421 | 304 | 279894 | 816 | 301723 |
| 200_40_130_9_D4 | 2038 | 47050 | 586 | 104261 | 1710 | 121485 |
| 200_40_133_15 | 1282 | 99266 | 183 | 285299 | 512 | 270201 |
| 200_40_45_15 | 1807 | 89642 | 267 | 266970 | 616 | 269754 |
| 200_40_45_9 | 2781 | 79979 | 198 | 273818 | 821 | 218708 |
| 200_40_90_9 | 2405 | 82177 | 173 | 292873 | 963 | 300258 |
| 200_40_91_15 | 1560 | 88233 | 179 | 250005 | 519 | 278582 |

Results obtained in the Tab. IV show that SLS [20] heuristic provides better results in DO mode in all of 36 dataset instances. It clearly shows that SLS heuristic is definitely better optimization method than greedy algorithm in this problem.

However the project definitions are compatible to each other between *.def* and *.mpp* formats, there are some small differences in cost result in CO, using the same method. It is because of the adjustment made when transferring *.mpp* to *.def* format. For the sake of simplicity, task's duration in .def has been rounded up to the integer values. It lead to the differences, because cost of performing project is a sum of each task's performance cost. While task's performance cost is computed as a multiplication of task duration and assigned resource's salary. As a result of rounding up, cost of each task has increased slightly, even though the same resource is assigned to it. Therefore the overall cost is slightly bigger for solutions obtained for *.def* files.

## VIII. Conclusions and further work

In this paper some novel difficulty indicators for instances of Multi–Skill Resource–Constrained Project Scheduling Problem have been presented. Furthermore the extended dataset has been presented and suggested as a benchmark for this problem, as no other benchmark dataset can be found that satisfies proposed constraints. Furthermore, those instances have been scheduled using greedy algorithm, to provide an initial platform for comparing results obtained by various researchers.

Proposed complexity estimations stand some first step in project scheduling data analysis. Guessing the project complexity could be helpful in parameters' tuning for various optimization methods. As more complex / difficult to schedule project is, the optimization process would potentially last longer for the same parameter configuration than for other instances. Hence, the decision maker could decide to change the parameters, e.g. by decrease number of method iterations. We managed to make those observations sure in our EA–based approach, where building schedule for the project with suffix D2 generally lasts longer than for the project with suffix D1.

The goal of presenting the dataset instance generator is to allow and encourage other researchers to focus on the problem and possible solutions and methods we propose. We still believe there is a lot to investigate and research. What is more, the dataset instance format we propose is very common in many industries, as the MS Project is a common standard.

We are also on the point of investigating approaches concentrated to different multi–objectiveness handling methods. Most of them we analyse are based on Pareto–front (like NSGA-II [22], [17] or other methods). One of the goals is to find a way how to provide a set of non–dominated results to the project manager, to delegate the matter of making decision which of those proposed solutions is the best, according to the specificity of the company it regards. E.g. in some industries the aim is to finish the project as soon as possible while in some others the most important is to perform it in the cheapest way. Still we would like to give the choice from a pool of some solutions.

## References

[1] Al–Anzi F.S., Al–Zamel K., Allahverdi A.; Weighted Multi–Skill Resources Project Scheduling, J. of Software Engineering & Applications (3), pp. 1125–1130, 2010.

[2] Blazewicz J., Lenstra J.K., Rinnooy Kan A.H.G.; Scheduling subject to resource constraints: Classification and complexity, Discrete Applied Mathematics (5), pp. 11-24, 1983.

[3] Drezet L.E., Billaut J.C.; A project scheduling problem with labour constraints and time–dependent activities requirements, Int. J. of Production Economics (112), pp. 217-225, 2008.

[4] Gonzalez F., Ramies Rios D., Multi–objective Optimization of the Resource Constrained Project Scheduling Problem (RCPSP) A heuristic approach based on the mathematical model, The Int. J. of Computer Science & Applications (TIJCSA) (2/2), pp. 1–13, 2013.

[5] Hegazy T., Shabeeb A.K., Elbeltagi E., Cheema T.; Algorithm for scheduling with multiskilled constrained resources, J. of Construction Engineering and Management (11-12/2000), pp. 414–421, 2000.

[6] Jaberi M., Jaberi M.; A Multi–objective Resource–Constrained Project–Scheduling Problem Using Mean Field Annealing Neural Networks, J. of Mathematics and Computer science (9), pp. 228–239, 2014.

[7] Kadrou Y., Najid N.M.; A new heuristic to solve RCPSP with multiple execution modes and Multi-Skilled Labor , IMACS Multiconference on Computational Engineering in Systems Applications (CESA), pp. 1302–1309, 2006,

[8] Kolisch R., Sprecher A., PSPLIB - A project scheduling problem library, European Journal of Operational Research (96), pp. 205–216, 1996.

[9] Kolisch R., Hartmann S., Experimental evaluation of state-of-the-art heuristics for the resource-constrained project scheduling problem, European Journal of Operational Research (127), pp. 394–407, 2000.

[10] Kolisch R., Hartmann S., Experimental investigation of heuristics for resource-constrained project scheduling: An update, European Journal of Operational Research (174), pp. 23-37, 2006.

[11] Latva-Koivisto A., M., Finding a complexity measure for business process models, Research Report, Mat-2.108, Individual Research Projects in Applied Mathematics, 2001.

[12] Li H., Womer K.; Scheduling projects with multi-skilled personnel by a hybrid MILP/CP benders decomposition algorithm, J. of Scheduling, (12), pp. 281-298, 2009.

[13] Luna F., Gonzalez-Alvarez, D. L., Chicano F., Vega-Rodriguez M. A.; The software project scheduling problem: A scalability analysis of multi-objective metaheuristics, Applied Soft Computing Vol. 15, pp.136–148, 2013.

[14] Myszkowski P. B., Skowroński M.E., Olech Ł., Oślizło K.; Hybrid Ant Colony Optimization in solving Multi-Skill Resource-Constrained Project Scheduling Problem, Soft Computing, DOI 10.1007/s00500-014-1455-x, 2014.

[15] Phruksaphanrat B.; Multi–Objective Multi–Mode Resource–Constrained Project Scheduling Problem by Preemptive Fuzzy Goal Programming, World Academy of Science, Engineering and Technology, Int. J. of Mechanical, Industrial Science and Engineering (8/3), pp. 99–103, 2014

[16] Santos M., Tereso A. P.; On the multi-mode, multi-skill resource constrained project scheduling problem - computational results, Soft Computing in Industrial Applications, Advances in Intelligent and Soft Computing (96), pp. 239–248, 2011.

[17] Sarker B.R, Yu J., Mungan D., Rahman M.A.A., Parveen S.; Pareto–optimal solution of a scheduling problem on a single machine with periodic maintenance and non pre–emptive jobs, Proceedings of the International Conference on Mechanical Engineering, pp. 1–5, 2007.

[18] Skowroński M. E., Myszkowski P. B., Specialized genetic operators for Multi-Skill Resource-Constrained Project Scheduling Problem, 19th Inter. Conference on Soft Computing Mendel 2013, pp. 57-62, 2013.

[19] Skowroński M. E., Myszkowski P. B., Kwiatek P., Adamski M., Tabu Search approach for Multi-Skill Resource-Constrained Project Scheduling Problem, Annals of Computer Science and Information Systems Volume 1, Proc. of the 2013 Federated Conference on Computer Science and Information Systems, pp. 153-158, 2013.

[20] Skowroński M. E., Myszkowski P. B., Podlodowski Ł., Novel heuristic solutions for Multi-Skill Resource-Constrained Project Scheduling Problem, Annals of Computer Science and Information Systems Volume 1, Proc. of the 2013 Federated Conference on Computer Science and Information Systems, pp. 159-166, 2013.

[21] Van Peteghem, Vanhoucke M., An experimental investigation of metaheuristics for the multi–mode resource–constrained project scheduling problem on new dataset instances, European Journal of Operational Research (235), pp.62-72, 2014.

[22] Vanucci C.S, Bicalho R., Carrano E.G., Takahashi R.H.C.; A Modified NSGA-II for the Multiobjective Multi–mode Resource-Constrained Project Scheduling Problem, WCCI 2012 IEEE World Congress on Computational Intelligence June, pp. 10–15, 2012.

[23] Wang Y., Chen D., Liu S., Zeng Q.; An Instance Generator for Project Scheduling Problems with Multi-Skilled Personnel Constraints, 2012 24th Chinese Control and Decision Conference (CCDC), pp. 3430–3435, 2012.

[24] Yannibelli V., Amandi A.; Hybridizing a multi–objective simulated annealing algorithm with a multi–objective evolutionary algorithm to solve a multi–objective project scheduling problem, Expert Systems with Applications (40), pp. 2421-2434, 2013.

DOI: 10.15439/2015F407
ACSIS, Vol. 5

# A two-level classifier for automatic medical objects classification

Przemyslaw Wiktor Pardel*, Jan G. Bazan*, Jacek Zarychta† and Stanislawa Bazan-Socha‡
*Interdisciplinary Centre for Computational Modelling, University of Rzeszów, Pigonia 1 Str., 35 - 310 Rzeszów, Poland
Email: ppardel@ur.edu.pl, bazan@ur.edu.pl
†Radiology consultant, Department of Pulmonology, Pulmonary Hospital,
Gladkie 1 Str., 34-500, Zakopane, Poland
Email: jzar@mp.pl
‡II Department of Internal Medicine, Jagiellonian University Medical College, Skawinska 8 Str., 31-066 Kraków, Poland
Email: mmsocha@cyf-kr.edu.pl

*Abstract*—**The goal of this paper is to describe the approach for automatic identifying human organs from a medical CT images and discuss results of its comparison to different classification methods. The main premise of this approach is the use of data sets together with the relevant domain knowledge. We test our approach on multiple CT images of chest organs (trachea, lungs, bronchus) and demonstrate usefulness and effectiveness of the resulting classifications. The presented approach can be used to assist in solving more complex medical problems. Keywords: CT images, concept approximation, classifiers, decision trees, medical object recognition, object classification, domain knowledge, organs identifying, medical system**

## I. Introduction

A DESIGN of human–machine interface is the most important aspect of computer aided interpretation of medical image exams. Assists include decision support, reminder and navigation techniques to help avoid diagnosis errors, content-based data mining capabilities, and access to reference libraries. Human–machine systems should take advantage of computer capabilities to increase physicians interpretation capabilities [1].

An automatic identification of medical objects visualized by Computed Tomography (CT) imagery (e.g., organs, blood vessels, bones, etc.), without any doubt, could be useful, to support solving many complex medical problems using computer tools. Our approach is based on a two-level classifier. On the lower level, our approach uses a classical classifier based on a decision tree that is calculated on the basis of the local discretization (see, *e.g.*, [2], [3]). This classifier is constructed and based on the features extracted from images using methods known from literature (see [4], [5] for more details). At a higher level of our two-level classifier, a collection of advisers works that is able to verify actions performed earlier by the lower-level classifier. This is possible by using domain knowledge injected to advisers. Each of the adviser is constructed as a simple algorithm based on a logical formula, that on input receives selected information extracted from a tested image and a decision returned by the lower-level classifier, and the output returns confirmation or negation for the suggestion generated by the lower-level classifier. It consists in the fact, that in a situation where the decision taken by the lower-level classifier, is clearly incompatible with domain knowledge, the adviser suggests to refrain from taking a decision. Thanks to this, increases the accuracy of such the two-level classifier, with a slight decrease in its coverage. To illustrate the method and to verify the effectiveness of presented classifiers, we have performed several experiments with the data sets obtained from Second Department of Internal Medicine, Collegium Medicum, Jagiellonian University, Krakow, Poland.

In the Section II, we describe the problem of medical image understanding. Second section present conception of design a system for automatic medical objects classification. Finally, we present the complete structure of two-level classifier, results of the comparison to other classification methods performed on medical data sets for the automatic classification of chest organs (see Section IV).

## II. Medical image understanding

A process of radiological interpretation generally includes the understanding of medical image content resulting in recognition of possible pathology symptoms, most often called detection, and assessment of comprehensive image information in a context of current clinical case-knowledge. It involves image-based detection of disease, defining disease extent, determining etiology of the disease process, assisting in designing of the clinical management plans for the patient, based on imaging findings, and following response to the therapy [6].

The other area of application of the automatic image understanding technique is deep and requires a detailed analysis of particularly difficult images, especially in case of doubts and difficulties in deciding on final diagnosis. A very important difference between all traditional methods of automatic image processing (or recognition) and the new paradigm for image understanding is that there is one directional scheme of the data flow in the traditional methods; there are two-directional interactions between signals (features) extracted from the image analysis and expectations resulting from the knowledge

This work was partially supported by the Polish National Science Centre grant DEC-2013/09/B/ST6/01568 and by the Centre for Innovation and Transfer of Natural Sciences and Engineering Knowledge of University of Rzeszów, Poland.

978-83-60810-66-8/$25.00©2015, IEEE

of image contents, as given by experts (physicians). The results of all analyses of medical image characteristics and objects visible in them, generated by computers, allow the physician to base his/her reasoning on much more reliable and quantifiable premises than just a visual assessment of that image, improving both the effectiveness of his/her activities, and the feeling of reliability and security. Finally, the increasing acceptance of techniques for the automatic recognition and classification of biological objects distinguished in medical images can help the doctor make the right diagnostic decisions, although these techniques sometimes require the doctor to be able to critically assess the automatically suggested categories, as every recognition technique carries some level of error, while nothing excuses the doctor's personal responsibility for his/her decisions [7].

Medical image analysis is one of the areas of computer vision where domain knowledge plays a very important role, because localized pixel information obtained from CT images is often ambiguous and unreliable [8]. The history of knowledge-based medical image analysis is older than the history of practical usage of CT imaging. One of the early studies in knowledge based medical image analysis was done by Harlow and Eisenbeisc [9] on radiographic image segmentation, when CT imaging was not yet available in hospitals. They proposed a top-down control system using a trees structured model description containing knowledge about locations and spatial relations of parts/organs of the human body. In his thesis work, Selfridge [10] discussed image understanding systems in general and divided the causes of difficulties into problems of model selection, segmentation techniques, and parameter setting [8].

We conclude that the automatic detection of organs is the first step to understand medical images and it is necessary to begin the process of proper medical diagnosis support. To understand the CT image correctly, a computer should detect and recognize all medical objects located on the image by using domain knowledge. The knowledge about objects located in the medical image, allows the correct identification of areas related to various medical problems. To understand medical image correctly, a computer should detect and recognize quite correctly all medical objects located on the image by using domain knowledge (extremely challenging task even for a man).

## III. Conception of design a system for automatic medical objects classification

### A. A general description

In order to understand the medical images, it is important to create a tool for understanding the interior of the human body on different levels of abstraction and tracking of interaction between the observed medical objects. The main issues to be addressed include problems with the quality of the medical image data, problems with domain knowledge descriptions and problems with modeling and exploration of the human body, which is very complex. The system should include the assumptions, such that the system should support work of doctors (not replace), expert always decide, system should allow for future sharing of knowledge and should naturally communicate in order to exchange knowledge (speech).

### B. "Low-Level" features (LLF)

There is no "*ideal set of features*" which characterize the object. Features are selected individually depending on the recognized objects. In the computer analysis of the images, extracted features from the image, can be assigned to one of the categories, such as non-transformed structural characteristics (*e.g.*moments, power, amplitude information, energy, etc.),transformed structural characteristics (*e.g.*frequency and amplitude spectra, subspace transformation methods, etc.), structural descriptions (formal languages and their grammars, parsing techniques, and string matching techniques) and graph descriptors (*e.g.*attributed graphs, relational graphs, and semantic networks) described in detail in [4] and [5]. In this publication we call these features as Low-Level Features (LLF). In total, for the purposes of the experiments we define 18 LLF features (see Table I).

TABLE I
"Low-Level" Features

| **Name** | **Description** |
|---|---|
| **DT** | Distance to the first image in the series (mm) |
| **SIZE** | Object size |
| **WIDTH** | Object width |
| **HEIGHT** | Object heihgt |
| **DFL** | The distance from the object to the left edge of the image |
| **DFT** | The distance from the object to the top edge of the image |
| **R1** | Size of the object located in the region R1 |
| **R2** | Size of the object located in the region R2 |
| **R3** | Size of the object located in the region R3 |
| **R4** | Size of the object located in the region R4 |
| **R5** | Size of the object located in the region R5 |
| **R6** | Size of the object located in the region R6 |
| **R7** | Size of the object located in the region R7 |
| **R8** | Size of the object located in the region R8 |
| **R9** | Size of the object located in the region R9 |
| **CIRCUIT** | Object circuit |
| **TFACTOR** | Object thickness factor |
| **SFACTOR** | Object shape factor |

### C. "Domain Knowledge" Features (DKF)

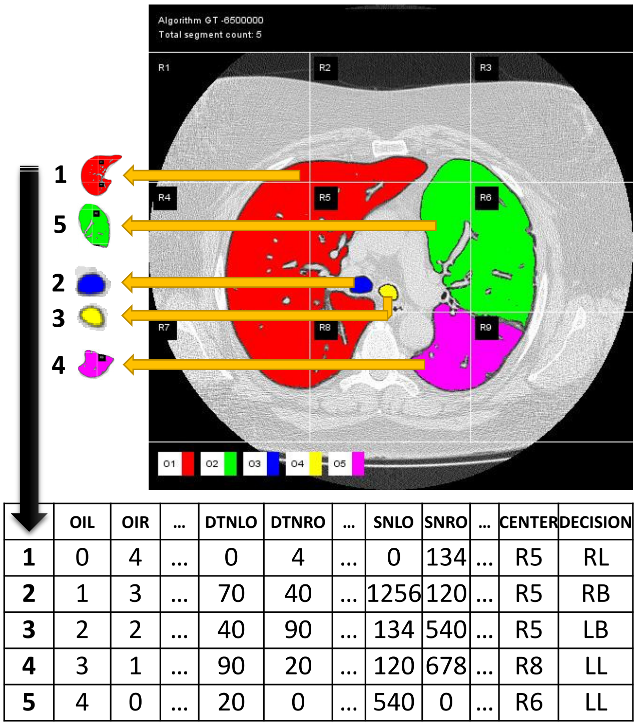
To understand the image, it is also necessary to define the additional features that will define the acquired domain knowledge from experts. We call these features Domain Knowledge Features (DKF). DKF can be assigned to one of the categories, such as:

- features used to describe domain knowledge about the number of objects that surround an analyzed object,
- features used to describe domain knowledge about the distance from analyzed object to surrounding objects,
- features used to describe domain knowledge about the size of objects that surround an analyzed object,
- features used to describe domain knowledge about position of an object.

In total, for the purposes of the experiments we define 13 DKF features (see Table II).

TABLE II
"DOMAIN KNOWLEDGE" FEATURES

| Name | Description |
|---|---|
| **CENTER** | Center of the object region (*e.g.*R1, R2 ...) |
| **OIL** | The number of objects on the right side |
| **OIR** | The number of objects on the left side |
| **OIA** | The number of objects above |
| **OIB** | The number of objects below |
| **DTNLO** | Distance to the nearest object on the left side |
| **DTNRO** | Distance to the nearest object on the right side |
| **DTNAO** | Distance to the nearest object above |
| **DTNBO** | Distance to the nearest object below |
| **SNLO** | The size of the nearest object on the left side |
| **SNRO** | The size of the nearest object on the right side |
| **SNAO** | The size of the nearest object above |
| **SNBO** | The size of the nearest object below |



| | OIL | OIR | ... | DTNLO | DTNRO | ... | SNLO | SNRO | ... | CENTER | DECISION |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0 | 4 | ... | 0 | 4 | ... | 0 | 134 | ... | R5 | RL |
| **2** | 1 | 3 | ... | 70 | 40 | ... | 1256 | 120 | ... | R5 | RB |
| **3** | 2 | 2 | ... | 40 | 90 | ... | 134 | 540 | ... | R5 | LB |
| **4** | 3 | 1 | ... | 90 | 20 | ... | 120 | 678 | ... | R8 | LL |
| **5** | 4 | 0 | ... | 20 | 0 | ... | 540 | 0 | ... | R6 | LL |

Fig. 1. Examples of "domain knowledge" features extraction

## D. Medical data

Our experiments were carried out on the data obtained from the clinical hospital Jagiellonian University Medical College in Kraków (the patients were diagnosed with asthma). The entire data set counted 26 patients (19 woman, 7 man). The average age of patients was 58.12 years (st.dev. 6.78 years, age range from 47 to 70 years). In all patients, volumetric CT torso scans were performed at both full inspiration and expiration with using 16-channel multi-detector CT scanner Toshiba (manufacturer's model name: Aquilion). The acquired data were reconstructed using a kernel (FC86) with 1 mm increments. Images were stored in the Digital Imaging and Communications in Medicine (DICOM) format. For each patient was taken 300 to 400 images (full inspiration) with a resolution of 512x512 pixels. The total size of the data set for the experiment count 9655 CT images.

From all images we select every fifth image (20% of all images, 5mm increments) to pre-processing. As a result of the segmentation process, we acquired 7491 objects for experiments. For all the objects we set LLF and DKF features, further all objects are classified by an expert to one of the 7 classes (chest organs) presented in the Table III.

TABLE III
OBJECT CLASSES

| Class | Object | Number |
|---|---|---|
| **TR** | Trachea | 671 (8,96%) |
| **RL** | Right lunge | 1621 (21,64%) |
| **LL** | Left lunge | 1616 (21,57%) |
| **RB** | Right main bronchi | 190 (2,54%) |
| **LB** | Left main bronchi | 211 (2,82%) |
| **LL+RL** | Object by gluing the left and right lungs | 55 (0,73%) |
| **OT** | Other objects | 3127 (41,74%) |

The entire data set was divided 20 times randomly into two sets - a set with training data and a set with test data (around 70% of the data getting into a training set - 18 patients, other (around 30%) into test set - 8 patients). All the experiments we conducted on these datasets.

# IV. EXPERIMENT RESULTS

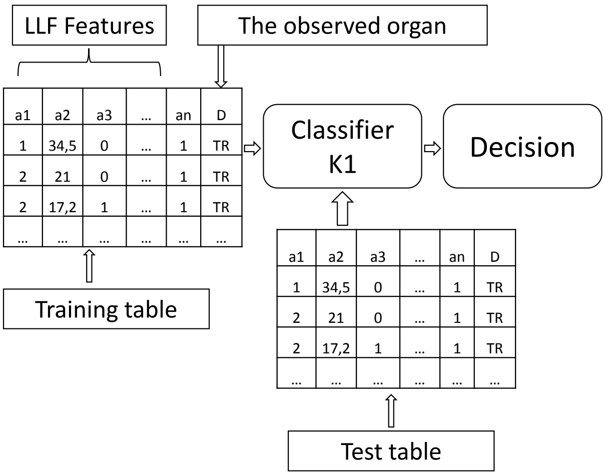
To verify the effectiveness of classification we prepare four methods. With using training data we built a classifier, which has been tested on test data. We designed a classifier to the automatic classification of chest organs. In method four we have implemented the two-level classifier in the IMPLA (Image Processing Laboratory), which is a continuation of the RSES-lib library (forming the kernel of the RSES system [11]), in the field of image processing. The IMPLA has developed recently in Interdisciplinary Centre for Computational Modelling, University of Rzeszów, Poland.

## A. Method 1

Method 1 is a method based on the decision tree with local discretization (LLF features, the quality of a given cut is computed as a number of objects pairs discerned by this cut and belonging to different decision classes, see, *e.g.*, [2], [3]). The method gave good results. Diagram of this method presented on Figure 2.

## B. Method 2

The second method was similar to the method 1 and based on the decision tree with local discretization. This method has used both LLF and DKF features.



| a1 | a2 | a3 | … | an | D |
|---|---|---|---|---|---|
| 1 | 34,5 | 0 | … | 1 | TR |
| 2 | 21 | 0 | … | 1 | TR |
| 2 | 17,2 | 1 | … | 1 | TR |
| … | … | … | … | … | … |

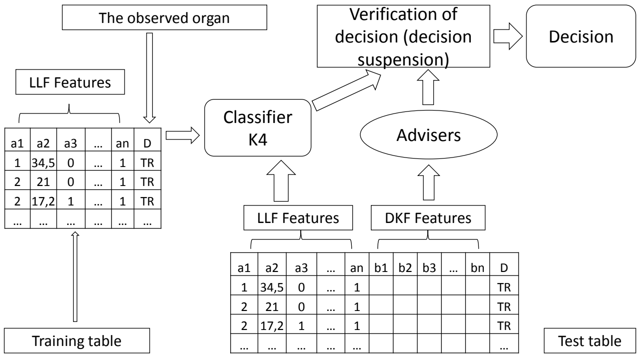
| a1 | a2 | a3 | … | an | D |
|---|---|---|---|---|---|
| 1 | 34,5 | 0 | … | 1 | TR |
| 2 | 21 | 0 | … | 1 | TR |
| 2 | 17,2 | 1 | … | 1 | TR |
| … | … | … | … | … | … |

Fig. 2. Diagram of method 1: The decision tree with local discretization

### C. Method 3

Method 3 was similar to the method 1 and based on the decision tree with local discretization. This method has used only DKF features.

### D. Method 4



| a1 | a2 | a3 | … | an | D |
|---|---|---|---|---|---|
| 1 | 34,5 | 0 | … | 1 | TR |
| 2 | 21 | 0 | … | 1 | TR |
| 2 | 17,2 | 1 | … | 1 | TR |
| … | … | … | … | … | … |

| a1 | a2 | a3 | … | an | b1 | b2 | b3 | … | bn | D |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 34,5 | 0 | … | 1 | | | | | | TR |
| 2 | 21 | 0 | … | 1 | | | | | | TR |
| 2 | 17,2 | 1 | … | 1 | | | | | | TR |
| … | … | … | … | … | | | | | | … |

Fig. 3. Diagram of method 4: Two-level Classifier with ”advisers”

Method 4 is a method based on two-level classifier with ”advisors” (Figure 3). In this approach classification decision is dependent on suggestions of domain knowledge advisers. DKA suggest decisions based on domain knowledge *e.g.*”*Left lung is located on the right side of medical image*”, ”*Object located on the left side of medical image is probably not a left lunge*”. We prepare 15 DKA for all chest organs. Advisers are divided into two groups:

- YES advisers - Advisers to advise on YES *e.g.*”*yes, this is probably the left lung*” (6 DKA),
- NO advisers - Advisers to advise on NO *e.g.*”*no, this is probably not the left lung*” (9 DKA).

Verification was followed on the basis of the DKF features *e.g.*if object center is located in region R3, R6 or R9 then YES adviser for right lunge take *false* decision. Advisers suggest what should be a decision (YES advisers) or suggested what should not be a decision (NO advisers). If any of the advisors suggested otherwise than the classifier (in some sense, the low-level classifier), decision was suspended (see [12]). All the decisions taken by the DKA pause the classifier decision where decisions are different. This is the direct reason for the decline coverage of the analyzed objects. By using domain knowledge we have obtained an improvement in the automatic classification of each chest organ. We presented the results of the experiments in the Table IV.

TABLE IV
COMPARISON OF THE RESULTS TRAIN&TEST (METHODS 1,2,3 AND 4)

| | **Method 1** | | **Method 2** | |
|---|---|---|---|---|
| Object | Acc | St.Dev. | Acc | St.Dev. |
| **TR** | **94,00%** | 3,61% | 93,97% | 4,99% |
| **RL** | **97,31%** | 0,98% | 97,56% | 0,83% |
| **LL** | **97,64%** | 0,92% | 97,56% | 1,11% |
| **RB** | **78,55%** | 6,04% | 78,12% | 4,97% |
| **LB** | **76,77%** | 6,00% | 75,74% | 8,80% |
| **LL+RL** | **87,95%** | 24,01% | 86,82% | 23,94% |
| **OT** | **94,73%** | 1,76% | 94,76% | 1,23% |

| | **Method 3** | | **Method 4** | | |
|---|---|---|---|---|---|
| Object | Acc | St.Dev. | Acc | St.Dev. | Coverage |
| **TR** | 79,61% | 6,45% | **99,90%** | 1,02% | 94,58% |
| **RL** | 81,66% | 2,89% | **98,32%** | 0,61% | 99,27% |
| **LL** | 80,73% | 3,96% | **98,44%** | 0,50% | 98,82% |
| **RB** | 60,77% | 11,65% | **88,19%** | 4,06% | 85,52% |
| **LB** | 49,33% | 7,63% | **85,04%** | 5,06% | 88,76% |
| **LL+RL** | 22,38% | 23,03% | **97,92%** | 3,50% | 94,09% |
| **OT** | 75,12% | 2,07% | **96,40%** | 1,27% | 97,57% |

### E. Comparison of classifiers by different paradigms

In Table V we give the results of experiments in applying different classification methods to our data. Those methods were developed in the following systems well known from literature: WEKA [13] and RSES [11]), and our Method 1 and 4. The coverage of all WEKA tested methods was equal 1.0 (every object was classified, without DKA). DKA approach can be used in the future in all presented WEKA methods. We used the various settings for C4.5 (WEKA) and k-NN (WEKA) methods. In Table V we presented the results of the best our approach. Tested methods gives different results depending on the class of decision.

Experimental results showed that the presented method of automatic classification of each chest organ gives good results and the results are comparable with results from other systems.

## V. CONCLUSIONS AND FURTHER WORKS

The results of experiments performed on medical data sets indicate that the presented approach seems to be promising. The use of domain knowledge significantly improved the quality of the medical object identification. The next steps will focus on the use of time dependencies between medical images (object tracking in time) and the addition of a classifier resolving conflicts between advisers.

TABLE V
COMPARISON RESULTS OF ALTERNATIVE CLASSIFICATION SYSTEMS

| Class | Result | Method 1 | Method 4 | C4.5 (WEKA) | k-NN (WEKA) | NaiveBayes (WEKA) | SVM (WEKA) | RandomForest (WEKA) |
|---|---|---|---|---|---|---|---|---|
| **TR** | ACC | 94,00% | **98,90%** | 95,31% | 97,92% | 96,27% | 93,03% | **98,76%** |
| | STD | 3,61% | 1,02% | 3,82% | 1,68% | 0,70% | 1,89% | 1,14% |
| | COV | 100% | 94,58% | 100% | 100% | 100% | 100% | 100% |
| **RL** | ACC | 97,31% | **98,32%** | 98,30% | **99,23%** | 92,59% | 89,58% | 98,79% |
| | STD | 0,98% | 0,61% | 0,92% | 0,47% | 1,21% | 0,80% | 0,52% |
| | COV | 100% | 98,82% | 100% | 100% | 100% | 100% | 100% |
| **LL** | ACC | 97,64% | **98,44%** | 97,81% | **99,08%** | 93,75% | 90,70% | 98,39% |
| | STD | 0,92% | 0,50% | 0,98% | 0,62% | 0,92% | 0,70% | 0,78% |
| | COV | 100% | 98,82% | 100% | 100% | 100% | 100% | 100% |
| **RB** | ACC | 78,55% | **88,19%** | 80,79% | 88,70% | **95,09%** | 0% | 86,46% |
| | STD | 6,04% | 4,06% | 8,33% | 2,92% | 3,66% | 0% | 4,50% |
| | COV | 100% | 85,52% | 100% | 100% | 100% | 100% | 100% |
| **LB** | ACC | 76,77% | **85,04%** | 79,80% | 81,90% | **96,64%** | 0% | 81,37% |
| | STD | 6,00% | 5,06% | 6,82% | 6,14% | 2,07% | 0% | 6,66% |
| | COV | 100% | 88,76% | 100% | 100% | 100% | 100% | 100% |
| **LL+RL** | ACC | 87,95% | **97,92%** | 95,06% | 93,90% | **100%** | 94,63% | 96,78% |
| | STD | 24,01% | 3,50% | 10,61% | 10,64% | 0% | 9,13% | 8,32% |
| | COV | 100% | 94,09% | 100% | 100% | 100% | 100% | 100% |
| **OT** | ACC | 94,73% | **96,40%** | 95,48% | 95,11% | 80,58% | 97,01% | **97,28%** |
| | STD | 1,76% | 1,27% | 0,82% | 1,56% | 2,99% | 1,32% | 1,08% |
| | COV | 100% | 97,57% | 100% | 100% | 100% | 100% | 100% |

The presented approach can be used in the future to support solving more complex medical problems. We plan to use the results of research, among other things, to treatment of an asthmatic airway remodeling (see, *e.g.*, [14] for more details) and develop more advanced methods of using domain knowledge to construct more effective classifiers.

## ACKNOWLEDGMENT

This work was partially supported by the Polish National Science Centre grant DEC-2013/09/B/ST6/01568 and by the Centre for Innovation and Transfer of Natural Sciences and Engineering Knowledge of University of Rzeszów, Poland.

## REFERENCES

[1] A. Przelaskowski, T. Podsiadly-Marczykowska, A. Wroblewska, P. Boninski, and P. Bargiel, “Computer-aided interpretation of medical images: Mammography case study,” *MG&V*, vol. 16, no. 3, pp. 347–375, Jan. 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1993447.1993457

[2] S. Nguyen, H., “Approximate boolean reasoning: Foundations and applications in data mining,” *LNCS Transactions on Rough Sets V*, vol. 4100, pp. 334–506, 2006.

[3] G. Bazan. J., S. Nguyen, H., H. Nguyen, S., P. Synak, and J. Wróblewski, “Rough set algorithms in classification problems,” in *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*, ser. Studies in Fuzziness and Soft Computing, L. Polkowski, T. Y. Lin, and S. Tsumoto, Eds. Heidelberg, Germany: Springer-Verlag/Physica-Verlag, 2000, vol. 56, pp. 49–88.

[4] A. Meyer-Baese and V. Schmid, “Chapter 2 - feature selection and extraction,” in *Pattern Recognition and Signal Analysis in Medical Imaging (Second Edition)*, second edition ed., A. Meyer-Baese and V. Schmid, Eds. Oxford: Academic Press, 2014, pp. 21 – 69. ISBN 978-0-12-409545-8

[5] J. Cytowski, J. Gielecki, and A. Gola, *Digital Medical Imaging. Theory. Algorithms. Applications*, ser. Problemy Współczesnej Nauki: Informatyka. Akademicka Oficyna Wydawnicza EXIT, 2008. ISBN 9788360434482 In Polish.

[6] A. Przelaskowski, “The role of sparse data representation in semantic image understanding,” in *Computer Vision and Graphics*, ser. Lecture Notes in Computer Science, L. Bolc, R. Tadeusiewicz, L. Chmielewski, and K. Wojciechowski, Eds. Springer Berlin Heidelberg, 2010, vol. 6374, pp. 69–80. ISBN 978-3-642-15909-1. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15910-7_8

[7] M. R. Ogiela and R. Tadeusiewicz, *Modern Computational Intelligence Methods for the Interpretation of Medical Images*, ser. Studies in Computational Intelligence. Springer, 2008, vol. 84. ISBN 978-3-540-75399-5

[8] M. Kobashi and L. G. Shapiro, “Knowledge-based organ identification from ct images,” *Pattern Recognition*, vol. 28, no. 4, pp. 475 – 491, 1995. doi: http://dx.doi.org/10.1016/0031-3203(94)00124-5

[9] C. A. Harlow and S. A. Eisenbeis, “The analysis of radiographic images,” *Computers, IEEE Transactions on*, vol. C-22, no. 7, pp. 678–689, July 1973. doi: 10.1109/TC.1973.5009135

[10] P. Selfridge, *Reasoning about Success and Failure in Aerial Image Understanding*, ser. Reports // ROCHESTER UNIV NY. University of Rochester. Department of Computer Science, 1981.

[11] G. Bazan, J. and M. Szczuka, “The Rough Set Exploration System,” *Transactions on Rough Sets*, vol. 3400, no. 3, pp. 37–56, 2005.

[12] P. Pardel, J. Bazan, J. Zarychta, and S. Bazan-Socha, “Automatic medical objects classification based on data sets and domain knowledge,” in *Beyond Databases, Architectures and Structures*, ser. Communications in Computer and Information Science, S. Kozielski, D. Mrozek, P. Kasprowski, B. Małysiak-Mrozek, and D. Kostrzewa, Eds. Springer International Publishing, 2015, vol. 521, pp. 415–424. ISBN 978-3-319-18421-0. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-18422-7_37

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: An update,” *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. doi: 10.1145/1656274.1656278. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[14] A. Niimi, H. Matsumoto, M. Takemura, T. Ueda, Y. Nakano, and M. Mishima, “Clinical assessment of airway remodeling in asthma,” *Clinical Reviews in Allergy And Immunology*, vol. 27, no. 1, pp. 45–57, 2004. doi: 10.1385/CRIAI:27:1:045

DOI: 10.15439/2015F382
ACSIS, Vol. 5

# Spatial information in classification of activity videos

Shreeya Sengupta*†, Hui Wang*‡, William Blackburn*§ and, Piyush Ojha*¶
*School of Computing and Mathematics, University Of Ulster,
Northern Ireland, UK, BT370QB
† sengupta-s@email.ulster.ac.uk
‡ h.wang@ulster.ac.uk
§ wt.blackburn@ulster.ac.uk
¶ pc.ojha@ulster.ac.uk

***Abstract*—Spatial information describes the relative spatial position of an object in a video. Such information may aid several video analysis tasks such as object, scene, event and activity recognition. This paper studies the effect of spatial information on video activity recognition. The paper firstly performs activity recognition on KTH and Weizmann videos using Hidden Markov Model and k-Nearest Neighbour classifiers trained on Histogram Of Oriented Optical Flows feature. Histogram of Oriented Optical Flows feature is based on optical flow vectors and ignores any spatial information present in a video. Further, in this paper, a new feature set, referred to as Regional Motion Vectors is proposed. This feature like Histogram of Oriented Optical Flow is derived from optical flow vectors; however, unlike Histogram of Oriented Optical Flows preserves any spatial information in a video. Activity recognition was again performed using the two classifiers, this time trained on Regional Motion Vectors feature. Results show that when Regional Motion Vectors is used as the feature set on the KTH dataset, there is a significant improvement in the performance of k-Nearest Neighbour. When Regional Motion Vector is used on the Weizmann dataset, performances of the k-Nearest Neighbour improves significantly for some of the cases and for the other cases, the performance is comparable to when oriented optical flows is used as a feature set. Slight improvement is achieved by Hidden Markov Model on both the datasets. As Histogram of Oriented Optical Flows ignores spatial information and Regional Motion Vectors preserves it, the increase in the performance of the classifiers on using Reginal Motion Vectors instead of Histogram of Oriented Optical Flows illustrates the importance of spatial information in video activity recognition.**

## I. Introduction

SPATIAL information describes the physical position of an object and its spatial relationship to other objects. It plays a crucial role in video activity recognition and may be very useful in differentiating between static and mobile activities. Mobile activities are activities where a person performing the activity moves along the field of view and static activities are activities where the person performing the activity remains at one place. Spatial information provides the position of a person and thus considering spatial information while activity recognition is expected to aid differentiation between static and mobile activities.

The importance of spatial information has also been studied by a group of researcher in Amsterdam [1]. According to them, the spatial extent of an object depends on the object to be classified itself. Spatial extent captures contextual information and for some objects the spatial extent is the whole scene whereas for some the extent is to a specific rigid boundary. For identifying functionalities such as walking or jumping, the more the context, the better it is. For identifying objects such as a car, a plane or a bus, the context is the object only.

In video activity recognition literature spatial information is often captured by various local space-time features as defined in [2], [3], [4], [5], [6], [7], [8], [9], [10], [11] and [12]. These local space-time features capture frame-wise spatial information by first detecting interest points with either interest point detectors (Harris detector, Hessian detectors, edge detector, corner detectors) or various sampling methods (dense sampling [13] or motion adaptive sampling [14]) for each frame, then spatio-temporal regions are defined around all the detected points in each frame and finally the spatio-temporal regions are described using one of the local space-time features. Other attempts to capture spatial information is contextual bag of words (CBOW) [15] and BOW with spatial pyramid [16]. Both [15] and [16] are extensions of the bag of words(BOW) feature set. BOW is a frequency based feature set which was initially used for text classification where it represents the occurrence of words in a text document. It has now been adapted in computer vision where it represents a video by counting the occurrence of a visual word in the video. Thus, BOW is a frequency based descriptor and it ignores any spatial relationship between the visual words. The extensions [15] and [16] were proposed to incorporate the missing spatial relationship explicitly in BOW.

Following the local space-time approaches this paper proposes a new feature set to preserve spatial information in activity recognition data. The new feature set, referred to as regional motion vectors (RMV), is based on optical flow vectors. Evaluation of the new set on KTH videos shows significant improvement in classification accuracy when compared with histogram of oriented optical flows (HOOF), an existing optical flow based feature set which does not preserve spatial information. The new feature set has also been assessed on the Weizmann dataset.

The next section (Section II) explains the proposed methodology, followed by the similar work section (Section III), the experimental setup section (Section IV), the results section (Section V), the conclusion (Section VI) section and finally the future works section (Section VII).

## II. Proposed Methodology

In this paper a new feature set based on optical flow vectors has been proposed. The new feature set is derived such that it preserves the spatial information present in a video. RMV was

978-83-60810-66-8/$25.00©2015, IEEE

derived by dividing a frame $f$ of resolution $n \times m$ into sub-regions using a grid of resolution $r \times s$, adding the magnitude of the optical flow vectors in each sub-region, and normalising the sum of RMVs to unity. Thus the overall relative motion $\psi$ of a sub-region $SR_{(a,b)}$ in a frame was computed as shown in Equation 1

$$\psi(SR_{(a,b)}) = \sum_{i=(a-1)r+1}^{ar} \sum_{j=(b-1)s+1}^{bs} |OFV(i,j)|/N, \quad (1)$$

where,
$a \in \{1, 2, \cdots, n_r\}$,
$b \in \{1, 2, \cdots, m_s\}$,
$n_r = \lfloor \frac{n}{r} \rfloor$,
$m_s = \lfloor \frac{m}{s} \rfloor$,
$OFV(i,j)$: Optical Flow Vector of the $i^{th}$ row and $j^{th}$ column,
$SR_{(a,b)}$: a sub-region $SR_{(a,b)}$,
$N$: normalisation constant.

Equation 1 was applied to all the sub-regions in a frame, thus producing a column vector $\Psi_t$ as shown in Equation 2:

$$\Psi_t = \{\psi(SR_{(a,b)})\}, a \in \{1, 2, \cdots, n_r\}, b \in \{1, 2, \cdots, m_s\}, \quad (2)$$

where,
$\psi(SR_{(a,b)})$: the motion of sub-region $SR_{(a,b)}$.

All such $\Psi_t$ were concatenated to form the RMV feature set (Equation 3):

$$RMV = \{\Psi_1, \Psi_2, \cdots, \Psi_T\}. \quad (3)$$

In this method, spatial information was preserved in the videos by dividing a frame into several sub-regions $SR_{(a,b)}$, $a \in \{1, 2, \cdots, n_r\}$ and $b \in \{1, 2, \cdots, m_s\}$. As the sub-regions were spatially correlated, any information extracted from these regions inherited the spatial relationship from the regions. Thus, the spatial information was not lost.

The next section (Section III) describes some of the existing features that are similar to the feature proposed in this paper.

## III. Similar work

Histogram of oriented optical flow (HOOF) proposed by Chaudhary et al. in [4] is a widely used feature set for video activity recognition. It is similar to RMV because both of them are calculated from raw optical flow vectors; however, unlike RMV, HOOF does not divide a frame into sub-regions. To extract HOOF, firstly optical flow vectors are calculated using either Horn-Schunck or Lucas-Kanade algorithm, then the flow vectors are binned into ninety angular bins, ranging from $-180°$ to $+180°$, according to their orientation and finally, the magnitude of the vectors in each of the bins is summed. Thus, while HOOF is a measure of motion in some specified directions (defined by the angular bin range) in each frame, RMV is a relative measure of motion of each sub-region in a frame. Also, while any spatial correlation among the flow vectors is lost in HOOF due to the binning strategy which ignores the spatial positioning of the flow vectors, RMV preserves such correlation by dividing a frame into several spatially correlated sub-regions.

Another feature set for representing a video which is very similar to the feature set proposed in this paper is the feature proposed by Janez Pers et al. in [11]. Similar to RMV proposed here, derivation of their representation also included dividing a frame of a video into various sub-regions. However, after the division, they calculated HOOF in each of the sub-regions unlike ours where only the relative motion of each of the sub-regions was calculated. Further, they converted the calculated HOOFs into a sequence of symbols and their final representation of a video was a sequence of symbols, the final representation of our method was a sequence of relative motions of various sub-regions in each frame of a video. Perz et al. proposed a frequency based representation of videos whereas this paper proposes a motion based representation of videos.

Raw optical flow vectors have also been used to derive space time appearance (STA) descriptor proposed in [17]. The computation of STA descriptors in [17] commenced by detecting interest regions in a video and then, the detected regions were divided into sub-regions. Dense optical flow vectors were calculated using the Farneback and TV-L1 optical flow algorithms and grid histograms representing the distribution of the optical flow vectors were computed in each sub-region. Grid histograms were concatenated to form the grid vectors and a weighted average of the grid vectors formed the order one STA (STA1) descriptors. Order two STA (STA2) descriptors were then obtained by combining the grid histograms to form component vectors and then binning the component vectors into k2 bins. The final feature was obtained by concatenating the STA2 descriptors into a vector. The only similarities between STA and RMV are the division of interest regions into sub-regions and use of optical flow vectors. While STA is again a frequency based approach which represents the distribution of optical flow vectors upto two orders, RMV does not represent any such distribution and only measures the motion of sub-regions in a frame.

While the features proposed in [11] and [17] preserve spatial information, HOOF in [4] does not preserve any spatial information. As the main aim of the proposed method was to preserve the spatial information in video data to aid activity recognition, the effectiveness of the proposed method has been studied by comparing its performance only with HOOF. HOOF is a feature set derived from optical flow vectors without any spatial information and the proposed method (RMV) is a feature set again derived from optical flow vectors but with spatial information. Thus, a comparison between the two methods is expected to illustrate the effectiveness of the proposed method as well as the importance of spatial information in video activity recognition.

The next section (Section IV) describes the experimental setup for comparing both the methods.

## IV. EXPERIMENTAL SETUP

RMV and HOOF were tested on videos from the KTH dataset [18] and the Weizmann dataset [19].

The KTH dataset is a video dataset consisting of six different human actions, namely boxing, hand-waving, hand-clapping, jogging, running and walking. These six actions were performed by twenty five subjects under four different scenarios: outdoors, outdoors with varying scale, outdoors with subjects wearing a variety of clothes and indoors. All the videos were taken over homogenous backgrounds with a static camera and a frame rate of twenty five frames per second. For this study, each video of this dataset was further divided into four sub-videos, and therefore, with twenty five people, six actions, four scenarios and four sub-videos, there are in total 2400 sub-videos in the dataset. Out of these, 120 sub-videos of each action were selected randomly, thus providing a total of 720 sub-videos for experimentation.

The Weizmann dataset [19] contains nine people performing ten different actions: gallop, jump, walk, run, gallop sideways, bend, one hand waving, jumping jack,two hand waving, jumping in place and skip. The actions were recorded at a resolution of 180 × 144.

For both the datasets in this study, the optical flow vectors for deriving HOOF were obtained using the Lucas-Kanade algorithm [20]. The optical flow vectors were then sorted into ninety angular bins, each $4^\circ$ wide, collectively covering the full angular range from $-180^\circ$ to $180^\circ$. The magnitudes of the optical flow vectors in each bin were added to produce a ninety dimensional optical flow vector (or histogram) for each frame. Thus for a $T$ frame sequence, we get a $90 \times T$ dimensional matrix referred to as HOOF. Principle component analysis was then used to reduce the data dimension to five, eight, twelve and sixteen.

For deriving regional motion vectors (RMV) feature set, the Lucas-Kanande algorithm was used to compute frame-wise optical flow vectors where each vector again had two dimensions - the magnitude and direction. Then, for the KTH dataset, instead of the angular bins, each frame was divided into sub-regions $SR_{(a,b)}$, $a \in \{1, 2, \cdots, n_r\}$ and $b \in \{1, 2, \cdots, m_s\}$ by using a patch of resolution $r = 10$ by $s = 20$. As the resolution of each frame was $n = 120$ by $m = 160$, so, each frame was divided into $(120/10) * (160/20) = 96$ regions. The vectors in each of these regions were grouped into one bin and their magnitudes were added. Vector sum of these vectors could also have been considered. However, as there was no significant difference, only the results obtained using magnitude have been listed. The value of $r$ and $s$ (size of the patch) could also have been varied. However, the size of the patch was chosen such that the number of bins was near to ninety - the number of bins in HOOF. To summarize, RMV produced a $96 \times 1$ dimensional column vector for each frame, and for a sequence of $T$ frames, a $96 \times T$ dimensional matrix. This matrix was known as regional motion vectors (RMV). Principle component analysis was again used to reduce the data dimension to five, eight, twelve and sixteen.

Similarly, RMV features for the Weizmann action videos were obtained by dividing each frame into sub-regions instead of angular bins. Again, a patch was used for the purpose, however, with a different resolution. The resolution of the patch used was $r = 18$ by $s = 16$ and since, the resolution of the frames were $n = 180$ by $m = 144$, $90 \times 1$ dimensional column vector was produced for each frame. For a sequence of $T$ frames, $90 \times T$ dimensional RMV matrix was produced. The dimension of the matrix was reduced to five, eight, twelve and sixteen using principal component analysis.

Once HOOF and RMV features were extracted, they were used with k-nearest neighbour (kNN) and hidden Markov model (HMM) classifiers. An unclassified pattern is assigned to the class of its nearest neighbour. The similarity between two *points* in the multi-dimensional space was defined either via their Euclidean distance ($EUC$) or by a neighbourhood counting similarity metric ($NCM$) [21]. These two measures can be extended directly to patterns, i.e. *sequences of points*. (When computing the Euclidean distance between sequences of unequal length, we truncate the longer sequence so that its length matches the shorter sequence.) Alternatively, $EUC$ and $NCM$ can be used as the underlying point-to-point similarity measures in dynamic time warping ($DTW$) [22], longest common subsequence ($LCS$) [23] or all common subsequences ($ACS$) [24] measures of similarity between sequences. Thus, eight measures of similarity were evaluated between sequences: $EUC$, $NCM$, $DTW + EUC$, $DTW + NCM$, $LCS + EUC$, $LCS + NCM$, $ACS + EUC$ and $ACS + NCM$. The performances of these measures were evaluated because some of these measures are widely used and are known to handle variation in time series data well. The kNN classifier along with the similarity measures were coded from scratch.

HMM was used with three Gaussian and six states and $EUC$ was the default similarity measure for this model. HMM was implemented using the Kevin Murphy toolbox [25] for HMM.

The classification regime was ten fold cross validation. The dataset was arranged such that the test set contained one video from each action category and the remaining videos from those categories were used as the training set. The following section (Section V) lists the performance of the two classifiers.

## V. RESULTS

This section lists the performance of all the classifiers using HOOF and RMV feature set on KTH and Weizmann videos. In this study HOOF is the optical flow feature set which lacks spatial information, i.e. any spatial relationship in the videos is lost when HOOF is extracted. On the contrary, when RMV feature is extracted, spatial correlations in the videos are also preserved. Thus, it is expected that a comparison of the performances of several classifiers(HMM and kNN in this study) using HOOF and RMV will illustrate the importance of considering spatial information during video activity recognition.

In this section, first the performance of the classifiers on the KTH dataset is listed, followed by the performance of the classifiers on the Weizmann dataset.

### *A. Performance of HMM and kNN on KTH dataset*

Table I presents the performance of kNN and HMM on the KTH dataset. In the table the first column represents the feature set used (RMV or HOOF) for activity recognition, the second column lists the classifiers used. It can be noted that several similarity measures ($EUC$, $NCM$, $DTW + EUC$, $DTW + NCM$, $LCS + EUC$, $LCS + NCM$, $ACS + EUC$ and $ACS + NCM$) have been explicitly specified with kNN in the second column. The measures indicate the method used with kNN to calculate similarity between two sequences while recognising activities. HMM was used with only $EUC$ distance measure and hence the measure has not been specified explicitly. The following columns three to six list the performance of both the classifiers on varying dimension of HOOF and RMV data. The varying dimensions have been indicated in the header as 5 PCs, 8 PCs, 12 PCs and 16 PCs. Here, PC stands for principle components and 5, 8, 12 and 16 stands for the number of principle components selected.

TABLE I. THE PERFORMANCE OF HMM & KNN WITH VARIOUS SIMILARITY MEASURES ON KTH DATASET USING BOTH THE HOOF AND THE RMV FEATURES. IN THE COLUMN HEADINGS, PC STANDS FOR PRINCIPLE COMPONENTS.

| Feature set | Classifiers | $5PCs$ | $8PCs$ | $12PCs$ | $16PCs$ |
|---|---|---|---|---|---|
| HOOF | HMM | 59 | 63 | 68 | 68 |
| | kNN + EUC | 59 | 64 | 67 | 66 |
| | kNN + NCM | 19 | 17 | 16 | 16 |
| | kNN + (DTW + EUC) | 60 | 67 | 70 | 72 |
| | kNN + (DTW + NCM) | 21 | 22 | 24 | 25 |
| | kNN + (LCS + EUC) | 26 | 25 | 25 | 24 |
| | kNN + (LCS + NCM) | 23 | 26 | 25 | 23 |
| | kNN + (ACS + EUC) | 22 | 21 | 21 | 21 |
| | kNN + (ACS + NCM) | 21 | 21 | 22 | 19 |
| RMV | HMM | 61 | 67 | 72 | 71 |
| | kNN + EUC | 86 | 88 | 89 | 90 |
| | kNN + NCM | 70 | 56 | 50 | 43 |
| | kNN + (DTW + EUC) | 77 | 77 | 78 | 79 |
| | kNN + (DTW + NCM) | 33 | 38 | 45 | 50 |
| | kNN + (LCS + EUC) | 34 | 38 | 33 | 16 |
| | kNN + (LCS + NCM) | 34 | 37 | 40 | 46 |
| | kNN + (ACS + EUC) | 33 | 35 | 33 | 33 |
| | kNN + (ACS + NCM) | 33 | 34 | 33 | 33 |

Following Table I, the performance of HMM and kNN on varying dimensional HOOF has been compared with their performance of varying dimensional RMV. The comparisons have been represented graphically in Figures 1, 2, 3 and 4. In the figures, classifiers are listed on the x-axis and the classification accuracy (activity recognition rate) achieved by them is listed on the y-axis.

Figure 1 compares the performance of HMM and kNN classifiers using five dimensional HOOF with their performance using five dimensional RMV data. It can be observed that performance of both HMM and KNN improves when RMV is used instead of HOOF. For HMM the improvement is marginal from $59\%$ when HOOF is used to $61\%$ when RMV is used. Significant improvement is noticed in cases where kNN was used with $EUC$ and $NCM$ similarity measures. In case of kNN with $EUC$ the change was from $59\%$ to $86\%$ and in case of kNN with $NCM$ it was from $18\%$ to $70\%$. The improvement in accuracy when RMV is used instead of HOOF indicates the importance of preserving spatial information in video data.

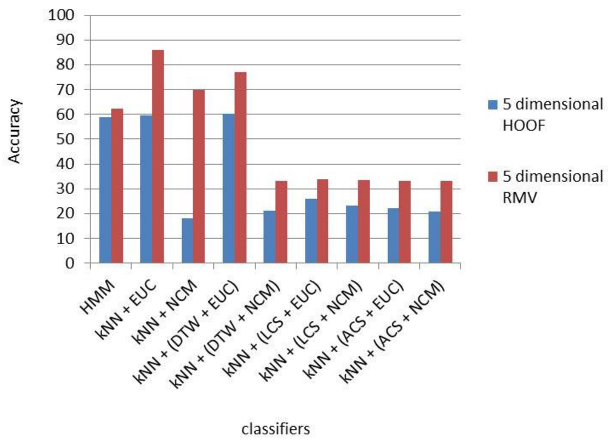


Fig. 1. Comparison of the performances of different classifiers on five dimensional HOOF and RMV data.

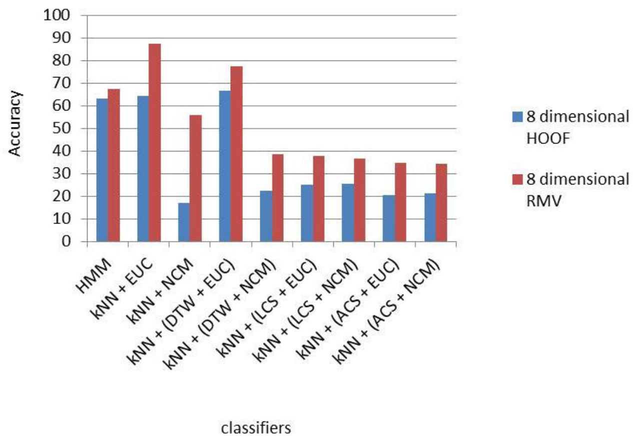


Fig. 2. Comparison of the performances of different classifiers on eight dimensional HOOF and RMV data.

Figure 2 presents a comparison similar to Figure 1. How-

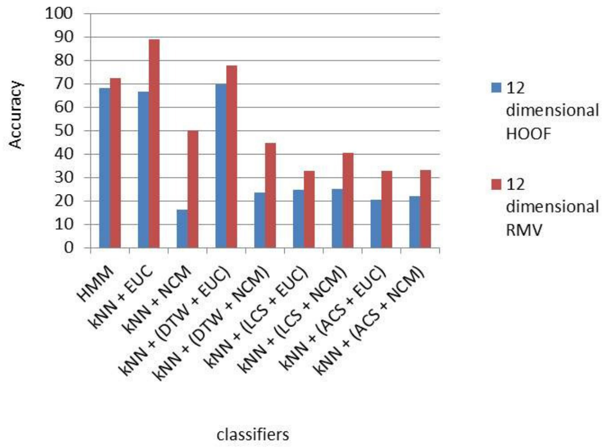


Fig. 3. Comparison of the performances of different classifiers on twelve dimensional HOOF and RMV data.

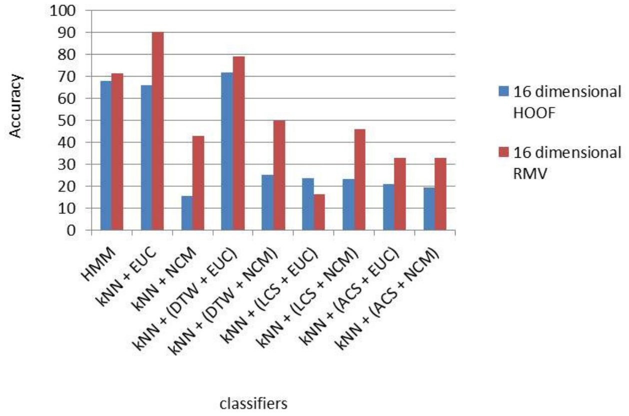


Fig. 4. Comparison of the performances of different classifiers on sixteen dimensional HOOF and RMV data.

ever, instead of five dimensional HOOF and RMV data, eight dimensional HOOF and RMV was used. It is observed that again an improvement in classification accuracy is achieved when RMV is used instead of HOOF. When RMV is used with kNN for activity recognition $EUC$ and $NCM$ similarity measures perform significantly better than when HOOF is used with kNN. $kNN + EUC$ achieved an accuracy of $64\%$ when HOOF is used (spatial information in videos is ignored) and it achieves an accuracy of $88\%$ when RMV is used (spatial information in KTH videos is considered). The accuracy obtained by $kNN + NCM$ in absence of spatial information (HOOF feature set is used) is $17\%$ and in presence of spatial information (RMV feature is used) is $56\%$. The results thus again show the importance of considering such spatial information.

Trend similar to Figure 1 and Figure 2 is also noted in Figure 3 and Figure 4, illustrating the importance of considering spatial information in videos.

It is also noted that in Figures 1, 2, 3 and 4 kNN with $DTW + EUC$ performs well consistently. When HOOF is used the accuracies are $60\%$, $67\%$, $70\%$ and $72\%$ for five, eight, twelve and sixteen dimensional HOOF data and when RMV is used the accuracies are $77\%$, $77\%$, $78\%$ and $79\%$ again for five, eight, twelve and sixteen dimensional data. Thus, for $DTW + EUC$ the accuracies are relatively stable through varying dimensional data. This was expected because $DTW$ is designed in such a way that it calculates a similarity score between two given sequences by matching each element of one sequence with every element of the other sequence. Such a matching facilitates comparing sequences having different number of frames which is very common in video data. Varying number of frames introduce a different type of variation which also hinders activity recognition. As $DTW$ handles such variation, its performance is relatively steady and superior to other similarity measures (for example when the data is five dimensional data, $DTW + EUC$ with kNN gives an accuracy of $60\%$ which is higher than other measures such as $NCM$ (18%), $LCS + EUC$ (26%), $LCS + NCM$ (23%), $ACS + EUC$ (22%) and $ACS + NCM$(21%).) during activity recognition.

However, it is also observed that although $DTW + EUC$ outperformed all the measures but $DTW + NCM$ performed poorly. This can be attributed to the different underlying *point to point similarity measure* that has been used with $DTW$. When the performance of $EUC$ and $NCM$ alone is compared, it is observed that when HOOF is used $EUC$ performs significantly better than $NCM$ with an accuracy of $59\%$ (accuracy obtained using $NCM$ is $18\%$). A possible explanation for such a performance of $NCM$ is its ability to work better on correlated data than uncorrelated data. This explanation is supported by the improvement in $NCM$'s performance when RMV, where the data is spatially correlated is used instead of HOOF, where the data is uncorrelated. The accuracy obtained by $NCM$ when RMV is used is $70\%$ and when HOOF is used is $18\%$. The behaviour of $NCM$ on correlated and uncorrelated data extends to $DTW$ when these two measures are used as the underlying point to point measure with $DTW$. Thus, $DTW + EUC$ performs superior to $DTW + NCM$. This behaviour of $NCM$ also supports considering spatial relationships and preserving spatial information during video activity recognition.

The next subsection lists and reviews the performance of kNN and HMM on Weizmann videos in the presence and absence of spatial information. Again, presence of spatial information is ensured by using RMV as the feature set and absence of the information is ensured by using HOOF.

### B. Performance of kNN and HMM on Weizmann dataset

In this subsection table II shows the performance of kNN and HMM using HOOF and RMV features. Similar to table I, column one shows the feature set being used, column 2 lists the classifiers and the rest of the columns (3-6) lists the obtained classification accuracy with varying dimensional HOOF and RMV.

Following the table are figures 5, 6, 7 and 8 which compare the performances of the classifiers while using five, eight, twelve and sixteen dimensional HOOF and RMV respectively.

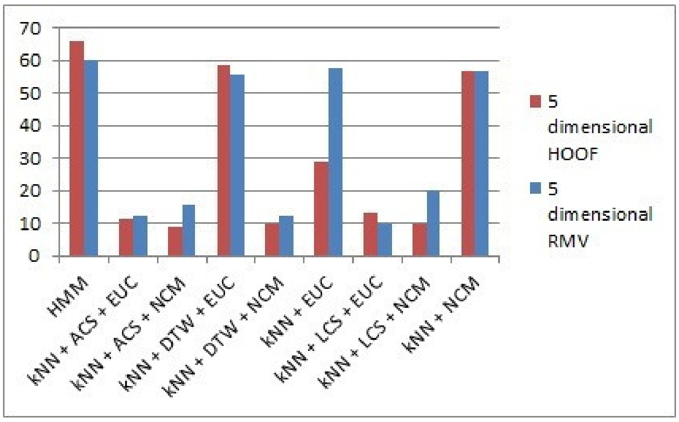


Fig. 5. Comparison of the performances of different classifiers on Weizmann dataset using five dimensional HOOF and RMV.

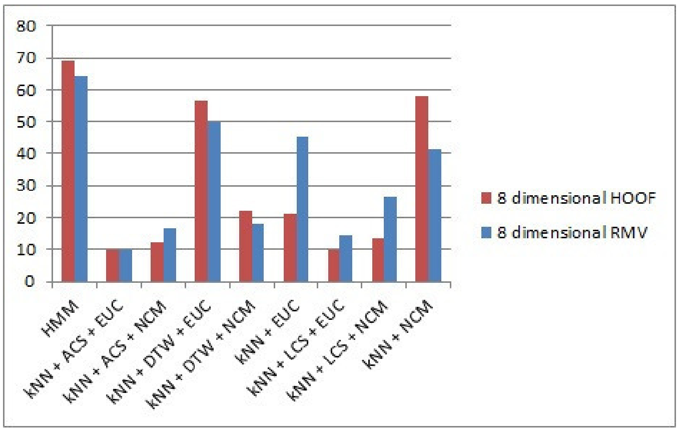


Fig. 6. Comparison of the performances of different classifiers on Weizmann dataset using eight dimensional HOOF and RMV.

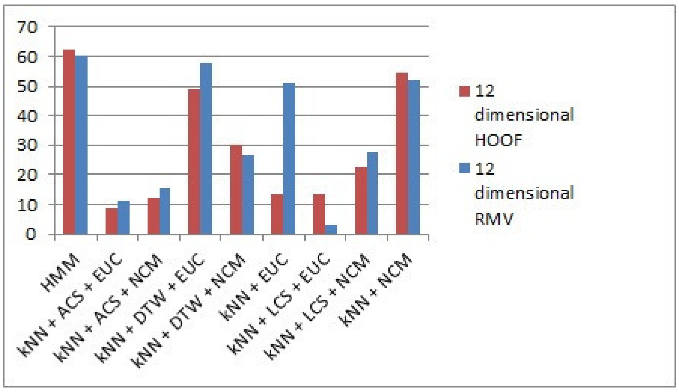


Fig. 7. Comparison of the performances of different classifiers on Weizmann dataset using twelve dimensional HOOF and RMV.

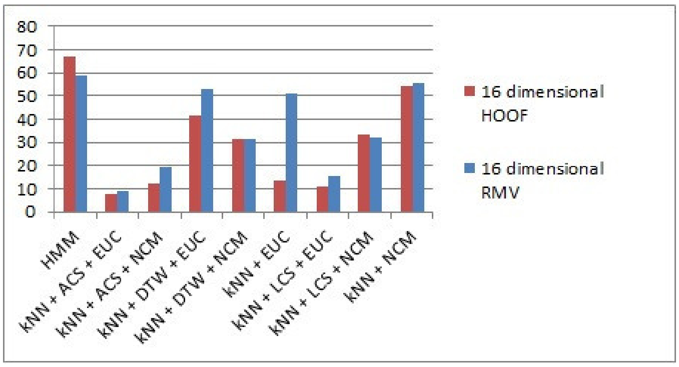


Fig. 8. Comparison of the performances of different classifiers on Weizmann dataset using sixteen dimensional HOOF and RMV.

TABLE II. THE PERFORMANCE OF HMM & KNN WITH VARIOUS SIMILARITY MEASURES ON WEIZMANN DATASET USING BOTH THE HOOF AND THE RMV FEATURES. IN THE COLUMN HEADINGS, PC STANDS FOR PRINCIPLE COMPONENTS.

| Feature set | Classifiers | $5PCs$ | $8PCs$ | $12PCs$ | $16PCs$ |
|---|---|---|---|---|---|
| HOOF | HMM | 66 | 69 | 62 | 67 |
| | kNN + EUC | 29 | 21 | 13 | 13 |
| | kNN + NCM | 57 | 58 | 54 | 54 |
| | kNN + (DTW + EUC) | 59 | 57 | 49 | 41 |
| | kNN + (DTW + NCM) | 10 | 22 | 30 | 31 |
| | kNN + (LCS + EUC) | 13 | 10 | 13 | 11 |
| | kNN + (LCS + NCM) | 10 | 13 | 22 | 33 |
| | kNN + (ACS + EUC) | 11 | 10 | 9 | 8 |
| | kNN + (ACS + NCM) | 9 | 12 | 12 | 12 |
| RMV | HMM | 60 | 64 | 60 | 59 |
| | kNN + EUC | 58 | 46 | 51 | 51 |
| | kNN + NCM | 57 | 41 | 52 | 56 |
| | kNN + (DTW + EUC) | 56 | 50 | 58 | 53 |
| | kNN + (DTW + NCM) | 12 | 18 | 27 | 31 |
| | kNN + (LCS + EUC) | 10 | 14 | 3 | 16 |
| | kNN + (LCS + NCM) | 20 | 27 | 28 | 32 |
| | kNN + (ACS + EUC) | 12 | 10 | 11 | 9 |
| | kNN + (ACS + NCM) | 16 | 17 | 16 | 19 |

From the figures, it can be seen that on several occasions, the performance of kNN increases when RMV is used instead of HOOF. An example is the improvement in classification accuracy from 29 to 58 for kNN+EUC when 5PC RMV is used instead of 5PC HOOF. Another example is the performance of kNN + (DTW+EUC) using 12 dimensional features. When 12 dimensional HOOF is used, the classification accuracy is 49, and when 12 dimensional RMV is used, the classification accuracy is 58. However, for some cases, the performance of the classifiers are either comparable or remains the same. For example,the highest classification accuracy obtained using kNN+(DTW+EUC) and 5 dimensional HOOF feature set is 59. However, when 5 dimensional RMV is used for the same case, the accuracy is 56. Then the performance of kNN + NCM on 5 dimensional HOOF is 57 which remains unchanged when 5 dimensional RMV is used.

Therefore, from the above results it can be observed that there is a significant improvement in the performance of the classifiers when RMV is used instead of HOOF on the KTH dataset. However, no such significant performance difference was noted for the classifiers when tested on Weizmann dataset using RMV and HOOF features. This can be attributed to the lack of direction information in the RMV feature set. Incorporating direction information in the feature set ensures that the direction (left to right or right to left) of mobile actions (jog, run, skip, walk) does not affect the classification / recognition results. As there are less number of mobile actions in the KTH dataset than the Weizmann dataset, lack of this information does not affect the overall performance of the classifiers on the KTH dataset. However, a closer look at the confusion matrix of the KTH dataset reveals that most of the misclassification is among the mobile activities, which in case of RMV can be attributed to the absence of direction information in the RMV feature set.

The competitive and sometimes better performance of all the classifiers on the Weizmann dataset does not reduce the significance of the proposed feature set. On the contrary, it shows the potential of the feature by performing well (on KTH dataset) and at par (on Weizmann dataset), even after the lack of direction information. This shows that beside direction, spatial relationships also play an important role in action classification and, thus, features preserving such spatial relationships are required.

Finally, it can be observed that the dimension of the data has been varied from five to sixteen. Data higher than sixteen dimension was not considered because the performance of the classifiers for most of the cases attains stability after eight dimensional HOOF and RMV data. Further, the computation time of kNN increases with increasing dimension and hence considering very high dimensional data is undesirable. Last but not the least, training HMM on such high dimensional data is not only time consuming but requires large number of training samples. As huge number of data may not always be available, data with dimension higher than sixteen is not considered.

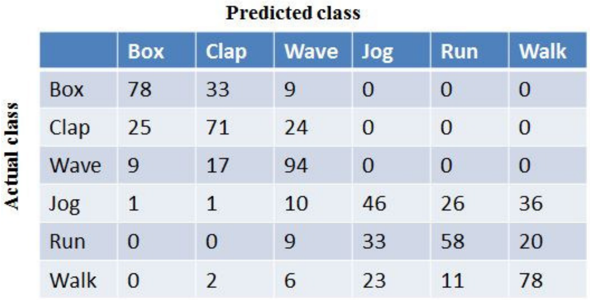
## VI. Conclusion

The main aim of this paper was to study the effects of spatial information in video data analysis. For this, the paper focussed on video activity recognition and videos from the KTH and Weizmann datasets, which are activity datasets, were selected for this purpose. Recognition was performed using two classifiers - kNN and HMM, trained firstly on HOOF and then on RMV features. HOOF is an optical flow based feature set which ignores any spatial information and therefore applying any classifier on such a feature set illustrates a scenario where spatial information has been ignored (absence of spatial information). RMV on the contrary is an optical flow based feature set where spatial information is preserved. Thus, any spatial relationships among objects in a video or object and background in a video is present in the feature set. A comparison of the performances of the classifiers with HOOF and RMV illustrated the effect of the presence and absence of spatial information on video activity recognition. Results show significant improvement in the performance of kNN classifier on KTH dataset, when spatial information is preserved. For example, kNN with EUC achieved an accuracy of 59% when trained with five dimensional HOOF data. This accuracy improved to 86% when five dimensional RMV data was used to train the kNN.

On the Weizmann dataset, when RMV is used in place of HOOF, the performance of the classifiers in some cases shows significant improvement. For example, while using kNN with EUC distance, the accuracy increases from 29% to 58% (when RMV is used instead of HOOF). Another example is when kNN is used with DTW+EUC, the accuracy increases from 49% to 58% when 12 dimensional RMV is used instead of 12 dimensional HOOF. For other cases, the performance of the classifiers are comparable. For example, the performance of kNN + NCM remains the same, at 57%, irrespective of the feature used.

## VII. Future work

Although the performance of kNN on using RMV, either increased or was comparable, improvement in the performance of HMM however was not very high. On the KTH dataset, with HOOF as the feature set, HMM achieved an recognition rate of 59% and with RMV feature set it obtained a recognition rate of 61%. Therefore, there is a scope of improving the performance of HMM further. Figures 9 and 10 shows the confusion matrices created by HMM when HOOF and RMV features are used respectively on the KTH dataset.

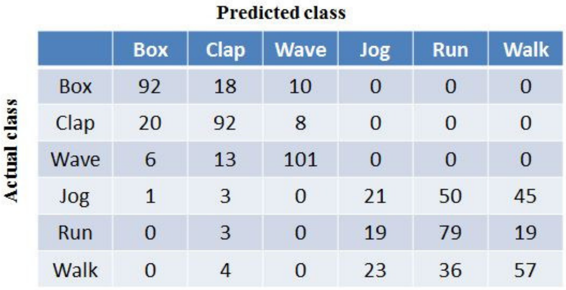


Predicted class

Actual class

| | Box | Clap | Wave | Jog | Run | Walk |
|---|---|---|---|---|---|---|
| Box | 78 | 33 | 9 | 0 | 0 | 0 |
| Clap | 25 | 71 | 24 | 0 | 0 | 0 |
| Wave | 9 | 17 | 94 | 0 | 0 | 0 |
| Jog | 1 | 1 | 10 | 46 | 26 | 36 |
| Run | 0 | 0 | 9 | 33 | 58 | 20 |
| Walk | 0 | 2 | 6 | 23 | 11 | 78 |

Fig. 9. Confusion matrix showing the performance of HMM when HOOF feature is used for activity recognition.



Predicted class

Actual class

| | Box | Clap | Wave | Jog | Run | Walk |
|---|---|---|---|---|---|---|
| Box | 92 | 18 | 10 | 0 | 0 | 0 |
| Clap | 20 | 92 | 8 | 0 | 0 | 0 |
| Wave | 6 | 13 | 101 | 0 | 0 | 0 |
| Jog | 1 | 3 | 0 | 21 | 50 | 45 |
| Run | 0 | 3 | 0 | 19 | 79 | 19 |
| Walk | 0 | 4 | 0 | 23 | 36 | 57 |

Fig. 10. Confusion matrix showing the performance of HMM when RMV feature is used for activity recognition.

The matrices reveal the following flaws while performing activity recognition:

- The classifier discriminates very well between static (boxing, clapping, waving) and mobile (jogging, walking and running) activities, less well between different static activities, and quite poorly between different mobile activities. Static activities in a video refer to activities where a subject is standing at one constant position. Mobile activities refer to activities where a subject is moving along the field of view throughout the video. Thus, while boxing, clapping and waving in the KTH dataset are static activities, jogging, running and walking are mobile activities.

- The pace of mobile activities increases naturally from walking to jogging to running. Intuitively, one would expect that walking is misclassified as jogging more frequently than it is misclassified as running, and, likewise, running is misclassified as jogging more frequently than it is misclassified as walking. However, this is not reflected in the confusion matrix which was obtained using RMV.

Future attempts to improve the performance of HMM can concentrate on solving these flaws.

Some other possible future works are as follows:

- RMV ignores direction of the flow vectors. For activity recognition direction is one of the important factors. Future attempts will try to incorporate direction into RMV.
- In RMV spatial information was incorporated by dividing a frame into sub-regions which are spatially correlated to each other. This division was done using a grid of resolution $r \times s$. The values of $r$ and $s$ were chosen as 10 and 20 respectively for the KTH dataset and as 18 and 16 respectively for the Weizmann dataset. However, these are not constant values and can be varied, i.e. a grid of different resolution can be chosen for the same dataset. Also, the values of $r$ and $s$ varies from one dataset to another. Future research can concentrate on studying the effect of varying $r$ and $s$ on a dataset and also on coming up with a more principled approach of selecting the values of $r$ and $s$.
- Next, converting any feature set to their BOW representation and then using them with SVM for activity recognition is the state of the art in the activity recognition field. Thus, in future, RMV can also be converted it to its BOW representation and used with SVM to further assess its potentiality in video activity recognition. Further, the performance of RMV in its BOW form can be compared with the performance of HOOF in its BOW format. As RMV captures spatial information, spatial information is also added to its BOW representation. Similarly, as HOOF does not capture spatial information, its BOW form does not inherit any such information. A comparison between the BOW of RMV and BOW of HOOF will further illustrate the importance of spatial information.
- Finally, BOW itself lacks any spatial information and previously attempts have been made to incorporate spatial information in BOW. Some of these works include spatial pyramid [16] and CBOW [15]. A comparison of RMV in its BOW form with these works ([16] and [15]) on activity data may be another interesting research direction.

## Acknowledgment

Shreeya Sengupta is supported by Vice Chancellor's Research Scholarship of University Of Ulster, UK.

## References

[1] J. R. R. Uijlings, A. Smeulders, and R. J. H. Scha, "What is the spatial extent of an object?" in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 770–777. [Online]. Available: http://www.huppelen.nl/publications/spatialExtentCvpr.pdf

[2] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, 1999, pp. 1150–1157 vol.2. [Online]. Available: http://dl.acm.org/citation.cfm?id=850924.851523

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *In CVPR*, 2005, pp. 886–893. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.177

[4] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1932–1939. [Online]. Available: http://dx.doi.org/10.1109/CVPRW.2009.5206821

[5] K. Brkić, A. Pinz, S. Šegvić, and Z. Kalafatić, "Histogram-based description of local space-time appearance," in *Proceedings of the 17th Scandinavian Conference on Image Analysis*, ser. SCIA'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 206–217.

[6] C. Tsai, "Bag-of-words representation in image annotation: A review," *ISRN Artificial Intelligence*, vol. 2012, p. 19 pages, 2012. [Online]. Available: http://dx.doi.org/10.5402/2012/376804

[7] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 925–931.

[8] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, Sep. 2008. [Online]. Available: http://dx.doi.org/10.1007/s11263-007-0122-4

[9] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, sep 2008, pp. 995–1004.

[10] F. M. Carrillo, A. Manzanera, and E. R. Castro, "A motion descriptor based on statistics of optical flow orientations for action classification in video-surveillance," in *Int. Conf. on Multimedia and Signal Processing (CMSP'12)*, Shanghai, China, december 2012.

[11] J. Pers, V. Sulic, M. Kristan, M. Perse, K. Polanec, and S. Kovacic, "Histograms of optical flow for efficient representation of body motion." *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1369–1376, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2010.03.024

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 1–8, 2008.

[13] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 3169–3176. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2011.5995407

[14] M. Grundmann, F. Meier, and I. A. Essa, "3d shape context and distance transform for action recognition." in *ICPR*. IEEE, 2008, pp. 1–4.

[15] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 4, pp. 381–392, 2011. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2010.2041828

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 2169–2178. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.68

[17] K. Brkic, S. Rasic, A. Pinz, S. Segvic, and Z. Kalafatic, "Combining spatio-temporal appearance descriptors and optical flow for human action recognition in video data," *CoRR*, vol. abs/1310.0308, 2013.

[18] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, 2004, pp. 32–36 Vol.3. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2004.747

[19] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2007.70711

[20] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision (ijcai),” in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI ’81)*, April 1981, pp. 674–679.

[21] H. Wang, “Nearest neighbors by neighborhood counting,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 942–953, June 2006. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2006.126

[22] P. Senin, “Dynamic Time Warping Algorithm Review,” Department of Information and Computer Sciences, University of Hawaii, Honolulu, Hawaii 96822, Tech. Rep. CSDL-08-04, Dec. 2008.

[23] D. S. Hirschberg, “A linear space algorithm for computing maximal common subsequences,” *Commun. ACM*, vol. 18, no. 6, pp. 341–343, Jun. 1975. [Online]. Available: http://doi.acm.org/10.1145/360825.360861

[24] H. Wang, “All common subsequences,” in *Proceedings of the 20th international joint conference on Artifical intelligence*, ser. IJCAI’07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 635–640.

[25] K. Murphy, “Hidden Markov Model (HMM) toolbox for matlab,” 1998.

DOI: 10.15439/2015F411
ACSIS, Vol. 5

# DiverGene: Experiments on Controlling Population Diversity in Genetic Algorithm with a Dispersion Operator

Anna Strzeżek', Ludwik Trammer', Marcin Sydow' "
'Polish-Japanese Institute of Information Technology
ul. Koszykowa 86, 02-008 Warszawa, Poland
"Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland
Email: astrzezek@gmail.com, ludwik@trammer.pl, msyd@poljap.edu.pl

***Abstract*—We present diverGene – a novel, diversity-aware population selection operator for genetic algorithm – to be used especially for particularly complex and multi-criteria optimisation problems. Genetic algorithm is one of the most known evolutionary algorithms for solving hard optimisation problems. Many attempts have been made to improve its convergence rate and quality of the result. In this paper we propose a novel extension of the selection operator that makes it possible to control the level of diversity in the population. We discuss its theoretical background, including its computational hardness and propose an efficient way of computing it. The approach is implemented and tested on three hard optimisation problems: Knapsack Problem, Travelling Salesman Problem and a relatively new Travelling Thief Problem that might be viewed as the composition of the latter two. We report experimental results that seem to indicate that the novel approach has a potential to improve the quality of the results for some hard optimisation problems.**

## I. Introduction

In this paper we present `diverGene` – a novel, diversity-aware population selection operator for genetic algorithm. The idea is based on the concept of the *dispersion* of the solutions modelled by means of pair-wise dissimilarity between the solutions.

The proposed operator seems to be especially useful for particularly complex optimisation problems of multi-criteria nature, where the solution space should be intensively explored due to the potential variety and mutual non-similarity of potential good solutions to the problem.

### A. Genetic Algorithm

The *genetic algorithm* [1] is a heuristic for finding satisfactory solutions to problems, using a process inspired by the natural selection. It does not guarantee to find the optimal solution and is intended to be used mainly in cases when finding the optimal solution directly would be too computationally expensive to be practical (more formally for NP-hard optimisation problems). In such cases search heuristics are often used to find a solution that is usually not the best possible one, but is still useful.

The process starts by randomly creating a set of feasible solutions to the problem. This set of solutions is called *population* and each solution in the population is called *an individual*, in reference to the biological origins of the algorithm. The initial state of the population is called *the initial (first) generation*. During the course of the algorithm execution, the subsequent generations are created, each one based on its immediate predecessor. While the solutions encoded in the initial generation are completely random, the quality of solutions in each subsequent generation should gradually get better, in an evolutionary fashion.

*1) Selection:* The process of selection is responsible for the decision on *which* individuals from the current generation should be used when creating the next generation (or, in other words, how many children should an individual have - if any). This step is essential to the process. Properly defined selection causes the best individuals to reproduce more intensively, resulting in a general increase of the solution quality in the population. On the other hand a selection that is *too* strong may lead to some problems - if the algorithm were to always simplistically select only a small subset of the very best solutions, there would be a high likelihood of losing the diversity and getting stuck in a local maximum. The problem is traditionally solved by introducing an element of randomness - the better the solution the greater the possibility it will be selected as a parent. As a consequence, other solutions also have a chance of being selected, while the overall solution quality generally increases.

*2) Mutation Operator:* The process of selection results in the increase of the *overall* solution quality in a population by generally eliminating weak solutions and reproducing stronger ones, but by itself it doesn't improve the quality of individual solutions. It's the mutation operator that enables the algorithm to explore other (possibly better) solutions. During the mutation stage there is a small chance of individual elements of a solution being randomly changed, resulting in a solution that is similar to one previously selected, but slightly different. Sometimes the changes are beneficial, but perhaps even more

This work is partially supported by Polish Ministry of Science and Higher Education grant N N519 578038.

978-83-60810-66-8/$25.00©2015, IEEE

often they make the solution weaker. During the subsequent selection stages the stronger solutions are more likely to be chosen than the weaker ones. As a consequence individuals with harmful mutations will eventually be eliminated, while individuals with beneficial mutations will become the parents for multiple newly created individuals, some of which will in turn be mutated, bringing the possibility of further improvements. The mutation operator corresponds to the concept of *random local search* in the solution space.

*3) Cross-over Operator:* Cross-over stage brings a possibility of two different solutions being combined. This allows for new individuals that include good parts of multiple previously known solutions. Without the cross-over stage there would be no sharing of good "ideas" within the population.

### B. Motivation

As discussed in the "Selection" section above, too much focus on promoting only the best solutions in the current population during the selection stage (at the expense of alternative candidate solutions) may result in inadequate exploration of the space of possible solutions, or, equivalently in getting stuck in a local optimum.

On the other hand too little focus on choosing good result will result in a weak evolutionary pressure and population not getting better over time. The balance is traditionally achieved by giving better quality solutions a greater probability of being selected, but allowing other solutions to also be occasionally selected by random chance. While this method allows for some additional diversity it does not guarantee it.

In this paper, we investigate whether controlling the level of diversity within a population at the expense of a slight drop in the overall population quality would have a positive effect on the overall performance of the algorithm. In particualr we want to explore the effect of this aproach on problems containing multiple interdependent sub-problems, using the Travelling Thief Problem [8] as an example. It seems that increased population diversity would especially improve the performance of the genetic algorithm for such compound hard optimisation problems.

### C. Contributions

The contributions of this paper are the following:

- we propose `diverGene` – a novel *diversity-aware* selection operator that makes it possible to control the balance between the exploration and exploatation of the solution space. The selection is based on both fitness and diversity among its members.
- we report development of a flexible software framework for testing the operator with various settings.
- we report experimental results on three hard optimisation problems: the Knapsack problem, the Travelling Salesman Problem and the Travelling Thief Problem (the last one being particularly noteworthy since it is composed of two interdependent sub-problems). The experiments concern the impact of the operator in various settings on the quality of the solutions.

### D. Contents

We make a brief overview of the related works in Section II.

We give some theoretical background concerning the concept of diversity and, in particular, modeling it by means of *dispersion* in Section III. In this section we also present diverGene, our diversity-aware selection operator, and explain that exact computation of our operator is a hard computational optimisation problem itself by linking it to a known NP-hard optimisation problem.

Because of this, in Section IV, we explain how our selection operator can be efficiently computed with a known poly-time approximation algorithm.

Section V contains the specifications of the used benchmark optimisation problems and describes the some technical representation and implementation details.

In Section VI we report experimental results that are discussed in Section VII. We conclude in Section VIII.

## II. Related Work

Genetic algorithm belongs to the class of evolutionary algorithms. First mentions of algorithm based on the mechanics of biological evolution dates back to the half of the previous century, but it became popular through the work of John Holland [1].

The concept of Diversity plays an important role in complex systems [2]. In particular, the diversity of population in biology is an important mechanism that supports better exploration of the environment. Therefore, the concept of diversity-awareness of population in biology is an inspiration for the domain of genetic algorithms to avoid a premature convergence (exploatation).

The Max-Sum Facility Dispersion Problem, that inspired our choice of a specific approach to diversify the population was studied in many papers in the domain of Operational Research. An example of such work is [3] that also discusses its computational hardness and efficient approximation algorithms for this problem. The connection of this problem to Web search result diversification with a new, parameterised objective function that takes into account a balance between the total value and pair-wise dispersion is discussed in [4] which also inspired our approach presented in this paper. Similar approach to diversification was proposed in the problem of semantic entity summarisation in [6] where another nature-inspired approach is applied to optimise the diversification problem itself. The study of the influence of the properties of the pair-wise dissimilarity function on the approximation factor in Facility Dispersion Problem is recently studied in [7].

The Travelling Thief Problem was introduced in [8] as a benchmark for testing heuristic algorithms on a problem that better resembles real life problems, that are often way more complex than well known “academic” benchmark problems like TSP or Knapsack. By its definition, the Travelling Thief Problem, is a kind of combination of the Knapsack Problem and TSP Problem.

## III. diverGene: Diversity-aware Selection Operator

In this section we introduce diverGene – a modified selection operator that makes it possible to control the desired level of diversity in the population in genetic algorithm. In the classical genetic algorithm, the selection operator is based mainly on the fitness function of an individual. Our diversity-aware selection operator takes the fitness score into account, but additionally aims at guaranteeing some level of the diversity of solutions in the selected population.

More precisely, *diversity-aware selection* can be viewed as an optimisation problem itself and can be defined as to select a sub-population $S$ of cardinality $k$ out of the whole current population that maximizes the following parameterised objective function:

$$f_d(S) = (1-\alpha)\sum_{p\in S} f(p) + \alpha \sum_{p,q\in S} dissimilarity(p,q) \quad (1)$$

Where $S$ is the selected, diverse sub-population and $\alpha$ is a parameter controlling the balance between the total $fitness$ (the $f()$ function) of the selected individuals and pair-wise $dissimilarity$ between them. $dissimilarity$ is a function returning a numerical value showing how different two individuals are. It's exact definition depends on the problem under consideration.

### A. A Link to Max-Sum Facility Dispersion Problem

Now we are going to explain that maximising the objective function 1 used in the definition of our population diversity operator at the beginning of Section III is equivalent to some well-known optimisation problem that is NP-hard but for which there exists a fast approximation algorithm. Due to this we use this approximation algorithm for efficiently compute the population diversity operator.

More precisely, the diversity selection operator described in Equation 1 is inspired by the Facility Dispersion Problem [3] and its recent applications in Web Search Diversification Problem [4].

In Facility Dispersion Problem, the input consists of a complete, undirected graph $G(V,E)$, an edge-weight function $d : V^2 \rightarrow R^+ \cup \{0\}$ that represents *pairwise distance* between the vertices and a positive natural number $k \leq |V|$. The task is to select a $k$-element subset $S \subseteq V$ that maximises the objective function $dispersion(S)$ that represents the notion of *dispersion* of the elements of the selected set $S$.

Facility Dispersion Problem was studied in operational research for modeling the problem of selecting a set of mutually-distant (dispersed) locations for $k$ obnoxious facilities like nuclear plants, ammunition dumps, etc.

In the most common variant of this problem, called Max-Sum Dispersion Problem, the $dispersion$ function to be maximised is defined as the total pair-wise distance between the selected locations (to be maximised):

$$dispersion(S) = \sum_{\{u,v\}\subseteq S} d(u,v) \quad (2)$$

The Max Sum Dispersion problem is NP-hard even if the distance function $d$ is a metric, but in such case there exists a polynomial-time algorithm of approximation factor of 2 that was presented in [3].

Recently, the Max-Sum Facility Dispersion problem has new applications in the *Web Search Result Diversification Problem* where the selected items represent documents (or pieces of information) to be returned by the search system and $d$ models the pairwise *dissimilarity* between the documents.

Given a set $V$ of documents to be potentially relevant to a user query, a number $p \in N^+, k < |V|$, a *document relevance* function $w : V \rightarrow R^+$ and pairwise *document dissimilarity* function $d : V^2 \rightarrow R^+ \cup \{0\}$, the task is to select a subset $S \subseteq V$ that maximises the value of a properly defined *diversity-aware relevance function*. In [4] the following parameterised, bi-criteria objective function (to be maximised) is proposed as the diversity-aware relevance function:

$$f_{div-sum}(\lambda, S) = (k-1)\sum_{v\in S} w(v) + 2\lambda \sum_{\{u,v\}\subseteq S} d(u,v) \quad (3)$$

where $\lambda \in R^+ \cup \{0\}$ is a parameter that controls the diversity/relevance-balance.

In the same work it is observed that a proper modification of $d$ to $d'$ (Equation 4) makes the described problem of maximising $f_{div-sum}(\lambda, S)$ equivalent to maximising $\sum_{\{u,v\}\subseteq S} d'_\lambda(u,v)$, where:

$$d'_\lambda(u,v) = w(u) + w(v) + 2\lambda d(u,v) \quad (4)$$

Thus, it makes the *result diversification* problem described above equivalent to the Max Sum Dispersion problem for $d'_\lambda$.

Now, it is not hard to see that the problem of maximising our dispersion-aware fitness function $f$ (for any value of $\alpha$) defined in Equation 1 is equivalent to the problem of maximising the objective function in Equation 3 for appropriate selection of the value of $\lambda$, thus the problems are equivalent.

To sum up, as the consequence, the problem of maximising the diversity-aware population fitness function defined in Equation 1 is NP-hard, as being equivalent to Max-Sum Dispersion Problem, but any approximation algorithm used for Max-Sum Dispersion can be used to optimise our problem.

Due to this, in the experiments reported in this paper we use the 2-factor greedy approximation algorithm for Max-Sum Facility Dispersion problem described in [3] to efficiently solve our problem. This is described in Section IV.

## IV. Dissimilarity selection

Our efficient implementation of diversity-aware selection operator is based on the algorithm mentioned at the end of the previous section. This approximation algorithm for finding subsets (in which weights of edges between those vertices are maximised) of vertices in a graph, can easily be adapted for choosing a subset of individuals in a population that are most different from each other. One just needs to think of individuals as vertices and the level of dissimilarity between two individuals as the weight value on edges between the vertices representing them.

In our case we wanted to take into account both dissimilarity and fitness of the two individuals, as can be seen in the function defined in Equation 1. We utilised the link to the previously studied problems of Max-Sum Facility Dispersion and Web search diversification described in the previous section to define a $dissimilarity\_fitness$ function that combines dissimilarity between two individuals and their respective fitness values:

$$df(u, v) = fitness(u) + fitness(v) + 2\lambda dissimilarity(u, v)$$

We used the function to calculate values of edges between the individuals on the graph.

More precisely, the algorithm that we utilize to choose a $k$-element subset of individuals from the previous population to control its guaranteed level of diversity works in the following way:

1) Create a list with all possible pairs of individuals in a population.
2) Choose the pair with the highest $dissimilarity\_fitness$ score.
3) Remove all pairs containing one of the individuals from the chosen pair.

Repeat steps 2–3 $k/2$ times, where $k$ is the desired size of the selected population. If $k$ is odd, add an arbitrary final individual to the selected sub-population. This greedy algorithm guarantees the approximation factor of 2 to the solution of such defined optimisation problem [3].

### A. *Combining two types of selection*

We create the new population of size $N$ by combining the results of two kinds of selection - classical *roulette selection* and our own *dissimilarity selection* described in Section IV. First we select $k$ individuals using dissimilarity selection and then achieve the desired population size $N$ by selecting the remaining $N - k$ of the individuals using standard roulette selection. The result for small values of $k$ can be viewed as classic selection enriched by a small pool of individuals retained for their uniqueness (in combination with their fitness, since $dissimilarity\_selection$ takes both into account).

## V. Experimental Implementation

We used the Python programming language to create a flexible framework to test our novel diversity-aware operator on various optimisation problems.

In this Section we report experiments performed on two classic optimisation problems - Knapsack Problem and Travelling Salesman Problem. We also report additional series of experiments concerning the combination of the latter two – the Travelling Thief Problem. The framework is designed so that in the continuation work, the plugins with support for other optimisation problems can be easily added to it.

### A. *Knapsack Problem*

The Knapsack Problem is one of the a classic NP-hard optimization problems. In this problem, there are $n$ items. Each item has a value ($p_i \in Q^+$) and a weight ($w_i \in Q^+$). The capacity ($W \in Q+$) of the knapsack is limited.

The items should be picked so that their total value is maximized while their total weight does not exceed the knapsack capacity:

$$f(\bar{x}) = \sum_{i=1}^{n} p_i x_i, while \sum_{i=1}^{n} w_i x_i \leq W$$

where $\bar{x} = (x_1, x_2, \ldots, x_n)$ and $x_i \in \{0, 1\}$ indicates whether the item is picked ($x_i = 1$) or not ($x_i = 0$).

*1) Representing the Knapsack Problem in GA Framework:* A solution is represented as a set of items in a knapsack. The most natural representation of any solution as a chromosome is in the form of a characteristic vector of the given subset of the items.

The mutation operator "flips" bits at random positions of the corresponding characteristic vector. Each position has an equal probability of being flipped (in our experiments the probability always equals $p_{km} = 0.01$). If the item was previously in the set it will be removed, otherwise it will be added to the set.

The cross-over operator exchanges the state between random items between two solutions. Each position has an equal probability of being exchanged (in our experiments the probability always equals $p_{kc} = 0.05$).

Both operators may result in a solution that exceeds the allowed capacity. In such case we fix the solution by removing random items from the set until the capacity is abided by.

The dissimilarity score is calculated by counting the number of items present in one solution but not present in the other (the *Hamming distance*).

*2) Knapsack Dataset:* The dataset for Knapsack problem was generated purposely for our experiments. Each instance contains 32 items with different values and weights (both ranging between 1 and 500). The knapsack capacity is set to 500.

### B. *Travelling Salesman Problem*

The Travelling Salesman Problem (TSP) is also a classic NP-hard optimization problem. In this paper, we consider a 2-dimensional euclidean variant. In this variant of TSP, there are $n$ cities ($c$) with their coordinates ($cx$ and $cy$). A salesman must visit each city exactly once and minimize the total length of the complete tour. The aim is to find a permutaion of the set containing all cities which minimizes the following equation:

$$f(\bar{c}) = \sum_{i=1}^{n-1} d(c_i, c_{i+1}) + d(c_n, c_1)$$

where $\bar{c} = (c_1, c_2, \ldots, c_n)$ represents the tour and $d(A, B)$ is a distance between the cities $A$ and $B$:

$$d(A, B) = \sqrt{(Ax - Bx)^2 + (Ay - By)^2}$$

*1) Representing TSP in GA Framework:* The *ordered crossover* and *reverse sequence mutation* operators were used for TSP. The ordered crossover operator, presented by Goldberg in [5], is a two-point crossover. Given two random crossover points parent chromosomes are split into three parts. Child chromosome inherits the left and right parts from the first parent, and the middle part from the second one. The elemetns in the right and left parts are removed and rotated to avoid the duplication of elements [5].

In the reverse sequence mutation operator the sequence of elements between two randomly chosen positions is reversed.

The dissimilarity operator takes into consideration the number of different elements on the same positions in both chromosomes and compares it to the total number of elements. For each position, the values from both individuals are compared. If they differ, the counter is incremented. After analyzing the whole chromosomes the counter value is divided by the size of a chromosome (number of all cities in the instance).

*2) TSP Dataset:* The dataset for TSP experiments was taken from the TSPLIB library [9] containing sample instances for TSP. Berlin52 set was chosen [9] for the experiments. This set provides coordinates of 52 locations. The optimal solution to this problem is known, and it has the value of 7542.

### C. Travelling Thief Problem

The *Travelling Thief Problem* is an optimisation problem related to the Knapsack Problem and the Travelling Salesman Problem (both described above). It was created to better model real-life problems, often more complex than the well-known benchmark NP-hard problems.

The Travelling Thief Problem consists of two interdependent sub-problems. The sub-problems can not be solved individually, because the solution to one sub-problem strongly influences the quality of the solution of the second sub-problem. To achieve satisfactory results one needs to solve both sub-problems at the same time, finding the best *combination* of solutions. This resembles the challenge often occurring in the case of solving complex real-life problems with multiple, often mutually contradictory, constraints.

We decided to include the Travelling Thief Problem in our experiments as it seems a natural next step of our work since it can be viewed as a specific *composition* of the two problems described above: the Knapsack Problem and TSP. Thus, the two components of the specification introduce a multi-criteria structure. Additional complexity of this problem is introduced by modeling the varying *speed*, that depends on the current contents of the knapsack, that influences the total cost of the solution, as explained below in this Section. Such a complex, multi-criteria optimisation specification seems to be a very natural domain of application for our diversity-aware population control approach as it enforces exploring multpile, mutually dissimilar alternatives for the solution.

There are several variants of the TTP problem. The one implemented by us, known as $TTP_1$, is most likely the most popular one. There are $n$ cities (represented by a distance matrix) and $m$ items (each having its own value, weight and set of cities in which it can be found). A thief has to visit each city exactly once, picking some items located in those cities and putting them in their knapsack (which has a maximum weight capacity $W$). Thief's speed depends on the weight of the knapsack:

$$V_C = V_{max} - W_C \frac{V_{max} - V_{min}}{W}$$

$V_C$ represents the current speed, $W_C$ represents the current weight of the knapsack and $W$ represents the knapsack's maximum weight capacity. $V_{max}$ and $V_{min}$ are parameters describing the thief's maximum and minimum speed, respectively. The natural interpretation is as follows: the heavier the knapsack, the slower the thief. The speed is important, because the objective function that is to be maximized depends not only on the value of the picked items, but also on the total time of the travel, as the thief has to pay for the time when the knapsack is being used:

$$G(x, z) = g(z) - R \cdot t(x, z)$$

Where $g$ is the sum of values of all items in the knapsack, $R$ is the knapsack payment fee (rent) for the time unit and $t$ is the function calculating the time of the travel for the given solution. This means the thief needs to maximise their profit - the value of the stolen items reduced by the knapsack fee.

A solution consists of a tour $x$ and item picking plan $z$ (recording the information concerning which items should be picked in which city - if it is picked at all). The $G$ function strongly interconnects them, so they can not be optimised separately.

*1) Representing TTP in GA Framework:* TTP representation in GA framework is a union of previously described Knapsack and TSP representations. The dissimilarity value is calculated by counting the number of different picks - the items picked in the different cities at different times.

*2) TTP Dataset:* We used the dataset for TTP experiments taken from the webpage of the optimisation group of the University of Adelaide [10]. The data provided on this site was used as a benchmark set for competitions in 2014 and 2015. More precisely, the `Berlin52_n51_bounded-strongly-corr_01` dataset was chosen [10]. This set provides coordinates of 52 locations and 51 items, one in each city. The optimum solution to this instance is not provided by the creators of the dataset nor yet known, to the best knowledge of the authors.

## VI. Experimental Results

In the experiments we measured the influence of the dissimilarity operator on the best solution of the objective function $f(S)$ for diffrent population sizes. The tests where performed in three groups for each problem:

1) The best solution of $f(S)$ as a function of $k$ for 10, 25, 50 and 100 individuals in population with $\alpha$ set to 0.35 (Fig. 1, 2 & 3)

2) The best solution of $f(S)$ as a function of $\alpha$ for 10, 25, 50, 100 and 150 individuals in population with $k$ set to 20% of population (Fig. 4, 5 & 6)
3) The best solution of $f(S)$ as a function of population size with $\alpha$ set to 0.35 and $k$ set to 20% of population (Fig. 7, 8 & 9)

Each test was run 10 times and the median of all the 10 results was computed and reported.

### A. Computational Platform

Since the computations involved in our experiments were quite extensive it was more natural to perform them on a specialised computational platform. Due to this, all calculations were performed on the online data science platform Sense [11]. Sense is a recently launched (released March 18, 2015) cloud platform for data science and big data analytics built by Sense, Inc. company. It provides an interface to multiple scripted analytic tools like Python, Node.js, SQL and many more, and allows to launch a new engine with dedicated CPU and RAM for every created job. It also makes sharing data, scripts and jobs and colaborating on the same project easy.

## VII. Discussion of the Results

### A. The Best Solution of $f(S)$ as a Function of $k$

For the Knapsack problem some improvements in quality of solutions can be observed only for small populations (Fig. 1). For 10 and 25 individuals using the diversity-aware selection operator increased the value of the best solution by 12-15%. For bigger populations (50 and 100 individuals), though, the positive changes are almost unnoticeable. A critical value of $k$ can be also observed for all populations. Exceeding $k = 0.7N$, where $N$ is the population size, causes a dramatic decrease of the best solution value. This could be interpreted as too much exploatation at the expense of exploration for this relatively easy[1] problem would deteriorate the performance.

More significant positive impact of diversity operator on the quality of solution can be observed for TSP (Fig. 2). For all population sizes the results were improved by 30-35% when 15-35% of individuals were selected using the diversity-aware selection operator. The best solution decrease starts earlier than in the Knapsack problem but is not so rapid as in the case of Knapsack - in the worst case the percentage decrease between the best value of the objective function without and with diversity-aware selection operator is about 8% while in the Knapsack problem it is between 30-40%.

In the TTP experiments (Fig. 3) even stronger positive impact of our diversity operator on the solution quality can be observed. It reaches up to 40-45%. Importantly and interestingly, the peak representing the best solutions is shift towards the case when more individuals are being selected using our diversity-aware selection operator (about 70-75%). After reaching $k = 1N$ the collected results are the same or slightly worse then for computations without diversity-aware selection operator. That seems to indicate that for this, particularly hard and complex, optimisation problem the population diversity can only improve the performance of the algorithm while it almost *never* makes the quality worse.

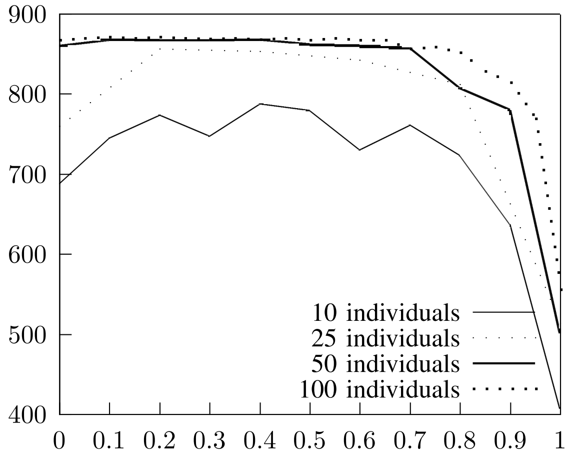


Fig. 1. $f(S)$ as a function of $k$ for Knapsack problem.

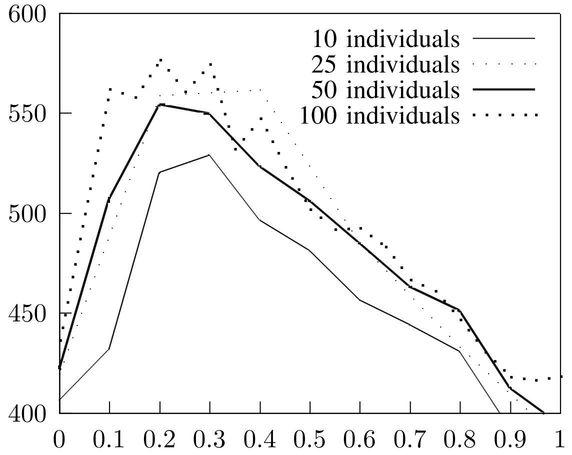


Fig. 2. $f(S)$ as a function of $k$ for TSP.

### B. The Best Solution of $f(S)$ as a Function of $\alpha$

Experiments performed for the Knapsack problem (Fig. 4) show that changes of $\alpha$ parameter have a positive impact on the best solution value only for the smallest tested population composed of 10 individuals (28% increase for $\alpha = 0.15$). For bigger populations adding $dissimilarity$ to $f(S)$ has no or only negative impact on the results.
For our current experimental settings, the TSP (Fig. 5) and TTP (Fig. 6) tests show no clear dependence between the results and changes of $\alpha$. This is an interesting signal that seems to be counter-intuitive (compared to the results concerning the dependence on the value of $k$ presented above) and should be

[1] Remind that Knapsack is a relatively "easy" computational problem among the NP-hard problems as there exists a pseudo-polynomial dynamic programming algorithm for it, what excludes it from the class of, so called, strongly NP-hard problems

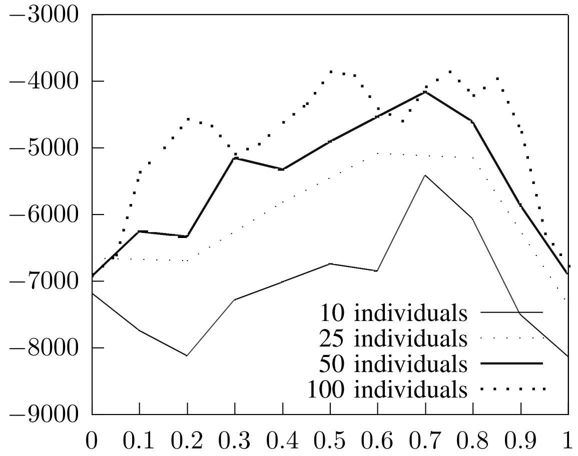


Fig. 3. $f(S)$ as a function of $k$ for TTP.

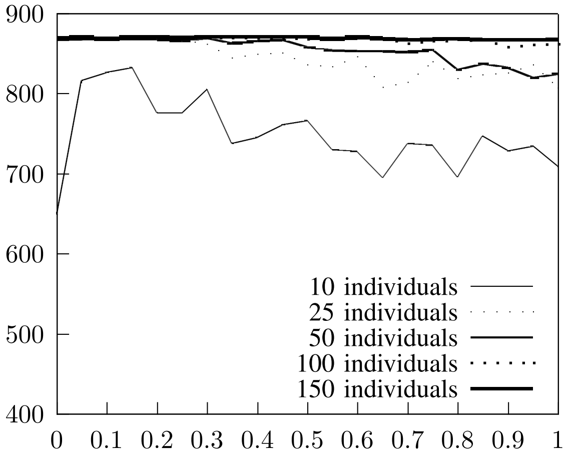


Fig. 4. $f(S)$ as a function of $\alpha$ for Knapsack problem.

further examined in the continuation work.

### C. The Best Solution of $f(S)$ as a Function of Population Size

Analysis of the last group of tests confirms the results observed in the first series of experiments (concerning the dependence of the solution quality on $k$). Namely, improvements in Knapsack problem solution (Fig. 7) are visible only for small populations and there is no change for bigger populations. For TSP (Fig. 8) a positive impact can be observed for all tested population sizes (10-200 individuals). The best solutions are equally improved throughout all the tested populations.

TTP results (Fig. 8) show, that for smaller populations there is a small postitve or even in some cases a negative impact on the objective function value. For bigger populations a positive impact is visible. It increases slightly with population size.

Most importantly, the reported experimental results clearly indicate that introducing diversity to the solution population can noticeably increase the quality of the solutions.

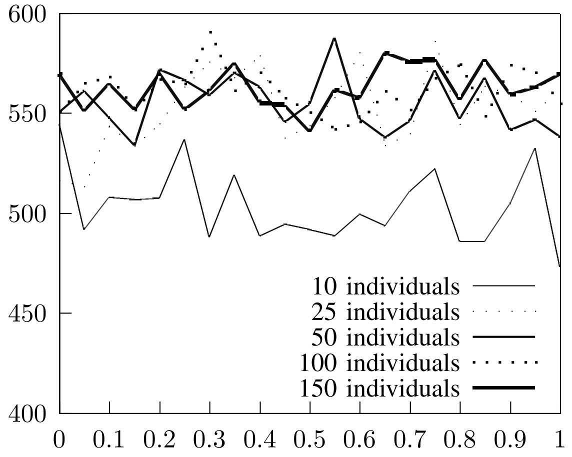


Fig. 5. $f(S)$ as a function of $\alpha$ for TSP.

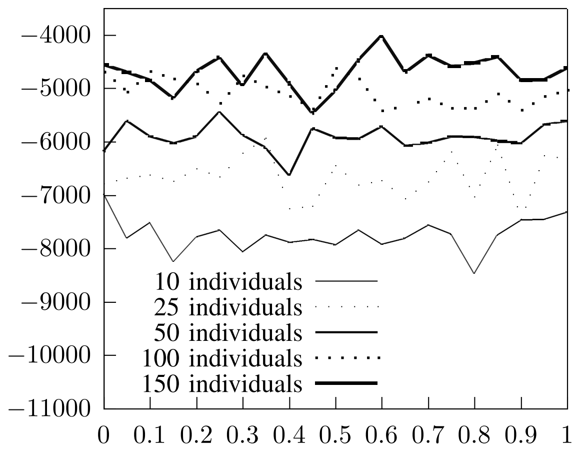


Fig. 6. $f(S)$ as a function of $\alpha$ for TTP.

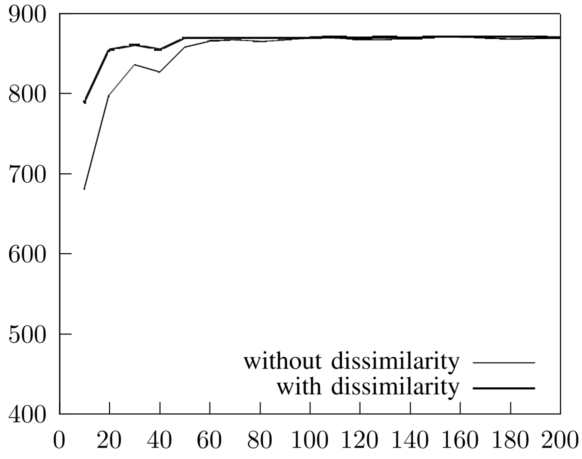


Fig. 7. $f(S)$ as a function of $|S|$ for Knapsack problem.

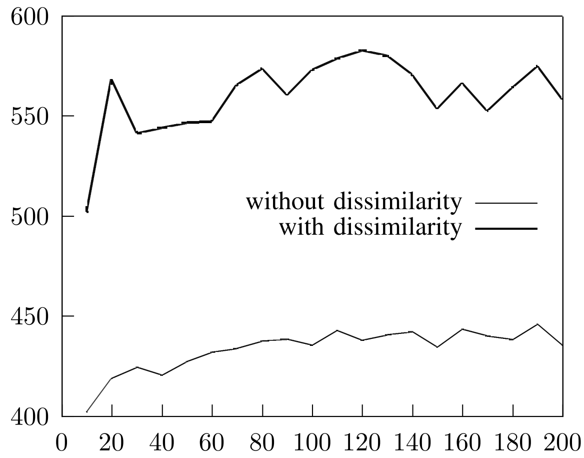


Fig. 8. $f(S)$ as a function of $|S|$ for TSP.

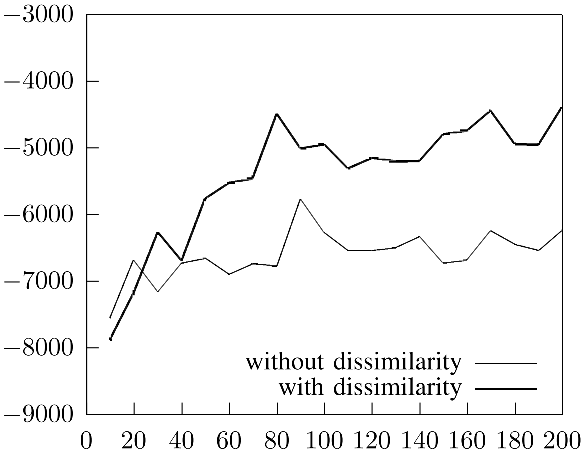


Fig. 9. $f(S)$ as a function of $|S|$ for TTP.

## VIII. Conclusions and Future Work

The experiments reported in this paper clearly indicate that controlling the population diversity in genetic algorithm in the way we proposed in this paper by using `diverGene` is a promising technique. In particular, picking a part of population using the diversity-aware selection operator can significantly improve results of the objective function $f(S)$ for the variant of the TSP problem that we studied and, even more significantly, for the TTP problem. There is also an interesting observation, that seems to be natural and intuitive, that the more complex and harder the optimisation problem the more improvement can be achieved by applying our diversity-aware selection operator.

More tests should be performed for the studied problems, including examining other implementations of dissimilarity operator and other settings.

Another interesting direction could be to study the convergence rate as the function of the population diversity.

It is a bit surprising that we did not observe much dependence of the quality of the solution on the value of the $\alpha$ parameter. This should be further investigated in the future work.

It would be also interesting to introduce the self-adaptation mechanism to our approach. More precisely, to make the algorithm itself dynamically controlling the values of the diversity-aware operator parameters (e.g. $k$ and $\alpha$) to better control the exploration/exploatation balance.

## IX. Acknowledgements

This work is partially supported by Polish Ministry of Science and Higher Education grant N N519 578038.

## References

[1] John H. Holland *Adaptation in Natural and Artificial Systems.* University of Michigan Press,1975

[2] Scott E. Page. Diversity and complexity. *Primers in Complex Systems.* Princeton, NJ: Princeton UniversityPress. x, 291 p. $ 19.95, 2011

[3] R. Hassin, S.Rubinstein, A. Tamir, “Approximation algorithms for maximum dispersion”, *Operations research letters* vol. 21/3, 1997, pp. 133–137.

[4] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, WWW ’09, pages 381–390, New York, NY, USA, 2009. ACM.

[5] David E. Goldberg. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning* (1st ed.). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[6] W.Kosiński, T.Kuśmierczyk, P.Rembelski, M.Sydow "Application of Ant-Colony Optimisation to Compute Diversified Entity Summarisation on Semantic Knowledge Graphs", Proc. of International IEEE AAIA 2013/FedCSIS Conference, Annals of Computer Science and Information Systems, Volume 1, pp. 69-76, ISSN 2300-5963, ISBN 978-1-4673-4471-5 (Web), 2013

[7] M.Sydow "Approximation Guarantees for Max Sum and Max Min Facility Dispersion with Parameterised Triangle Inequality and Applications in Result Diversification" (extended journal version) Mathematica Applicanda Vol. 42, no. 2, pp. 241-257, DOI: 10.14708/ma.v42i0.547, Print ISSN: 1730-2668; On-line ISSN: 2299-4009, Polish Mathematical Society, 2014

[8] M. R. Bonyadi, Z. Michalewicz, L. Barone (2013, June). The travelling thief problem: the first step in the transition from theoretical problems to realistic problems. In Evolutionary Computation (CEC), 2013 IEEE Congress on (pp. 1037-1044). IEEE.

[9] http://www.iwr.uni-heidelberg.de/groups/comopt/software/ /TSPLIB95/tsp/berlin52.tsp.gz

[10] http://cs.adelaide.edu.au/ optlog/CEC2014COMP_InstancesNew/berlin52-ttp.rar

[11] http://wwww.sense.io

[12] Mohammad Reza Bonyadi, Zbigniew Michalewicz and Luigi Barone The travelling thief problem: The first step in the transition from theoretical problems to realistic problems. *IEEE Congress on Evolutionary Computation.* IEEE, pp. 1037–1044, 2013

DOI: 10.15439/2015F174
ACSIS, Vol. 5

# New similarity index based on the aggregation of membership functions through OWA operator

Amine AÏT YOUNES, Frédéric BLANCHARD, Michel HERBIN
Université de Reims Champagne Ardenne, France
CReSTIC,
Email: {amine.ait-younes, frederic.blanchard, michel.herbin}@univ-reims.fr

***Abstract*—In the field of data analysis, the use of metrics is a classical way to assess pairwise similarity. Unfortunately the popular distances are often inoperative because of the noise, the multidimensionality and the heterogeneous nature of data. These drawbacks lead us to propose a similarity index based on fuzzy set theory. Each object of the dataset is described with the vector of its fuzzy attributes. Thanks to aggregation operators, the object is fuzzified by using the fuzzy attributes. Thus each object becomes a fuzzy subset within the dataset. The similarity of a reference object compared to another one is assessed through the membership function of the fuzzified reference object and an aggregation method using OWA operator.**

## I. Introduction

ASSESSING the similarity between samples is a key of success in data analysis process. Many methods rely on similarity indices. Most of clustering ones uses pairwise comparisons when aggregating or separating samples [11], [12]. In the framework of case-based reasoning, solving problem needs for searching similar cases and assessing their similarities [10]. Recommender systems also deal with similarity between objects [18]. Thus the search for pairwise similarity indices remains an active field of research [2], [14]. The choice of similarity measures depends on the representation of objects we compare [5] [6]. In this paper, we restrict the scope of this study to the comparison of vector data.

When data is described with multidimensional vectors, the use of metrics remains the classical way to assess pairwise similarity [19], [6]. Unfortunately, database objects could have qualitative features making difficult to obtain standardized quantitative vectors of attributes from the objects. Thus the popular distances (Euclidean distance, Mahalanobis distance, Minkowski metric, Cosine distance, Correlation distance,... ) become often inoperative. Moreover noise or vagueness can corrupt data and the curse of dimensionality is also an obstacle in processing queries in high-dimensional space [3]. These drawbacks lead us to propose a similarity index which is not based on a distance function or a metric in the data space.

To overcome these difficulties, the fuzzy set theory gives a framework to design similarity indices [17], [20]. In this paper, the dataset forms the context for our pairwise comparisons between objects.

Each object $X_i$ of the dataset is fuzzified. Let $\tilde{X}_i$ be the fuzzy set obtained from the crisp object $X_i$, $\tilde{X}_i$ is the fuzzy version of $X_i$. The membership degree of the crisp object $X_j$ to the fuzzy set $\tilde{X}_i$ is considered as the similarity value from $X_j$ to the reference object $X_i$. Therefore the similarity indices we propose are only fuzzy membership functions. Note that such similarity indices based on membership functions do not necessarily define symmetric relations.

The challenge using our approach becomes to obtain a fuzzification of an object $X_i$ within the dataset [13]. To achieve this goal, the attributes of the object $X_i$ are considered as fuzzy numbers or fuzzy quantities. Then each crisp object $X_j$ is described with a vector of membership degrees relative to the fuzzy attributes of $X_i$. Thanks to fuzzy logic operators, we aggregate these membership degrees to obtain the aggregated membership value of $X_j$ to the fuzzy set $\tilde{X}_i$. The critical issue of this approach is the aggregation method we use. This communication proposes to adapt OWA operators [16] to define our aggregation method.

The paper is organized as follows.

In The sections 2 and 3 we present our approach for the fuzzification of the attributes. The section 4 exposes the methodology we use to evaluate the sensibility and specificity of each attributes. After the fuzzification of the data, we present in the section 5 the aggregation procedure used to build a new similirity index using an Ordered Weighted Aggregation operator. Before concluding, we present in the section 6 a comparison of our new pairwise similarity indice with the popular metrics.

## II. Domain of fuzzification

Let $E$ be a set of $n$ objects defined by:

$$E = \{X_i \ / \ 1 \leq i \leq n\} \tag{1}$$

where $X_i$ are the $n$ objects of $E$.

Each object is described by a vector of $p$ attributes. The object $X$ is represented by the $p$-tuple $(x_{ik})$ with $1 \leq k \leq p$ where $x_{ik}$ is the value of the $k$-th attribute of the object $X_i$. These $p$ attributes are either quantitative or qualitative. If the $k$-th attribute is quantitative, then its values lie within an interval $[a_k, b_k]$ of $\Re$. If the $k$-th attribute is qualitative, then its values are within a set $\{v_1, v_2, v_3, ...v_l\}$ of $l$ values. In both cases, we call $D_k$ the set we use to define the $k$-th attribute.

The domain of definition $D$ of $E$ is defined by:

978-83-60810-66-8/$25.00©2015, IEEE

$$D = \prod_{1 \leq k \leq p} D_k \tag{2}$$

Then we have: $E \subset D$ with: $\#E = n$. In the following each object becomes a fuzzy subset of $E$ relatively to its attributes. Thus $E$ is called domain of fuzzification.

## III. Fuzzification of the attributes

This section is devoted to the fuzzification of an object $X_i$ within the dataset $E$. Although the fuzzification of an attribute is itself beyond the scope of this document, we firstly describe the way we used to fuzzify each attribute value of the object $X_i$. Thus we obtain $k$ fuzzy attributes for $X_i$. Then we merge these fuzzy attributes to build the fuzzy object $\tilde{X}_i$ defined in $E$.

Let $X_i$ be an arbitrary reference object of the data set $E$. Let $x_{ik}$ be the value of the $k$-th attribute of $X_i$. The values of attributes are often imprecise and the meaning could be vague. Therefore it is convenient to represent such imprecise or vague values by fuzzy sets. Thus $x_{ik}$ is represented by a fuzzy subset of $D_k$. The membership function $m_k^i$ of this fuzzy subset is defined by:

$$\begin{array}{llll} m_k^i \colon & D_k & \longrightarrow & [0,1] \\ & x & \longmapsto & m_k^i(x) \end{array} \tag{3}$$

In this paper, these fuzzy sets are normalized with $m_k^i(x_{ik}) \leq 1$.

In this paper, we propose a simple and empirical approach of the data fuzzification. Each numeric value is represented by a conventional trapezoidal membership function defined by $(a, b, c, d)$ with (cf. fig. 1):

$$m_k^i(x) = \begin{cases} 0 & \text{if } x < a_i \\ \frac{x - a_i}{b_i - a_i} & \text{if } a_i \leq x < b_i \\ 1 & \text{if } b_i \leq x < c_i \\ \frac{d_i - x}{d_i - c_i} & \text{if } c_i \leq x < d_i \\ 0 & \text{if } d_i \leq x \end{cases} \tag{4}$$

Let $\overline{x}_k$ and $\sigma_k$ be respectively the mean and the standard deviation of the $k$-th attribute within $D_k$. If $dev_k^i$ is the deviation between $x_{ik}$ and $\overline{x}_k$ (i.e. $dev_k^i = |x_{ik} - \overline{x}_k|$), an empirical study leads us to propose:

$$\begin{cases} a_i = x_{ik} - \sigma_k - 0.5\, dev_k^i \\ b_i = x_{ik} - 0.5\, \sigma_k - 0.1\, dev_k^i \\ c_i = x_{ik} + 0.5\, \sigma_k + 0.1\, dev_k^i \\ d_i = x_{ik} + \sigma_k + 0.5\, dev_k^i \end{cases} \tag{5}$$

If the $k$-th attribute is qualitative, $x_{ik}$ is fuzzified using a degree of membership for each possible value of the attribute. Then $m_k^i$ is defined by $l$ values $(m_k^i(v_1), m_k^i(v_2), ...m_k^i(v_l))$ within $D_k$.

This paper proposes to use $m_k^i$ to fuzzify the object $X_i$ within $E$ in respect with its $k$-th attribute. The membership function of the object $X_i$ is defined by:

$$\begin{array}{llll} \mu_k^i \colon & E & \longrightarrow & [0,1] \\ & X_j & \longmapsto & \mu_k^i(X_j) = m_k^i(x_{jk}) = \mu_{jk}^i \end{array} \tag{6}$$

with $1 \leq j \leq n$ and $1 \leq k \leq p$.

If the value of $x_{ik}$ is not set, then we propose to define $\mu_k^i$ by simply $\mu_k^i(X_j) = \frac{1}{2}$ in order to ensure the robustness of the proposed approach.

At this stage of the communication, several points should be noted. Each object $X_i$ gives rise to $p$ fuzzy subsets of $E$.

Each subset is associated with an attribute. These fuzzy subsets are defined with reference to the object $X_i$. They are normalized because $\mu_{jk}^i \leq 1$.

We propose to consider the membership degrees $\mu_{jk}^i$ (with $X_j \in E$) as similarity values from $X_j$ to the reference $X_i$ with respect to the $k$-th attribute. If $\mu_{jk}^i = 1$, then $X_j$ and $X_i$ are considered as similar with respect to the $k$-th attribute. In contrast, if $\mu_{jk}^i = 0$, then $X_j$ and $X_i$ are considered as dissimilar with respect to the attribute. The more $\mu_{jk}^i$ is close to 1, the larger the similarity from $X_j$ to $X_i$ for the $k$-th attribute. Thus the membership function $\mu_{ik}$ (with $1 \leq k \leq p$) is considered as a similarity index to $X_i$ with respect to its attribute $x_{ik}$.

We can see in "fig. 1" that $x_{j_3k}$ is considered as similar to $x_{ik}$ but $x_{j_1k}$ is not comparable to $x_{ik}$. This similarity value is asymmetric. In "fig. 2" $x_{j_4k}$ is not comparable to $x_{ik}$ : $\mu_k^i(x_{j_4k}) = 0$ but $\mu_k^{j_4}(x_{ik}) > 0$.

The membership functions $\mu_{ik}$ give $p$ indices of similarity to $X_i$ within the set $E$. Let us define two characteristics of these indices that we call the sensibility and the specificity to the similarity with $X$.

Let $sens_{ik}$ be the mean of $\mu_{jk}^i$ when $X_j \in E$:

$$sens_{ik} = \frac{1}{n} \sum_{X_j \in E} \mu_{jk}^i \tag{7}$$

The value $sens_{ik}$ lies between 0 and 1. It assesses an average similarity between the reference object $X_i$ and the whole dataset $E$ in respect with the $k$-th attribute.

If $sens_{ik}$ is close to 1, then the $n$ similarity values $\mu_{jk}^i$ are also rather close to 1. Then the $n$ objects $X_j$ of $E$ are rather similar to $X_i$. In this case, the values $\mu_{jk}^i$ are sensitive indicators of the similarity to $X_i$. Since these values are rather equal to 1 or close to 1, then the value $\mu_{jk}^i$ becomes highly symptomatic of a dissimilarity (non-similarity) with $X_i$ when the indicator of similarity $\mu_{jk}^i$ is close to 0. Thus the $k$-th attribute is considered as an attribute sensitive to the similarity with $X_i$.

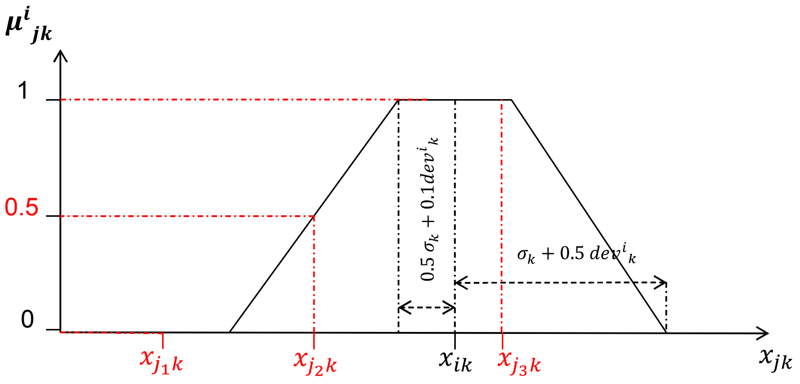


Fig. 1. Fuzzy representation.

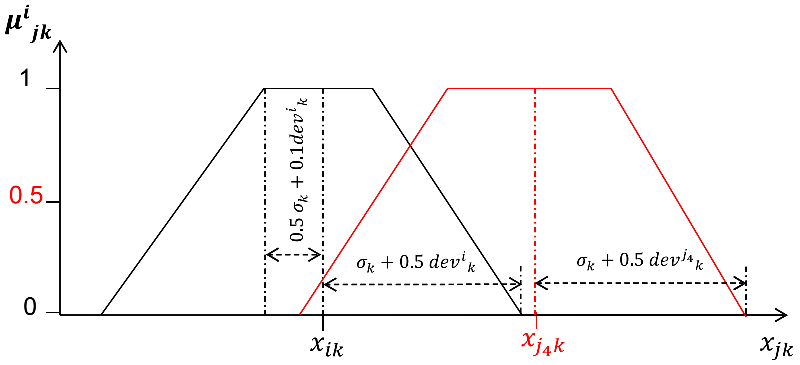

$\mu^i{}_{jk}$
1
0.5
0
$0.5\ \sigma_k + 0.1dev^i{}_k$
$\sigma_k + 0.5\ dev^i{}_k$
$\sigma_k + 0.5\ dev^{j_4}{}_k$
$x_{ik}$
$x_{j_4k}$
$x_{jk}$

Fig. 2. Asymmetric similarity values.

In contrast, if $sens_{ik}$ is close to 0, the $n$ similarity values $\mu^i_{jk}$ are also rather close to 0. Then the $n$ objects $X_j$ are rather dissimilar (non-similar) to $X_i$. In this case, the values $\mu^i_{jk}$ are specific indicators of the similarity to $X_i$. Since these values are rather equal to 0 or close to 0, then the value $\mu^i_{jk}$ becomes highly symptomatic of a similarity to $X_i$ when the indicator of similarity $\mu^i_{jk}$ is close to 1. Thus the $k$-th attribute is considered as an attribute specific of the similarity with $X_i$.

When we consider the $k$-th attribute, $sens_{ik}$ is a coefficient of the sensibility of this attribute to the similarity with $X_i$ and 1-$sens_{ik}$ is a coefficient of the specificity of the attribute for the similarity with $X_i$. These two coefficients characterize the $k$-th attribute with reference to the object $X_i$ within $E$.

Let us consider an example (see Table I) to explain these coefficients. The dataset $E$ has six objects $X_1$, $X_2$, $X_3$, $X_4$, $X_5$ and $X_6$. Each object is described with four attributes. The reference object is $X_1$. The four attribute values of $X_1$ are fuzzified. The four membership functions $\mu^1_1$, $\mu^1_2$, $\mu^1_3$ and $\mu^1_4$ indicate the degrees of similarity to $X_1$. In this example, the sensitivities of the four attributes are respectively 0.767, 0.400, 0.583 and 0.680. The 1st attribute is the most sensitive one. Only $X_6$ has 1st attribute value dissimilar from the one of $X_1$. Thus the 1st attribute reveals the dissimilarity with $X_1$. The specificities of the four attributes are respectively 0.233, 0.600, 0.417 and 0.320. The 2nd attribute is the most specific one. Only $X_2$ has 2nd attribute value similar to the one of $X_1$. Thus the 2nd attribute reveals the similarity with $X_1$.

TABLE I

Sensitivity and Specificity of the Attributes in Respect with the Reference Object $X_1$: Example of 6 Objects with 4 Fuzzy Attributes, $\mu_k$ are the Degrees of Membership to $X_1$ Relative to the $k$-th Attribute with $1 \leq k \leq 4$

| Objects | Fuzzy attributes | | | |
|---|---|---|---|---|
| | $\mu^1_1$ | $\mu^1_2$ | $\mu^1_3$ | $\mu^1_4$ |
| $X_1$ | 1 | 1 | 1 | 1 |
| $X_2$ | 0.9 | 1 | 0.3 | 0.9 |
| $X_3$ | 0.9 | 0.1 | 0.4 | 0.7 |
| $X_4$ | 0.9 | 0.1 | 0.5 | 0.5 |
| $X_5$ | 0.9 | 0.1 | 0.6 | 0.3 |
| $X_6$ | 0 | 0.1 | 0.7 | 0.2 |
| sensitivity | 0.767 | 0.400 | 0.583 | 0.680 |
| specificity | 0.233 | 0.600 | 0.417 | 0.320 |

## IV. Aggregation with OWA Operators

Let us consider the reference object $X_i$ in $E$. The fuzzy subset defined by the membership function $\mu^i_{jk}$ depends on the value $x_{ik}$ of the $k$-th attribute of $X_i$. We propose to aggregate these $p$ fuzzy subsets taking into account all the attributes. The goal is to fuzzify the reference object $X_i$ within $E$ defining a new membership function $\mu^i$ fusing the functions $\mu^i_k$.

The aggregation operators give a classical way to merge the fuzzy subsets in $E$. Let $aggreg$ be an aggregation operator. The function $\mu^i$ is defined by:

$$\begin{array}{llll} \mu^i: & E & \longrightarrow & [0,1] \\ & Y & \longmapsto & \mu^i(X_j) = \underset{1\le k\le p}{aggreg}(\mu^i_k(X_j)) \end{array} \tag{8}$$

The aggregation operators are well studied in literature [7], [8]. The minimum is the reference operator to obtain a conjunction and the maximum is the one for a disjunction. The operators used in this paper are a tradeoff between the conjunction (AND) and the disjunction (OR).

If the similarity index $\mu^i_k$ is very sensitive ($sens_{ik}$ close to 1), then the similarity index $\mu^i_k$ should contribute to $\mu^i$ using a conjunction operator. Indeed, a conjunction seems desirable because significant information is obtained when $\mu^i_k$ is close to 0. In contrast, if the similarity index $\mu^i_k$ is very specific ($sens_{ik}$ close to 0), a disjunction operator seems preferable because significant information is obtained when $\mu^i_k$ is close to 1.

In this paper, the tradeoff between conjunction and disjunction is defined using an Ordered Weighted Aggregation (OWA operators proposed by R. Yager [16]).

Let us describe the function $\mu^i$ obtained when using such an OWA operator.

For an object $X_j$ in $E$, the membership degrees $\mu^i_k(X_j)$ are ordered by decreasing order. We obtain :

$$\mu^i_{(1)}(X_j) \ge \mu^i_{(2)}(X_j) \ge \mu^i_{(3)}(X_j) \ge ... \ge \mu^i_{(p)}(X_j)$$

The aggregation is defined by:

$$\mu^i(X_j) = \sum_{1\le k\le p} w_k \times \mu^i_{(k)}(X_j) \tag{9}$$

We denote $W$ the weighting vector :

$$W = [w_1, w_2, \dots, w_p] \tag{10}$$

with $\sum_{1\le k\le p}(w_k) = 1$ and $w_k \in [0,1]$

The weights are not associated with attributes but with their ordered positions. The challenge is to determine the weights.

The conjunction operator (i.e. the minimum) is obtained if :

$$W_* = [0, 0, \dots, 1] \tag{11}$$

The disjunction operator (i.e. the maximum) is obtained if :

$$W^* = [1, 0, \dots, 0] \tag{12}$$

The ordinary average is recovered if :

$$\bar{W} = \left[\frac{1}{p}, \frac{1}{p}, \dots, \frac{1}{p}\right] \tag{13}$$

Pérez and Lamata [15] discuss the weights determination by means of linear functions that we use in this paper.

To emphasize the conjunction, we propose to use decreasing weights $w_{min}$ defined by using the linear orders of Borda [4] where :

$$w_{min}(k) = \frac{2k}{p(p+1)} \tag{14}$$

where $1 \le k \le p$.

To emphasize the disjunction, we propose to use increasing weights $w_{max}$ defined by increasing linear orders with:

$$w_{max}(k) = \frac{2(p+1-k)}{p(p+1)} \tag{15}$$

where $1 \le k \le p$.

Then we have:

$$w_{min}(k) + w_{max}(k) = \frac{2}{p} \tag{16}$$

For example, if $p = 4$ $\begin{cases} W_{min} = \left[\frac{2}{20}, \frac{4}{20}, \frac{6}{20}, \frac{8}{20}\right] \\ W_{max} = \left[\frac{8}{20}, \frac{6}{20}, \frac{4}{20}, \frac{2}{20}\right] \end{cases}$ (17)

The membership function $\mu^i_k$ is a similarity index with $X_i$ in respect with the $k$-th attribute. $sens_{ik}$ describes the sensitivity of the index. In this paper, $sens_{ik}$ is considered as the ANDness of the attribute and 1-$sens_{ik}$ defines the specificity i.e. the ORness of the attribute. Then the weights of OWA operator we use are defined by:

$$w_k = C\left(\ (sens_{(ik)})w_{min}(k) + (1 - sens_{(ik)})w_{max}(k)\ \right) \tag{18}$$

where $C$ is the coefficient for obtaining $\sum_{1\le k\le p}(w_k) = 1$.

The membership degrees $\mu^i(X_j)$ with $X_j \in E$ are obtained through these weights. Such an OWA operator in $E$ permits us to define a similarity index $sim_i$ from the reference $X_i$ to the other objects $X_j$ of $E$ by:

$$sim(X_i, X_j) = \mu^i(X_j) \tag{19}$$

Note that the similarity index we propose is not necessarily symmetrical (cf. fig:fuzzy2). In fact $sim(X_i, X_j)$ is not always equal to $sim(X_j, X_i)$.

## V. Comparison with popular metrics

The way we describe to design pairwise similarity between multidimensional data leads us to propose a new pairwise similarity index based on fuzzy logic operators. This section is devoted to the assessment of the new similarity index we propose. First we define a criterion to compare the similarity indices. Second we apply this criterion for comparing our new pairwise similarity index with more classical indices based on the popular metrics.

### A. Assessment of a similarity index

Such a pairwise similarity indices are often used in the context of data clustering [9]. In this paper, we take the problem upside down using the clusters for assessing the similarity indices. The clusters define a partition of $E$, we call $C_{X_i}$ the cluster to which the object $X_i$ belongs and we call $sim$ a similarity index between two objects of $E$.

Let us consider all pairs of objects $(X_i, X_j)$ within the sample data $E$. If the two objects $X_i$ and $X_j$ belong to the same cluster, then an optimal similarity index from $X_i$ to $X_j$ should be equal to one (i.e. $X_i$ and $X_j$ are similar). On the contrary, if the two objects $X_i$ and $X_j$ belong to two different clusters, then an optimal similarity index from $X_i$ to $X_j$ should be equal to zero (i.e. $X_i$ and $X_j$ are dissimilar). Thus we define the intra-cluster similarity of $sim$ with:

$$intra(sim) = \frac{1}{n_1} \sum_{C_{X_i}=C_{X_j}} sim(X_i, X_j) \tag{20}$$

where $n_1$ is the number of couples $(X, Y)$ where $X$ and $Y$ belong to the same cluster. The inter-cluster similarity is defined with:

$$inter(sim) = \frac{1}{n_2} \sum_{C_{X_i} \neq C_{X_j}} sim(X_i, X_j) \tag{21}$$

where $n_2$ is the number of couples $(X_i, X_j)$ where $X_i$ and $X_j$ belong to two different clusters.

The similarity index $sim$ is optimal for the clusters when $intra(sim) = 1$ and $inter(sim) = 0$. Therefore we define a criterion to evaluate $sim$ with:

$$crit(sim) = intra(sim) - inter(sim) \tag{22}$$

The value $crit(sim)$ lies always between -1 and 1.

The higher $crit(sim)$, the more optimal $sim$ with respect to the clusters.

We propose to use this criterion to assess our new pairwise similarity index.

### B. Applications

This paper proposes a new way to evaluate the similarity between multidimensional vector data. The most classical way consists in using the popular metrics when data are quantitative. In this paper we consider Euclidean distance, Manhattan distance, Chebyshev distance, Canberra distance and Mahalanobis distance (see Table II). In fact, these distances are dissimilarity indices that we transform into similarity indices with:

$$simil(X, Y) = 1 - \frac{dist(X, Y)}{\max\limits_{A,B \in E} dist(A, B)} \tag{23}$$

where $dist$ is the distance that we use.

Then we have five similarity indices we call $Euclidean$, $Manhattan$, $Chebyshev$, $Canberra$, and $Mahalanobis$ which are based on their three respective popular metrics.

We compare these similarity indices based on distances with two indices we propose based on the aggregation operators of membership functions. The first one is called $simOWA$ that is the index which is described in this paper. It is based on OWA operators. The second one replaces the OWA operator with the arithmetic mean of the membership functions. This second one is called $Arithmetic$.

The similarity indices are computed using the databases from *Machine Learning Repository of UCI* [1].

In this paper we propose to use six numerical multivariate clustering databases that are $iris$, $wine$, $ecoli$, $glass$, $seeds$ and $haberman$. The number of attributes lies between 3 and 15. The number of objects lies between 100 and 500. The number of clusters lies between 3 and 10. $iris$ is the classical database that has 150 iris plants with 4 attributes and three clusters. The $wine$ recognition database has 178 objects with 13 attributes and three clusters. $ecoli$ is the database of sites of protein localization, it has 336 objects with 7 attributes and eight clusters. The $glass$ identification database has 214 objects with 9 attributes and seven clusters. The $seeds$ database of wheat varieties has 210 objects with 7 attributes and three clusters. $Haberman$'s survival database has 306 objects with 3 attributes and two clusters.

The results obtained are in Table III. We can see that the similarity indices proposed in this paper (SimOWA and Arithmetic) are better than the others in 5 cases and ranked second for one of them ($glass$ Database). In 5 cases of 6, the similarity indice based on the OWA operator gives us better results than the one based on the arithmetic mean.

## VI. Conclusion

The approach we propose has a significant advantage, it allows us to deal with imperfection that is a general case with real data. In medicine and biology, data is often imprecise mainly due to the inherent variability of biological data. In physics, data is also imperfect and it is usual to assign a value from a sensor with the accuracy of the measurement. Qualitative data is also imprecise or vague. Thus the use of the fuzzy set theory is relevant in this context of imperfect data.

In this paper we propose a simple method of fuzzification for imperfect multidimensional data. With this fuzzification we define a new similarity indice that will allow us in future works to identify the main features of the dataset and build a robust classification. We can also use this approach to compare a new object with the existing data, for example by finding the nearest objects.

## References

[1] Bache, K., Lichman, M., *UCI* Machine learning repository. http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences (2013)

[2] Barrena, M., Jurado, E., Marquez, P. and Pachon, C., *A flexible framework to ease nearest neighbor search in multidimensional data spaces*, Data and Knowledge Engineering, 69, pp. 116–136 (2010)

TABLE II
DEFINITION OF THE MOST POPULAR METRICS BETWEEN QUANTITATIVE DATA

| Popular metrics | |
|---|---|
| Euclidean | $dist(X,Y) = \sqrt{\sum_{1 \leq k \leq p} (x_k - y_k)^2}$ |
| Manhattan (city block) | $dist(X,Y) = \sum_{1 \leq k \leq p} \lvert x_k - y_k \rvert$ |
| Chebyshev | $dist(X,Y) = \max_{1 \leq k \leq p} \lvert x_k - y_k \rvert$ |
| Canberra | $dist(X,Y) = \sum_{1 \leq k \leq p} \frac{\lvert x_k - y_k \rvert}{\lvert x_k \rvert + \lvert y_k \rvert}$ |
| Mahalanobis | $dist(X,Y) = \sqrt{(X-Y)^T C^{-1} (X-Y)}$ |

TABLE III
COMPARISON OF SIMILARITY INDICES

| | Database | | | | | |
|---|---|---|---|---|---|---|
| | $iris$ | $wine$ | $ecoli$ | $glass$ | $seeds$ | $haberman$ |
| Number of objects | 150 | 178 | 336 | 214 | 210 | 306 |
| Number of attributes | 4 | 13 | 7 | 9 | 7 | 3 |
| Number of clusters | 3 | 3 | 8 | 7 | 3 | 2 |
| Index of similarity | | | | | | |
| Euclidean | 0.336 | 0.175 | 0.230 | 0.098 | 0.275 | 0.020 |
| Manhattan | 0.331 | 0.176 | 0.210 | 0.097 | 0.272 | 0.026 |
| Chebyshev | 0.344 | 0.175 | 0.211 | 0.085 | 0.258 | 0.017 |
| Canberra | 0.422 | 0.222 | 0.142 | **0.166** | 0.238 | 0.048 |
| Mahalanobis | 0.113 | 0.047 | 0.078 | 0.064 | 0.080 | 0.027 |
| Arithmetic | 0.506 | 0.264 | 0.245 | 0.142 | 0.447 | **0.052** |
| SimOWA | **0.510** | **0.270** | **0.289** | 0.151 | **0.451** | 0.044 |

[3] Bohm, C., Berchtold, S. and Keim, D.A., *Searching in high-dimensional spaces: index structures for improving the performance of multi-media databases*, ACM Computing Surveys, 33(3), pp. 322–373 (2001)

[4] de Borda, M., *Memoire sur les elections au scrutin*, Academie Royale des Sciences, Paris (1784)

[5] Cha, S.H., *Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions*, International Journal of Mathematical Models and Methods in Applied Sciences, 1(4), pp. 300–307 (2007)

[6] Cunningham, P., *A Taxonomy of Similarity Mechanisms for Case-Based Reasoning*, IEEE Trans. Knowledge and Data Engineering, Vol. 21 (11), pp. 1532–1543, (2009)

[7] Detyniecki, M., *Mathematical aggregation operators and their application to video querying*, Research Report, LIP6, Paris (2001)

[8] Dubois, D. and Prade, H., *On the use of aggregation operations in information fusion processes*, Fuzzy Sets and Systems, 142, pp. 143–161 (2004)

[9] Fred, A.N.L. and Jain, A.K., *Learning Pairwise Similarity for Data Clustering*, 18th International Conference on Pattern Recognition (ICPR 2006), vol. 1, pp. 925–928 (2006)

[10] De Mantaras, R.L., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., et al., *Retrieval, reuse, revision, and retention in case-based reasoning*, Knowledge Engineering Review, 20(3), pp. 215–240 (2005)

[11] Jain, A K., Murty, M.N. and Flynn, P.J., *Data clustering: a review*,ACM Computing Surveys, 31(3), pp. 264–323 (1999)

[12] Jain, A., *Data clustering: 50 years beyond k-means*, Pattern Recognition Letters, 31(8), pp. 651–666 (2010)

[13] Nourizadeh, A., Blanchard, F., Aït Younes, A., Delemer, B. and Herbin, M., *Data Analysis of Insulin Therapy in the Elderly Type 2 Diabetic Patients*, Studia Informatica Universalis, 11(3), pp. 32–49 (2013)

[14] Novak, D., Batko, M. and Zezula, P., *Large-scale similarity data management with distributed metric index*, Information Processing and Management, 48(5), pp. 855–872 (2012)

[15] Perez, E.C. and Lamata, M.T., *OWA weights determination by means of linear functions*, Mathware and Soft Computing, 16, pp. 107–122 (2009)

[16] Yager, R., *On ordered weighted averaging aggregation operators in multicriteria decision making*, IEEE Trans. Systems, Man and Cybern., 18(1), pp. 183–190 (1988)

[17] Yager, R., *Fuzzy logic methods in recommender systems*, Fuzzy Sets and Systems, 136, pp. 133–149 (2003)

[18] Zenebe, A. and Norcio, A.F., *Representation, similarity measures and aggregation methods using fuzzy sets for content-based recommender systems*, Fuzzy Sets and Systems, 160, pp. 76–94 (2009)

[19] Zerzucha, P. and Walczak, B., *Concept of (dis)similarity in data analysis*, Trends in Analytical Chemistry, 38, pp. 116–128 (2012)

[20] Zwick, R., Caristein, E. and Badescu, D.V., *Measures of similarity among fuzzy concepts: A comparative analysis*, International Journal of Approximate Reasoning, 1, pp. 221–242 (1987)

DOI: 10.15439/2015F90
ACSIS, Vol. 5

# Transformation of nominal features into numeric in supervised multi-class problems based on the weight of evidence parameter

Eftim Zdravevski[*], Petre Lameski[†], Andrea Kulakov[‡], Slobodan Kalajdziski[§]
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: {[*]eftim.zdravevski, [†]petre.lameski, [‡]andrea.kulakov, [§]slobodan.kalajdziski}@finki.ukim.mk

***Abstract*—Machine learning has received increased interest by both the scientific community and the industry. Most of the machine learning algorithms rely on certain distance metrics that can only be applied to numeric data. This becomes a problem in complex datasets that contain heterogeneous data consisted of numeric and nominal (i.e. categorical) features. Thus the need of transformation from nominal to numeric data. Weight of evidence (WoE) is one of the parameters that can be used for transformation of the nominal features to numeric. In this paper we describe a method that uses WoE to transform the features. Although the applicability of this method is researched to some extent, in this paper we extend its applicability for multi-class problems, which is a novelty. We compared it with the method that generates dummy features. We test both methods on binary and multi-class classification problems with different machine learning algorithms. Our experiments show that the WoE based transformation generates smaller number of features compared to the technique based on generation of dummy features while also improving the classification accuracy, reducing memory complexity and shortening the execution time. Be that as it may, we also point out some of its weaknesses and make some recommendations when to use the method based on dummy features generation instead.**

***Keywords*—Weight of Evidence, WoE, dummy features, data transformation, nominal features, categorical features, heterogeneous data**

## I. Introduction

CLASSIFICATION is one of the most researched problems in the data mining community. The data mining process is consisted of business and data understanding, data preparation, modeling, evaluation and deployment [1]. One of the crucial parts of this process is the data preparation which has a very large influence over the success or failure of the classification. Data preparation is not a trivial task and is dependent on the nature of the data. There are many problems that need to be addressed during the data preparation such as the existence of outliers, errors, noise, missing values etc. It is important that the data is processed and transformed correctly so that the problems that exist in the raw data are eliminated or their influence reduced as much as possible. For this we use different data transformation methods.

Depending on the data type of the features different transformations are suitable. Some of the most common methods for data transformation are well described in [2] and [3]. Transforming numeric data, be it continuous or discrete, can be done in a variety of ways. Then again, transformations of nominal and categorical data are not as extensively researched. This issue is also highlighted in [4]. A very common way for transformation of nominal features is by generating dummy features (i.e. varables). It is characterized by simplicity, independence of the data domain, and ease of implementation. Authors in [5, 3, 6] recommend to generate dummy features when there is no mapping of nominal to numeric data. After the data transformation the distance between the dummy features of the instances can be calculated in different ways, as described in [7]: Euclidean distance, Hamming distance, Jaccard distance, Levenshtein distance, etc. In this paper we treat binary dummy features as numeric features, and the distances are calculated intrinsically in the learning algorithm.

Another similarity measure for nominal features is proposed in [8]. It gives greater weight to uncommon feature value matches in similarity computations and makes no assumptions about the underlying distributions of the feature values. Application of this measure in an unsupervised setting to define the similarity metric between pairs of objects is proposed in [9].

The term frequency-inverse document frequency (TF-IDF), as described in [10] and [11], is often used in text mining problems as a numerical statistic which estimates the importance of a word to a document in a collection of documents. In a similar manner this weight can be used to transform arbitrary nominal values into numerical just as it assigns weight to words in text mining and information retrieval. This approach is mentioned in [12], but using TF-IDF as nominal data transformation technique during the preprocessing phase has not be widely researched.

The weight of evidence parameter, originally defined in [13], can be used for estimating the evidence in support of a hypothesis. Additionally it can be used for transformation of nominal data. Its applicability with examples is also discussed in [14] and [2]. There are some computational limitations of this method and we have addressed them in [15], so it can be used even when the preconditions are not met. In [16] we present an application of this transformation for calculation of the information value of features, which is consequently applied for feature selection.

One of the most serious drawbacks of this method is that it is applicable only to binary classification problems. In this paper we are providing theoretical foundations in order to extend the applicability of the WoE method to multi-class

978-83-60810-66-8/$25.00©2015, IEEE

problems. In order to evaluate its advantages, limitations and drawbacks we are comparing it against the technique that generates dummy features. The reason are using this technique as benchmark is because it is most widely adopted in both literature and practice. With this in mind, we have tested both methods on binary and multi-class classification problems and we have trained Support Vector Machines (SVMs) with different types of kernels and a feed forward back-propagation neural network. Finally, we summarize the findings of our research and discuss the applicability of the proposed extension of the weight of evidence parameter and give some recommendations of when to use it.

## II. Transformation with generation of dummy features

This section describes in more detail the method for transformation of nominal into numeric features based on generation of dummy features. By using this technique $n$ new dummy features are generated from a nominal feature that has $n$ different values, when $n > 2$. When $n = 2$ only one dummy feature is generated. Equation (1) gives the total number of generated dummy features where $v_i$ is the number of different values of the $i$-th nominal features. Each of the generated features can have a value of 0 or 1, depending on the occurrence of a particular value of the original feature. This approach has been published for the first time in [17] and was mainly used in regression analysis. Also, [18] covers many aspects related to regression analysis with dummy variables. Over time, this technique has been added to many software packages as a common stage before applying various machine-learning algorithms. When the number of nominal features and the number of different values they can have is small, this transformation leads to good performance of the algorithms. The problems arise when there are a lot of nominal features that can have many different values. These kinds of situations lead to rapid increase of the number of generated dummy features, which in turn, slows down machine-learning algorithms. In fact, the memory requirements or time complexity of algorithms can expand to that degree that they cannot be executed in a reasonable time on the computers that we have today. This issue can be partially addressed by discarding some of the generated features based on their predictive power. By doing that, some potentially useful information in the discarded features is consciously thrown away.

$$z_{dummy} = \sum_{i=0}^{n} v_i \tag{1}$$

## III. Weight of evidence

Decision making is mostly based on estimating the probability that one event might occur. The complexity of decision making varies from trivial decisions to some complex ones that require more involved processing of data from multiple sources. The outcome of this probabilistic decision making depends on facts that might even have inter dependencies [19]. For each decision we need to weight the influence of the facts that contribute to it. This provides us means of mapping the risk associated with a given choice or fact on a linear scale.

The concept of numerically weighting evidence was first introduced in [13] and is a result of the work performed by Alan Turing and I.J. Good during World War II. It is a statistical quantitative method for evaluating the facts (evidence) in support of a hypothesis. The weighting is performed with the parameter Weight of Evidence (WoE)[2]. It is a great tool for estimating the relative risk based on the available data. In this section we describe the mathematical background of WoE when used for binary classification problems, originally described in [15] and subsequently in [20].

Equation (2) defines the weight of evidence (WoE) of the $i$-th value of the feature $A$, where $N_i^A$ is the number of data points (i.e. instances) that were labeled as negative, and $P_i^A$ is the number of data points that were labeled as positive for the $i$-th value of the feature $A$. $SN$ is the total number of negatively labeled data points, $PN$ is the total number of positively labeled data points in the training set, and $n^A$ is the number of different values for the feature $A$.

$$WoE_i^A = ln\left(\frac{\frac{N_i^A}{SN}}{\frac{P_i^A}{SP}}\right) = ln\left(\frac{N_i^A}{P_i^A}\right) - ln\left(\frac{SN}{SP}\right) \tag{2}$$

Values of $SN$ and $SP$ can be calculated with (3) and (4), respectively:

$$SN = \sum_{i=1}^{n^A} N_i^A \tag{3}$$

$$SP = \sum_{i=1}^{n^A} P_i^A \tag{4}$$

WoE, as illustrated in the second part of (2), has two components: a variable component and a constant component. These numbers are independent of the machine learning algorithm that is going to be applied in the data mining process. They are calculated in the preprocessing phase. The variable component is calculated based on the data points that have a particular value of feature $A$ and the constant component is based on the whole sample (the training data set). In real-time systems, these values should be calculated at regular tyme intervals based on the dynamic of the new data input. There are various statistics that can be monitored so we can determine the need of updating the WoE parameters.

Equation (2) implies that the values for $N_i^A$ and $P_i^A$ have to be different than zero, and given that they represent counts, these constraints transform to $N_i^A > 0$ and $P_i^A > 0$. However, these preconditions are not always met in real datasets thus imposing limited applicability of the WoE parameter. The following section reviews the adjustments of the WoE technique that overcome these preconditions as they were proposed in [15].

## IV. Calculation of Weight of evidence for binary problems when preconditions are not met

As mentioned in the definition of WoE, in order to calculate WoE, we must satisfy the constraints $P_i^A > 0$ and $N_i^A > 0$.

These conditions must be met for any value of the feature so that the WoE can be computed. To be able to compute WoE for all values of the nominal feature that needs to be transformed, we need to make some adjustments in the way WoE is calculated for the cases where the preconditions are not met. The different types of unsatisfied preconditions as proposed in [15] are listed below.

Case 1: *The number of positively labeled data points is zero ($P_i^A = 0$) and the number of negatively labeled data points is zero ($N_i^A = 0$).* This is a trivial case when there are no data points (i.e. instances) with the $i$-th value of the feature $A$, even though this value is valid for this feature. In such case we assume that $WoE_i^A$ is zero, meaning that this particular value of the feature $A$ will have no impact on any transformations nor will change the calculation of some other parameters that are dependent on WoE. In fact, this value could be even deleted from the possible set of values for the current attribute. However, because it does not have effect on anything, it could be retained in the set of possible values in case some data points from the data sets have the $i$-th value of variable $A$, as well as for future analysis.

Case 2: *The number of positively labeled data points is zero ($P_i^A = 0$) and the number of negatively labeled data points is greater than zero ($N_i^A > 0$).* There are no positively labeled data points, and only negatively labeled data points with the $i$-th value of the feature $A$. In order to apply (2), we propose to use the value $P_i^A = 1$ for the positively labeled data points. At the same time, we propose to add the appropriate number of negatively labeled data points, so the overall ratio of the added positively and negatively labeled data points will be equal to the ratio of positively and negatively labeled data points in the whole data set ($SN/SP$). In the following equations let us denote the artificially added positive data points (in this case only one) with $\delta p_i^A = 1$ and with $\delta n_i^A$ denotes the added negative data points. These artificial "additions" of data points does not involve actual additions of instances in any data set, rather it only alters the number of counted data points of particular type for the purpose of the calculations. If instead of one "added" data point ($\delta p_i^A = 1$) we were to add more, then we would need to add more negatively labeled data points $\delta n_i^A$ compared to what we are adding now. This, in turn, may pose a problem because the artificial data points may become greater than the actual data points that are negatively labeled. Equations (5) and (6) define how the number of added data points will be calculated, as well as, the proposed estimate of WoE with (10).

$$\frac{\delta p_i^A}{\delta n_i^A} = \frac{SP}{SN} \tag{5}$$

After applying $\delta p_i^A = 1$ in (5), we can calculate $\delta n_i^A$ with (6):

$$\delta n_i^A = \frac{SN}{SP} \tag{6}$$

Now $P_i^A$ and $N_i^A$ that were defined in section III can be modified to include the artificially added data points:

$$\Delta P_i^A = P_i^A + \delta p_i^A = 1 \tag{7}$$

$$\Delta N_i^A = N_i^A + \delta n_i^A = N_i^A + \frac{SN}{SP} \tag{8}$$

Then, instead of calculating WoE with (2) using $N_i^A$ and $P_i^A$, we can calculate it with their modified values $\Delta N_i^A$ and $\Delta P_i^A$, defined in the two previous equations:

$$WoE_i^A = ln\left(\frac{\Delta N_i^A}{\Delta P_i^A}\right) - ln\left(\frac{SN}{SP}\right) \tag{9}$$

And if we apply (7) and (8) in (9), finally we get the proposed estimate of WoE for this case of unsatisfied preconditions:

$$WoE_i^A = ln\left(\frac{N_i^A \times SP + SN}{SN}\right) \tag{10}$$

Case 3: *The number of negatively labeled data points is zero ($N_i^A = 0$) and the number of positively labeled data points is greater than zero ($P_i^A > 0$).* There are only positively labeled data points with the $i$-th value of the variable $A$. We propose to add one data point that is labeled as negative, so we can use $N_i^A = 1$ when applying (2), and to add the appropriate number of positively labeled data points, so the overall ratio of the added positively and negatively labeled data points will be equal to the ratio of positively and negatively labeled data points in the whole data set ($SN/SP$). As in the previous case, these artificial "additions" of data points are virtual because the instances in the data sets are left intact, rather we only alter the number of counted data points of a particular type. As for why we are using only one artificial "addition" the same observation as in Case 2 stands. In this case the number of artificially added negative data points is one ($\delta n_i^A = 1$), so (5) can be transformed as (11):

$$\delta p_i^A = \frac{SP}{SN} \tag{11}$$

Now $P_i^A$ and $N_i^A$ that were defined in section III can be modified to include the artificially added data points:

$$\Delta P_i^A = P_i^A + \delta p_i^A = P_i^A + \frac{SP}{SN} \tag{12}$$

$$\Delta N_i^A = N_i^A + \delta n_i^A = 1 \tag{13}$$

Finally, for this case of unsatisfied preconditions, instead of calculating WoE with (2) using $N_i^A$ and $P_i^A$, we can calculate it with their modified values $\Delta N_i^A$ and $\Delta P_i^A$, defined in the two previous equations:

$$WoE_i^A = ln\left(\frac{SP}{P_i^A \times SN + SP}\right) \tag{14}$$

In this section we reviewed the enhancement [15] for estimation of WoE for cases in which the it can not be calculated by using the original (2). With this approach we add some data points so that the number of positive and negative labeled data points is always positive. This has a very small influence on the data and does not change the overall distribution of the dataset. There are several benefits from using the proposed method for WoE estimation:

- It would be computable for all features and all values in the data set, meaning that WoE could be used to transform the nominal features into numeric.
- The computed WoE could be used for binning of some values of the features.
- Information value of all features could be computed, and later it could be used for feature selection.
- Many classification algorithms have preference of numeric over nominal features, and sometimes the distance between different data points cannot be estimated if the values of the features are nominal. After we calculate the WoE values, the data points can be compared in terms of WoE.

The proposed transformation could degrade the performance of the classification model when significant noise is present. The estimated risk in such cases could defer from the real risk. Noisy data, however, poses a serious problem in the data mining in general and should be addressed before performing any kind of data transformation and using some machine learning algorithm.

## V. One-vs-all generalization for multi-class problems of the weight of evidence parameter

One of the most constraining properties of the Weight of evidence parameter is that it is computable only for binary classification problems. On the other hand, many real data mining and machine learning applications require classification into more than two classes. In order for the WoE parameter to be applicable for such cases, the algorithm for its calculation needs to be modified accordingly. One way to achieve this is by representing the multi-class classification problem as a set of binary problems. After that, we can calculate the WoE values separately for each of the binary subproblems. In [21] and [22] is applied a similar approach, known as one-vs-all or one-vs-rest, for generalization of many machine-learning algorithms that natively support only two classes (e.g. SVMs). By applying the one-vs-all technique we can also generalize the WoE transformation for multi-class problems. We have followed this idea originally in [23], but without substantial formal definition nor significant empirical evidence. Here we explore the idea of one-vs-all generalization of the WoE transformation in more depth.

Algorithm 1 is repeated for each of the $m$ classes. With this algorithm from a dataset with $k$ nominal features and $m$ classes we generate $z_{woe}$ new numeric features, as defined with (15).

**Algorithm 1** One-vs-all generalization for multi-class problems of the Weight of evidence parameter

```
for $i = 1 \to m$ do
    Temporary label with $TempClass^1$ all instances that were originally labeled with $Class^i$).
    Temporary label with $TempClass^2$ all instances that were originally labeled with some class different than $Class^i$).
    Calculate the WoE parameters for all instances and all their nominal features using the temporary labels ($TempClass^1$ and $TempClass^2$).
    Transform all $k$ nominal features using the calculated WoE parameters in the previous step. This produces $k$ new numeric features.
    Add the $k$ generated numeric features to the transformed dataset.
    Remove the temporary labels of all instances and revert them to their original labels (classes).
end for
```

$$z_{woe} = \begin{cases} m \times n, & m > 2 \\ n, & m = 2 \end{cases} \tag{15}$$

The same algorithm can be applied for transformation of the numeric attributes in the original dataset into another numeric features as well. However, given the fact that there are plenty of algorithms for transformation of numeric features, we are not focusing on that kind of application of the WoE transformation.

## VI. Results

In this section we present the experiments that we conducted using the proposed method for data transformation. The first four subsections describe the performance metric and the cross-validation process that we used across all subsequent tests, how and when we performed feature selection and how we evaluated the performance. The following subsections describe how we have applied the weight of evidence transformation on the nominal features in some of the datasets obtained from the UCI Repository of Machine Learning Databases [24].

### A. Performance metric

The choice of performance metric to evaluate any transformation is very important. Several research papers indicate the fact that in some cases for a given dataset, the learning method that obtains the best model according to a given measure, is not the best method if a different measure is used. In [25] is shown that Naive Bayes and pruned decision trees are very similar in predictive accuracy. Later on, applying the same algorithms, in [26] the same authors show that Naive Bayes is significantly better than pruned decision trees in terms of AUC ROC. As it is said in [27] the different results cannot be explained by slightly different implementations or variants of machine learning algorithms, but on the fact that the two measures (accuracy and AUC ROC) evaluate different things. The predictive accuracy is perhaps the most popular metric for classification problems and many researchers have been publishing papers that show the performance of various algorithms, techniques and transformations in terms of predictive accuracy on the same datasets that we also use in this paper. Another property of accuracy is that it can be calculated for multi-class problems in the same way as it is calculated for binary problems, while other metrics does not natively support multi-class problems. Because of these reasons we have also decided to compare our results in terms of predictive accuracy.

Additionally, we wanted to compare the execution times of both transformations so we can have insight of that aspect too. We have implemented all algorithms in Matlab and all tests have been performed on Windows 7 Professional SP1 64 bit running on a virtual Intel Xeon X5680 at 3.33 Ghz with 8 GB RAM. We have used the built-in implementations of FFNN and SVMs in Matlab. All listed execution times consist of the data transformation, the training of the machine learning algorithm and making the predictions.

### B. Cross-validation

Some of the available datasets in the research community are consisted of training, validation and testing subset. However more often they are consisted of only one set that should be used for training, validation and testing, so the task of splitting the original dataset is left to the data analysts and researchers. The most common practice when evaluating performance in such cases is to perform cross-validation. There are few alternatives of how can cross-validation be performed and each of them has advantages and disadvantages. Stratified 10-fold cross-validation in [28] is recommended as the best model selection method, as it tends to provide less biased estimation of the accuracy. Following this recommendation we have decided to perform $k$-fold cross-validation while using different values for $k$: 2, 4, 6, 8 and 10. Usually the cross-validation is performed after the data is preprocessed. This means that all data cleaning, data transformations, and removal of outliers have to be performed, and then different subsets from the processed datasets are selected as training, validation and testing. This approach is suitable when the data transformations depend only on the actual values of features that should be transformed. Such transformations techniques are all mathematical functions that can be applied on numerical features or generation of dummy features from nominal features.

However, if the data transformations additionally depend on the relationship between the instances in the training set this approach is not suitable. Such relationship is present in the Weight of Evidence transformation because the transformed values depend on the set of values of the features that are being transformed and the set of values of all other instances in the dataset, as it is evident in (2). To review, the transformed value depends on the number of instances labeled with a particular class, and the number of instances that have a particular value of the original feature and are labeled with a particular class. These counts can vary for different subsets of the original dataset, thus a different transformed values can be obtained for the same original value. Ideally, if the dataset is large enough and if the distribution of values of the nominal features is nearly uniform, then the transformed values for different subsets of the dataset would not vary much. However, such conditions do not occur in real datasets, hence the need of to modify the cross-validation process. Algorithm 2 describes how the k-fold cross-validation is performed with repeated transformation of nominal features. By applying this algorithm for cross-validation, only the information that is present in the current set of training folds is used for data transformation.

**Algorithm 2** K-fold cross-validation with repeated transformation of nominal features

```
Randomly shufle the dataset
Randomly assign $Fold_i$ $(i = 1 \to k)$ to each instance in the shuffled dataset
for $i = 1 \to k$ do
    Assign all instances belonging to $Fold_i$ to $TestSet_i$.
    Assign all instances not belonging to $Fold_i$ to $TrainingSet_i$.
    Calculate $WoE_i$ parameters for all nominal features in $TrainingSet_i$.
    Transform all values of all nominal features in $TrainingSet_i$ using the transformation values $WoE_i$, thus obtaining $TransformedTrainingSet_i$.
    Transform all values of all nominal features in $TestSet_i$ using the transformation values $WoE_i$, thus obtaining $TransformedTestSet_i$.
    Train $Model_i$ using $TransformedTrainingSet_i$.
    Validate $Model_i$ using $TransformedTestSet_i$, thus obtaining $Results_i$.
end for
Aggregate $Results_i (i = 1 \to k)$
```

### C. Feature selection

In the machine learning literature there are a lot of published papers and books about various feature selection techniques, both for nominal and numeric features. Such paper is [29], where the most popular algorithms are described and illustrated with examples. This work focuses on data transformation techniques and therefore we have performed some basic feature selection, and have not tested more advanced methods. First thing that is performed for all datasets that we have worked on, all single-value features were removed. Also if such features were generated with one of the applied data transformation techniques they were also removed. This is because these features have no information value i.e. their entropy is zero, therefore having no predictive power regardless of which machine learning algorithm would later be applied.

In the case when some nominal feature has many different values and the dummy variables generation technique is used, a lot of dummy features would be generated. This can pose a serious problem for many machine learning algorithms because they would demand enormous amount of computer power, even to the extent that they would not be usable. To overcome this problem we have used a simple feature selection that restricts the number of dummy features that can be generated. Namely, only for the values of the nominal features that occur in more than 5% of the instances a dummy feature is generated, while for all other less frequently occurring values we generate only one common dummy feature. Choosing the value of 5% as a threshold provided good balance between retaining a small number of features while not loosing too much information. This simple algorithm is very effective in helping prevent the generation of vast amount of dummy features. Of course, a more intelligent feature selection can be applied but we did not want to defocus from the main topic of the paper.

### D. Performance evaluation

After applying the proposed transformation we have transformed all nominal features in the datasets into numeric and then we have trained few machine learning algorithms. In order to have basis for comparison, we have also generated dummy features from all nominal features in the original datasets and afterwards we have trained the same machine learning algorithms. To decrease the impact of randomness while making the splits of the data into cross-validation folds we have repeated the whole process multiple times and then we have aggregated the results. The following subsections describe each of the datasets, the performed data transformations and the performance in more detail. The performance is analyzed in terms of accuracy and execution time.

### E. The Annealing dataset

The annealing dataset was obtained from the UCI Repository of Machine Learning Databases [24] and it contains data for a classification problem. More particularly, it contains 798 instances described with 6 numeric and 32 nominal features. All instances are labeled with one of the 5 possible classes. Only the nominal features of the dataset were subject to transformation, while the numeric features were not transformed in any way.

First, we have generated dummy features for all different values of all nominal features in the original dataset. Note that for nominal features that have only 2 different values we do not generate dummy features rather we only convert their values to 0s and 1s, because these features are already dummy features with differently encoded values. By doing that we have generated 64 dummy features. From them 7 had the same value for all instances in the training and test sets, therefore they were removed. Finally, together with the original 6 numeric features the dataset 1 is comprised of 63 numeric features.

Then we have applied the proposed WoE transformation thus obtaining $32 \times 5 = 160$ new numeric features, as defined with (15). From them 40 had the same value for all instances in the training and test sets, meaning that there is no information value in them, therefore they were removed. Finally, together with the original 6 numeric features the dataset 2 is comprised of 126 numeric features.

Both datasets were tested using a feed forward back propagation neural network (FFNN), and SVM with a linear, quadratic, polynomial and RBF kernel. The training and test partitions of the datasets were obtained using k-fold cross validation and 2, 4, 6, 8 and 10 were used as $k$ values. An important thing to mention is that a same value of the original feature can be transformed into different values. This situation arises from the nature of the WoE values - they depend not only on the original values but also on the distribution of the other values of the same feature in the training dataset. This in turn require the WoE transformation to be performed for each training fold separately. In fact for k-fold cross validation the WoE transformation would be performed $k$ times, once for each training fold combination. Because the instances that belong to the folds are chosen randomly, the performance can vary and may not be consistent. Therefore, the whole process was repeated 10 times for each value of $k$. At the end we have calculated the average, minimum, maximum and standard deviation of the accuracies and the execution times for each algorithm and each transformation type using the data from the 10 repetitions.

Table I shows the accuracies for different values of $k$ of the 10 repetitions when a FFNN was trained with both datasets. We can see that for all values of $k$ the WoE transformed dataset produced better average accuracy. Next, table II shows the execution times when a FFNN was trained with both datasets. We can see that both transformations together with the training and test phase needed were completed in similar time, even though the WoE transformed dataset had about twice more features.

The next algorithm we trained was SVM with linear kernel. Tables III and IV show the results of this algorithm using both datasets. For it we can note that both the accuracy and execution time are similar to each other for both transformations, but are much better than the FFNN.

After that, we have trained a SVM with RBF (i.e. Radial Basis Function) kernel. Tables V and VI show the results of this algorithm using both datasets. For this algorithm we can note that both the accuracy and the execution time are similar to each other for both transformations, are significantly better than the FFNN and worse than the SVM with linear kernel.

The following algorithm that we have trained was a SVM with Polynomial kernel. Tables VII and VIII show the results of this algorithm using both datasets. For this algorithm we can note that the accuracy is slightly better when dummy transfomation is used and execution time is slightly better when WoE transformation is used. Overall the performance is similar to the SVM with linear kernel and to the performance of SVM with Quadratic kernel, which are shown with tables IX and X.

Tables XI and XII show the results of a SVM with MLP (i.e. multilayer perceptron) kernel using both datasets. For this algorithm we can note that the accuracy is slightly better when dummy transfomation is used and execution time is slightly better when WoE transformation is used. Overall the performance is similar to the SVM with RBF kernel.

Finally, we can conclude that for this particular dataset both transformations lead to similar performance in terms of accuracy and execution time when various machine learning algorithms are applied. We want to point out again the fact that the dataset has 1000 instances that are non-uniformly distributed into 5 classes. As a consequence the execution of the transformations is very fast and therefore they are very similar. Additionally the number of different values of the nominal features is very small therefore the advantages of the WoE transformation can not be exploited. For datasets like this one, where the number of instances, the number of nominal features and the number of different values are fairly small, both transformations provide similar results. However, in such cases we recommend applying the dummy transformation because it is easier to implement and it is a lot simpler to interpret and understand.

TABLE I. ANNEALING DATASET. ***Accuracy*** *from 10 repetitions of k-fold cross-validation with* ***FFNN***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.7617 | 0.7628 | 0.7718 | 0.7862 | 0.7239 |
| **Max** | 0.8307 | 0.8364 | 0.8342 | 0.8273 | 0.8230 |
| **Mean** | 0.7951 | 0.8025 | 0.7981 | 0.8081 | 0.7882 |
| **StDev** | 0.0244 | 0.0229 | 0.0191 | 0.0127 | 0.0242 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.7550 | 0.7728 | 0.7030 | 0.7917 | 0.7773 |
| **Max** | 0.8408 | 0.8954 | 0.8731 | 0.8375 | 0.8430 |
| **Mean** | 0.7989 | 0.8258 | 0.8114 | 0.8116 | 0.8113 |
| **StDev** | 0.0258 | 0.0335 | 0.0420 | 0.0139 | 0.0207 |

TABLE II. ANNEALING DATASET. ***Execution time*** *in seconds from 10 repetitions of k-fold cross-validation with* ***FFNN***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 1.75 | 4.56 | 7.32 | 10.58 | 13.45 |
| **Max** | 9.48 | 6.25 | 10.99 | 12.66 | 18.95 |
| **Mean** | 2.83 | 5.43 | 9.02 | 11.62 | 14.88 |
| **StDev** | 2.23 | 0.53 | 0.96 | 0.63 | 1.49 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 1.96 | 5.33 | 8.64 | 12.12 | 14.35 |
| **Max** | 2.74 | 6.45 | 10.27 | 13.38 | 18.56 |
| **Mean** | 2.31 | 5.89 | 9.21 | 12.63 | 15.93 |
| **StDev** | 0.26 | 0.37 | 0.44 | 0.41 | 1.25 |

TABLE III. ANNEALING DATASET. ***Accuracy*** *from 10 repetitions of k-fold cross-validation with* ***SVM with linear kernel***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.9755 | 0.9866 | 0.9677 | 0.9732 | 0.9888 |
| **Max** | 0.9889 | 0.9922 | 0.9922 | 0.9922 | 0.9922 |
| **Mean** | 0.9825 | 0.9889 | 0.9884 | 0.9892 | 0.9906 |
| **StDev** | 0.0042 | 0.0017 | 0.0071 | 0.0054 | 0.0009 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.9755 | 0.9611 | 0.9866 | 0.9866 | 0.9900 |
| **Max** | 0.9911 | 0.9944 | 0.9944 | 0.9944 | 0.9922 |
| **Mean** | 0.9850 | 0.9882 | 0.9914 | 0.9919 | 0.9915 |
| **StDev** | 0.0041 | 0.0092 | 0.0025 | 0.0021 | 0.0007 |

TABLE IV. ANNEALING DATASET. ***Execution time*** *in seconds from 10 repetitions of k-fold cross-validation with* ***SVM with linear kernel***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 1.32 | 2.04 | 4.22 | 6.12 | 6.00 |
| **Max** | 2.07 | 3.80 | 6.21 | 10.46 | 13.36 |
| **Mean** | 1.49 | 2.67 | 5.44 | 7.69 | 9.06 |
| **StDev** | 0.21 | 0.49 | 0.62 | 1.38 | 1.92 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 1.55 | 3.34 | 4.74 | 5.75 | 8.22 |
| **Max** | 2.56 | 5.80 | 6.69 | 9.64 | 11.13 |
| **Mean** | 2.01 | 4.30 | 5.80 | 7.79 | 9.39 |
| **StDev** | 0.33 | 0.71 | 0.65 | 1.15 | 0.79 |

TABLE V. ANNEALING DATASET. ***Accuracy*** *from 10 repetitions of k-fold cross-validation with* ***SVM with RBF kernel***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.8664 | 0.8953 | 0.9053 | 0.9087 | 0.9064 |
| **Max** | 0.8953 | 0.9199 | 0.9198 | 0.9243 | 0.9243 |
| **Mean** | 0.8814 | 0.9081 | 0.9139 | 0.9180 | 0.9187 |
| **StDev** | 0.0086 | 0.0070 | 0.0044 | 0.0054 | 0.0050 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.8641 | 0.8942 | 0.9031 | 0.9076 | 0.9064 |
| **Max** | 0.8931 | 0.9176 | 0.9176 | 0.9242 | 0.9243 |
| **Mean** | 0.8782 | 0.9067 | 0.9126 | 0.9173 | 0.9177 |
| **StDev** | 0.0088 | 0.0067 | 0.0044 | 0.0054 | 0.0048 |

TABLE VI. ANNEALING DATASET. ***Execution time*** *in seconds from 10 repetitions of k-fold cross-validation with* ***SVM with RBF kernel***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 1.72 | 4.09 | 6.44 | 8.79 | 11.45 |
| **Max** | 2.03 | 4.41 | 7.10 | 11.14 | 15.50 |
| **Mean** | 1.83 | 4.29 | 6.73 | 9.44 | 12.05 |
| **StDev** | 0.08 | 0.10 | 0.24 | 0.60 | 1.16 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 1.77 | 4.15 | 6.26 | 9.08 | 11.14 |
| **Max** | 1.97 | 4.63 | 7.18 | 10.44 | 13.06 |
| **Mean** | 1.84 | 4.42 | 6.74 | 9.49 | 11.89 |
| **StDev** | 0.06 | 0.13 | 0.28 | 0.42 | 0.48 |

TABLE VII. ANNEALING DATASET. ***Accuracy*** *from 10 repetitions of k-fold cross-validation with* ***SVM with Polynomial kernel***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.9788 | 0.9889 | 0.9900 | 0.9900 | 0.9878 |
| **Max** | 0.9889 | 0.9933 | 0.9933 | 0.9933 | 0.9933 |
| **Mean** | 0.9860 | 0.9914 | 0.9922 | 0.9920 | 0.9916 |
| **StDev** | 0.0028 | 0.0020 | 0.0011 | 0.0010 | 0.0015 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.9577 | 0.9655 | 0.9654 | 0.9666 | 0.9655 |
| **Max** | 0.9755 | 0.9755 | 0.9733 | 0.9744 | 0.9710 |
| **Mean** | 0.9674 | 0.9701 | 0.9695 | 0.9698 | 0.9689 |
| **StDev** | 0.0056 | 0.0028 | 0.0024 | 0.0020 | 0.0019 |

TABLE VIII. ANNEALING DATASET. ***Execution time*** *in seconds from 10 repetitions of k-fold cross-validation with* ***SVM with Polynomial kernel***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.98 | 1.84 | 2.64 | 3.44 | 4.33 |
| **Max** | 1.11 | 2.14 | 2.89 | 4.35 | 4.93 |
| **Mean** | 1.04 | 1.97 | 2.78 | 3.71 | 4.50 |
| **StDev** | 0.04 | 0.08 | 0.09 | 0.24 | 0.19 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.89 | 1.67 | 2.40 | 3.23 | 3.96 |
| **Max** | 1.03 | 2.86 | 2.78 | 3.84 | 4.60 |
| **Mean** | 0.95 | 1.87 | 2.57 | 3.45 | 4.20 |
| **StDev** | 0.04 | 0.34 | 0.13 | 0.17 | 0.17 |

TABLE IX. ANNEALING DATASET. ***Accuracy*** *from 10 repetitions of k-fold cross-validation with* ***SVM with Quadratic kernel***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.9766 | 0.9911 | 0.9922 | 0.9900 | 0.9833 |
| **Max** | 0.9933 | 0.9956 | 0.9955 | 0.9956 | 0.9956 |
| **Mean** | 0.9874 | 0.9933 | 0.9948 | 0.9947 | 0.9934 |
| **StDev** | 0.0043 | 0.0014 | 0.0012 | 0.0017 | 0.0036 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.9744 | 0.9889 | 0.9922 | 0.9889 | 0.9844 |
| **Max** | 0.9900 | 0.9944 | 0.9944 | 0.9955 | 0.9956 |
| **Mean** | 0.9840 | 0.9918 | 0.9939 | 0.9929 | 0.9923 |
| **StDev** | 0.0043 | 0.0017 | 0.0009 | 0.0023 | 0.0034 |

TABLE X. ANNEALING DATASET. ***Execution time*** *in seconds from 10 repetitions of k-fold cross-validation with* ***SVM with Quadratic kernel***

| | **Dummy transformation** | | | | |
|---|---|---|---|---|---|
| | **k=2** | **k=4** | **k=6** | **k=8** | **k=10** |
| **Min** | 0.97 | 1.73 | 2.41 | 3.33 | 4.05 |
| **Max** | 1.18 | 2.21 | 2.80 | 3.90 | 4.45 |
| **Mean** | 1.04 | 1.91 | 2.64 | 3.53 | 4.21 |
| **StDev** | 0.06 | 0.15 | 0.12 | 0.17 | 0.12 |

| | **WoE transformation** | | | | |
|---|---|---|---|---|---|
| | **k=2** | **k=4** | **k=6** | **k=8** | **k=10** |
| **Min** | 0.87 | 1.61 | 2.30 | 3.07 | 3.63 |
| **Max** | 0.99 | 1.92 | 2.57 | 3.65 | 4.18 |
| **Mean** | 0.94 | 1.74 | 2.43 | 3.26 | 3.81 |
| **StDev** | 0.04 | 0.10 | 0.07 | 0.15 | 0.18 |

TABLE XI. ANNEALING DATASET. ***Accuracy*** *from 10 repetitions of k-fold cross-validation with* ***SVM with MLP kernel***

| | **Dummy transformation** | | | | |
|---|---|---|---|---|---|
| | **k=2** | **k=4** | **k=6** | **k=8** | **k=10** |
| **Min** | 0.8664 | 0.8953 | 0.9053 | 0.9087 | 0.9064 |
| **Max** | 0.8953 | 0.9199 | 0.9198 | 0.9243 | 0.9243 |
| **Mean** | 0.8814 | 0.9081 | 0.9139 | 0.9180 | 0.9187 |
| **StDev** | 0.0086 | 0.0070 | 0.0044 | 0.0054 | 0.0050 |

| | **WoE transformation** | | | | |
|---|---|---|---|---|---|
| | **k=2** | **k=4** | **k=6** | **k=8** | **k=10** |
| **Min** | 0.8641 | 0.8942 | 0.9031 | 0.9076 | 0.9064 |
| **Max** | 0.8931 | 0.9176 | 0.9176 | 0.9242 | 0.9243 |
| **Mean** | 0.8782 | 0.9067 | 0.9126 | 0.9173 | 0.9177 |
| **StDev** | 0.0088 | 0.0067 | 0.0044 | 0.0054 | 0.0048 |

### *F. The PAKDD 2010 dataset*

The 14th Pacific-Asia Knowledge Discovery and Data Mining conference (PAKDD 2010) together with NeuroTech Ltd. and the Center for Informatics of the Federal University of Pernambuco (Brazil) co-organized a data mining competition [30]. This credit risk assessment problem comes from the private label credit card operation of a major retail chain. The company has been operating its private label card for over 10 years and has applied two different methods for risk assessment with the application's acceptance rate varying from 50% to 75% within this period. Each accepted application turns the applicant into a client and gives him/her the access to credit for purchasing on the retail chain to be billed 10 to 40 days after the purchase, on a monthly basis on a fixed month day. After his/her credit acceptance, a client would take some time to make their first purchase and receive their first bill. During the first year of using the card, the set of monthly bills and payment behavior is collected and used for credit risk assessment. If the client had any monthly defaults (delays longer than the agreed payment periods) he is labeled as bad, otherwise as good client. The goal is to exploit the information that was available when the applicant applied for credit and try to predict whether he would be a good or bad client. To achieve this we have used the training data consisting of various kinds of information for the applicants like age, profession, sex, marital status, monthly income etc. and the label (good/bad) to build prediction models.

TABLE XII. ANNEALING DATASET. ***Execution time*** *in seconds from 10 repetitions of k-fold cross-validation with* ***SVM with MLP kernel***

| | **Dummy transformation** | | | | |
|---|---|---|---|---|---|
| | **k=2** | **k=4** | **k=6** | **k=8** | **k=10** |
| **Min** | 1.74 | 4.19 | 6.38 | 9.13 | 11.33 |
| **Max** | 1.94 | 4.82 | 7.09 | 10.22 | 12.67 |
| **Mean** | 1.84 | 4.42 | 6.70 | 9.42 | 11.76 |
| **StDev** | 0.07 | 0.22 | 0.19 | 0.29 | 0.34 |

| | **WoE transformation** | | | | |
|---|---|---|---|---|---|
| | **k=2** | **k=4** | **k=6** | **k=8** | **k=10** |
| **Min** | 1.73 | 4.25 | 6.27 | 8.93 | 11.03 |
| **Max** | 2.01 | 6.07 | 7.09 | 11.28 | 12.43 |
| **Mean** | 1.87 | 4.65 | 6.69 | 9.59 | 11.77 |
| **StDev** | 0.08 | 0.49 | 0.24 | 0.62 | 0.39 |

For the purpose of the competition there are three available datasets, but the labels (i.e. target class) of only one of them are made publicly available. This dataset is named as Modeling and has 50000 instances, distributed as 26% vs. 74% per class. The dataset is real and collected manually during a long time period therefore some of the instances have missing, invalid or inconsistent data, some of the columns have no information value or have the same values for all instances etc. We have addressed these issues by removing some of the instances and the columns. Additionally, we have generated few nominal features that capture interactions between the original features: sex, marital status, age group, profession etc. After this stage of data cleaning and preparation we ended up with a dataset with 11 numeric and 24 nominal features and a total of about 42000 instances that were similarly distributed in respect to the target class.

First, we have generated dummy features for all different values of all nominal features in the original dataset. Note that for nominal features that have only 2 different values we do not generate dummy features rather we only convert their values to 0s and 1s, because these features are already dummy features with differently encoded values. By doing that we have obtained around 4630 dummy features. Obviously a dataset with such high number of features makes it difficult to train models on it. That is why we have performed a simple feature selection as described in subsection VI-C, which resulted in a dataset with about 140 dummy features. Finally, together with the original 11 numeric features the dataset 1 is comprised of about 150 numeric features. The exact number of features varies depending on the random split during the cross-validation.

Then we have applied the proposed WoE transformation and we generated for each nominal features one numeric feature which resulted in 24 new numeric features, as defined with (15). Together with the original 11 numeric features the dataset 2 is comprised of 35 numeric features.

Both datasets were tested using a feed forward back propagation neural network, and SVM with a linear, quadratic, polynomial and RBF kernel. The training and test partitions of the datasets were obtained using k-fold cross validation and 2, 4, 6, 8 and 10 were used as $k$ values. The whole process was repeated 10 times for each value of $k$, so the influence of randomness during the cross-validations can be neglected. The official performance metric of the competition is area under the receiver operating curve (AUC ROC) [26], but due to the reasons discussed in subsection VI-A, we have decided to compare the results in terms of accuracy. The first results were not very useful because the accuracy of the classifiers built from both datasets was in the most cases around 74%. Additionally in many cases the machine learning algorithms could not converge, meaning we could not compare the both transformations. The WoE transformed dataset usually provided better accuracy, but the overall results were insufficient to make firm conclusions. The fact that the accuracy was similar to the percentage of the more common class and the confusion matrices undoubtedly showed that the unbalanced dataset introduces serious problems for all machine learning algorithms. The algorithms trained on the dummy transformed dataset converged twice less often than the ones trained on the WoE transformed dataset, and in the rare cases when they both converged the WoE transformation almost always provided better results.

In order to make better comparisons of the transformations, it was obvious that a balanced dataset would be more suitable. While we can artificially generate more instances of the less common class, the most common way to balance a dataset is by "throwing away" some of the instances labeled with the more common class. When data is balanced, accuracy rates tend to decline [31]. If we balance the dataset by reducing the training set size, then this can lead to the degeneracy of the model because we are neglecting potentially useful training instances. Nevertheless, we have opted for this option because it was easier to implement the shrinking, but mostly because this way we can train algorithms much faster because of the smaller training sets. By doing this we have obtained a balanced dataset with around 22000 instances and afterwards we have performed all tests described in the previous paragraph. Important to mention here is that by repeating the whole tests 10 times we were able to select different subsets of instances of the more common class for each repetition, thus mitigating the issues of thrown-away information to some extent.

From all algorithms that were tested only the FFNN converged for all values of $k$ and all repetitions. The achieved results are shown in tables XIII and XIV. We can conclude that when FFNN was trained the WoE transformation produced 6% to 9% better accuracy in 10% to 40% more time than the dummy transformation.

After the training was performed for the SVM with RBF kernel we noticed that convergence could not be acheived when $k$ is 4, 6, 8 and 10 when the dummy transformed dataset is used. Similary, when the WoE transformed dataset was used the SVM converged when $k$ was 2 and in six of the repetitions when $k$ was 4. The results that we were able to collect are listed in tables XV and XVI. The results for SVM with MLP kernel were so similar that the averages look exactly the same despite the fact that the individual values were different. This can be seen in tables XVII and XVIII. From them we can conclude that when SVM with RBF or MLP kernel was trained the WoE transformation produced about 2% better accuracy about three times faster than the dummy transformation and additionally the SVMs were able to converge more often. Compared to the the FFNN, the SVMs produced about 20% better accuracy for $k = 2$ but were significantly slower - about 6 times.

The other SVMs with linear, polynomial and quadratic kernel could not converge for none of the repetitions and different $k$ values. We realize that these issues might be addressed by parameter tuning, but we cannot concentrate on this issue at the moment because it is not the focus point of this paper.

TABLE XIII. **PAKDD 2010 DATASET.** ***Accuracy*** *from 10 repetitions of k-fold cross-validation with* ***FFNN***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.5122 | 0.5443 | 0.5430 | 0.5478 | 0.5536 |
| **Max** | 0.5944 | 0.5900 | 0.5903 | 0.5935 | 0.5860 |
| **Mean** | 0.5547 | 0.5711 | 0.5697 | 0.5739 | 0.5706 |
| **StDev** | 0.0246 | 0.0144 | 0.0147 | 0.0149 | 0.0088 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 0.5028 | 0.5472 | 0.5610 | 0.5868 | 0.5702 |
| **Max** | 0.6453 | 0.6561 | 0.6596 | 0.6377 | 0.6567 |
| **Mean** | 0.5894 | 0.6214 | 0.6158 | 0.6111 | 0.6123 |
| **StDev** | 0.0504 | 0.0328 | 0.0262 | 0.0149 | 0.0227 |

TABLE XIV. **PAKDD 2010 DATASET.** ***Execution time*** *in seconds from 10 repetitions of k-fold cross-validation with* ***FFNN***

**Dummy transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 29.59 | 84.85 | 141.67 | 183.26 | 242.82 |
| **Max** | 45.16 | 103.84 | 309.29 | 223.04 | 272.56 |
| **Mean** | 35.54 | 93.86 | 169.14 | 207.04 | 257.37 |
| **StDev** | 4.67 | 5.61 | 47.23 | 12.04 | 8.53 |

**WoE transformation**

| | k=2 | k=4 | k=6 | k=8 | k=10 |
|---|---|---|---|---|---|
| **Min** | 41.54 | 104.24 | 168.24 | 251.99 | 302.14 |
| **Max** | 48.07 | 297.96 | 200.57 | 279.83 | 388.43 |
| **Mean** | 44.21 | 135.97 | 190.84 | 267.72 | 338.81 |
| **StDev** | 1.69 | 54.34 | 9.04 | 6.94 | 21.50 |

TABLE XV. **PAKDD2010 BALANCED DATASET.** ***Accuracy*** *from 10 repetitions of k-fold cross-validation with* ***SVM with RBF kernel***

| **Dummy trans.** | | | **WoE trans.** | |
|---|---|---|---|---|
| | **k=2** | | **k=2** | **k=4** |
| **Min** | 0.7300 | **Min** | 0.7409 | 0.8044 |
| **Max** | 0.7400 | **Max** | 0.7531 | 0.8089 |
| **Mean** | 0.7336 | **Mean** | 0.7453 | 0.8063 |
| **StDev** | 0.0029 | **StDev** | 0.0035 | 0.0015 |

TABLE XVI. **PAKDD2010 BALANCED DATASET.** ***Execution times*** *in seconds from 10 repetitions of k-fold cross-validation with* ***SVM with RBF kernel***

| **Dummy trans.** | | | **WoE trans.** | |
|---|---|---|---|---|
| | **k=2** | | **k=2** | **k=4** |
| **Min** | 713.34 | **Min** | 254.66 | 851.89 |
| **Max** | 968.45 | **Max** | 274.56 | 1054.31 |
| **Mean** | 803.70 | **Mean** | 262.75 | 954.75 |
| **StDev** | 82.49 | **StDev** | 6.22 | 88.40 |

TABLE XVII. **PAKDD2010 BALANCED DATASET.** ***Accuracy*** *from 10 repetitions of k-fold cross-validation with* ***SVM with MLP kernel***

| **Dummy trans.** | | | **WoE trans.** | |
|---|---|---|---|---|
| | **k=2** | | **k=2** | **k=4** |
| **Min** | 0.7300 | **Min** | 0.7409 | 0.8044 |
| **Max** | 0.7400 | **Max** | 0.7531 | 0.8089 |
| **Mean** | 0.7336 | **Mean** | 0.7453 | 0.8063 |
| **StDev** | 0.0029 | **StDev** | 0.0035 | 0.0015 |

TABLE XVIII. **PAKDD2010 BALANCED DATASET.** ***Execution times*** *in seconds from 10 repetitions of k-fold cross-validation with* ***SVM with MLP kernel***

| **Dummy trans.** | | | **WoE trans.** | |
|---|---|---|---|---|
| | **k=2** | | **k=2** | **k=4** |
| **Min** | 713.34 | **Min** | 254.66 | 851.89 |
| **Max** | 968.45 | **Max** | 274.56 | 1054.31 |
| **Mean** | 803.70 | **Mean** | 262.75 | 954.75 |
| **StDev** | 82.49 | **StDev** | 6.22 | 88.40 |

## VII. Conclusion

In this study we have proposed a data transformation method based on the weight of evidence parameter. This technique is applicable in binary and multivariate supervised learning problems particularly for the nominal and categorical features. We have tested this technique on two real datasets. To verify our results, we have also generated dummy features from all nominal features in the same datasets and afterwards we have trained the same machine learning algorithms (i.e. feed forward neural networks and support vector machines with different kernels).

The analysis of the results show that in datasets in which the number of instances, the number of nominal features and the number of different values are fairly small, both transformations provide similar results in terms of predictive performance and execution time. For this reason in such cases we recommend applying the dummy transformation, because is easier to implement and it is a lot simpler to interpret and understand. However, it was the opposite case when we applied both transformations to a significantly larger dataset (with more than 10000 instances) that has more nominal features and they have more different values. Very often machine learning algorithms could not be trained on the dummy transformed dataset because of the memory complexity or because convergence could not be achieved. On the same dataset, but transformed with WoE, the same algorithms achieved convergence and produced significantly better results both in terms of predictive performance and execution time.

The presented method can be used without prior knowledge of the nature of the datasets. In our future work, we plan to compare the WoE transformation with other techniques for data transformations for nominal features. Also we will train a larger set of classifiers and will apply these transformations on other datasets as well. Additionally when we train the classifiers we also need to investigate the effect of parameter tuning, for instance when training SVMs. Our goal is to make firm and well-explained conclusions of which transformation is most suitable for what kinds of datasets, so researchers and practitioners can make apply these transformations more confidently without having to try out all possible combinations before deciding which one is most suitable.

## Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss.Cyril and Methodius University, Skopje, Macedonia.

## References

[1] C. Shearer, "The crisp-dm model: the new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–19, 2000.

[2] R. Anderson, *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford: Oxford University Press, 2007. ISBN 9780199226405

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. doi: 10.1145/1656274.1656278. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[4] E. Tuv and G. Runger, "Scoring levels of categorical variables with heterogeneous data," *Intelligent Systems, IEEE*, vol. 19, no. 2, pp. 14–19, Mar 2004. doi: 10.1109/MIS.2004.1274906

[5] M. Hofmann and R. Klinkenberg, Eds., *RapidMiner: data mining use cases and business analytics applications*, ser. Chapman & Hall/CRC data mining and knowledge discovery series. Boca Raton: CRC Press, 2014, no. 33. ISBN 9781482205497

[6] T. W. Miller, *Modeling techniques in predictive analytics: business problems and solutions with R*. Upper Saddle River, New Jersey: Pearson Education, Inc, 2014. ISBN 9780133412932

[7] M. Deza, *Encyclopedia of distances*. Dordrecht : New York: Springer Verlag, 2009. ISBN 9783642002335

[8] D. W. Goodall, "A new similarity index based on probability," *Biometrics*, vol. 22, no. 4, pp. pp. 882–907, 1966. [Online]. Available: http://www.jstor.org/stable/2528080

[9] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 14, no. 4, pp. 673–690, Jul 2002. doi: 10.1109/TKDE.2002.1019208

[10] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004. doi: 10.1108/00220410410560582. [Online]. Available: http://dx.doi.org/10.1108/00220410410560582

[11] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Trans. Inf. Syst.*, vol. 26, no. 3, pp. 13:1–13:37, Jun. 2008. doi: 10.1145/1361684.1361686. [Online]. Available: http://doi.acm.org/10.1145/1361684.1361686

[12] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*, ser. Lecture Notes in Computer Science, C. Nédellec and C. Rouveirol, Eds. Springer Berlin Heidelberg, 1998, vol. 1398, pp. 137–142. ISBN 978-3-540-64417-0. [Online]. Available: http://dx.doi.org/10.1007/BFb0026683

[13] I. J. Good, *Probability and the Weighing of Evidence*. C. Griffin & Co., London, UK, 1950.

[14] E. P. Smith, I. Lipkovich, and K. Ye, "Weight-of-evidence (woe): Quantitative estimation of probability of impairment for individual and multiple lines of evidence," *Human and Ecological Risk Assessment: An International Journal*, vol. 8, no. 7, pp. 1585–1596, 2002. doi: 10.1080/20028091057493. [Online]. Available: http://dx.doi.org/10.1080/20028091057493

[15] E. Zdravevski, P. Lameski, and A. Kulakov, "Weight of evidence as a tool for attribute transformation in the preprocessing stage of supervised learning algorithms," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, July 2011. doi: 10.1109/IJCNN.2011.6033219. ISSN 2161-4393 pp. 181–188.

[16] E. Zdravevski, P. Lameski, A. Kulakov, and D. Gjorgjevikj, "Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets," in *Computer Science and Information Systems (FedC-*

*SIS), 2014 Federated Conference on,* Sept 2014. doi: 10.15439/2014F500 pp. 387–394.

[17] D. B. Suits, “Use of dummy variables in regression equations,” *Journal of the American Statistical Association,* vol. 52, no. 280, 1957.

[18] M. A. Hardy, *Regression with dummy variables,* ser. *Sage university papers series.* Newbury Park: Sage Publications, 1993, no. no. 07-093. ISBN 0803951280

[19] N. Chater and M. Oaksford, Eds., *The probabilistic mind: prospects for Bayesian cognitive science.* Oxford ; New York: Oxford University Press, 2008. ISBN 9780199216093

[20] E. Zdravevski, P. Lameski, and A. Kulakov, “Towards a general technique for transformation of nominal features into numeric features in supervised learning,” in *Proceedings of the 9th Conference for Informatics and Information Technology (CIIT 2012).* Faculty of Computer Science and Engineering (FCSE) and Computer Society of Macedonia, 2012.

[21] E. L. Allwein, R. E. Schapire, and Y. Singer, “Reducing multiclass to binary: A unifying approach for margin classifiers,” *The Journal of Machine Learning Research,* vol. 1, pp. 113–141, 2001.

[22] R. Rifkin and A. Klautau, “In defense of one-vs-all classification,” J. Mach. Learn. Res., vol. 5, pp. 101–141, Dec. 2004. [Online]. Available: http: //dl.acm.org/citation.cfm?id=1005332.1005336

[23] E. Zdravevski, P. Lameski, and A. Kulakov, “Advanced transformations for nominal and categorical data into numeric data in supervised learning problems,” in *Proceedings of the 10th Conference for Informatics and Information Technology (CIIT 2013).* Faculty of Computer Science and Engineering (FCSE) and Computer Society of Macedonia, 2013.

[24] K. Bache and M. Lichman, “UCI machine learning repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[25] J. Huang, J. Lu, and C. Ling, “Comparing naive bayes, decision trees, and svm with auc and accuracy,” in *Data Mining,* 2003. ICDM 2003. Third IEEE International Conference on, Nov 2003. doi: 10.1109/ICDM.2003.1250975 pp. 553–556.

[26] J. Huang and C. Ling, “Using auc and accuracy in evaluating learning algorithms,” *Knowledge and Data Engineering, IEEE Transactions on,* vol. 17, no. 3, pp. 299–310, March 2005. doi: 10.1109/TKDE.2005.50

[27] C. Ferri, J. Hernndez-Orallo, and R. Modroiu, “An experimental comparison of performance measures for classification,” *Pattern Recognition Letters,* vol. 30, no. 1, pp. 27 – 38, 2009. doi: http://dx.doi.org/10.1016/j.patrec.2008.08.010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865508002687

[28] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence* - Volume 2, ser. IJCAI’95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. ISBN 1-55860-363-8 pp. 1137–1143. [Online]. Available: http://dl.acm.org/citation.cfm?id=1643031.1643047

[29] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.,* vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968

[30] “Pacific-asia knowledge discovery and data mining competition 2010,” http://sede.neurotech.com.br/PAKDD2010/, accessed: 2015-06-05.

[31] D. Olson, “Data set balancing,” in *Data Mining and Knowledge Management, ser. Lecture Notes in Computer Science,* Y. Shi, W. Xu, and Z. Chen, Eds. Springer Berlin Heidelberg, 2005, vol. 3327, pp. 71–80. ISBN 978-3-540-23987-1. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-30537-8 8

DOI: 10.15439/2015F89
ACSIS, Vol. 5

# Parallel computation of information gain using Hadoop and MapReduce

Eftim Zdravevski*, Petre Lameski†, Andrea Kulakov‡, Sonja Filiposka§, Dimitar Trajanov¶, Boro Jakimovski‖
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: *eftim.zdravevski@finki.ukim.mk, †petre.lameski@finki.ukim.mk, ‡andrea.kulakov@finki.ukim.mk,
§sonja.filiposka@finki.ukim.mk, ¶dimitar.trajanov@finki.ukim.mk, ‖boro.jakimovski@finki.ukim.mk

***Abstract*—Nowadays, companies collect data at an increasingly high rate to the extent that traditional implementation of algorithms cannot cope with it in reasonable time. On the other hand, analysis of the available data is a key to the business success. In a Big Data setting tasks like feature selection, finding discretization thresholds of continuous data, building decision threes, etc are especially difficult. In this paper we discuss how a parallel implementation of the algorithm for computing the information gain can address these issues. Our approach is based on writing Pig Latin scripts that are compiled into MapReduce jobs which then can be executed on Hadoop clusters. In order to implement the algorithm first we define a framework for developing arbitrary algorithms and then we apply it for the task at hand. With intent to analyze the impact of the parallelization, we have processed the FedCSIS AAIA'14 dataset with the proposed implementation of the information gain. During the experiments we evaluate the speedup of the parallelization compared to a one-node cluster. We also analyze how to optimally determine the number of map and reduce tasks for a given cluster. To demonstrate the portability of the implementation we present results using an on-premises and Amazon AWS clusters. Finally, we illustrate the scalability of the implementation by evaluating it on a replicated version of the same dataset which is 80 times larger than the original.**

***Keywords*—*Hadoop, MapReduce, information gain, parallelization, feature ranking***

## I. Introduction

THE volume of data that needs to be processed has increased significantly in recent years. Most of the organizations in the world base their decisions on the data they collect and they need large volumes of data to be processed in as little time as possible. Over the years many ideas have been developed for solving the Big Data challenge. Increasing the processing power is the logical way to go but this has proven to be effective up to a certain point. After that the hardware scaling is not yet effective enough. The idea of distributing the computation has become popular in recent years since the publications of Google's approaches for MapReduce [1] in 2004 and the concept of Big Table [2] in 2006. Other companies have followed similar paths introducing open-source solutions. One such system is Apache Hadoop that contains a set of algorithms for distributed processing, storage of large datasets on computer clusters, scheduling etc. It is a framework that is employed by industry leaders like Yahoo, Facebook, Ebay, Adobe, etc [3].

Machine learning algorithms such as decision trees [4], neural networks [5], Naive Bayes [6, 7] and many others can automatically analyze data and make conclusions, predictions or even find patterns that otherwise cannot be detected. The main drawback of these algorithms is the degrading performance in presence of redundant and irrelevant features. Other algorithms such as Support Vector Machines are able to cope with this problem to some extent, however, this ability increases the computational time so much that the algorithm doesn't give result in reasonable time. This has already been confirmed in the literature [8, 9, 10]. One way to resolve this is to perform feature selection [11, 9, 12], defined as the task of selection of feature subsets that describe the hypothesis at least as well as the original set. In [13] the most widely used methods for feature selection are introduced.

The rest of this paper is organized as follows. First, in section II we review the most recent approaches to parallelization of various algorithms. Afterwards in section III we describe the definition and applications of information gain. Next, in section IV we present we describe the services in the Hadoop ecosystem and then we present a framework for parallelization of algorithms. Thereupon, in section V we apply it for parallel and distributed computation of information gain based on MapReduce. Next, in section VI we present the experimental setup and the obtained results. Finally, in section VII we discuss the contribution of our work and our plans for further research.

## II. Related work

This section describes some of the most recent work on parallelizing different algorithms with MapReduce. The general approaches and limitations of different data mining algorithms when applied to massive datasets are described in [14]. Here some common data mining problems are explained from a Big Data perspective, but a MapReduce implementation is given only for some common problems like matrix manipulation and joins between tables.

A good overview of the parallel programming paradigms and frameworks in the Big Data era is presented in [15]. Here the authors describe the MapReduce paradigm, but more importantly introduce the frameworks that are built on top of it like: Pig Latin for processing data flows, Hive for non-real time querying of partitioned tables, and Spark and Twister for iterative parallel algorithms.

The authors in [16] address the problem of efficient feature evaluation for logistic regression on very large data sets. Here they present a new forward feature selection heuristic

978-83-60810-66-8/$25.00©2015, IEEE

that ranks features by their estimated effect on the resulting model's performance. They test the method on already available datasets from UCI, but also generate artificial datasets for which they know the logistic regression coefficients. They use that to evaluate the selected features.

By using the MapReduce paradigm in [17] a data intensive parallel feature selection method is proposed. In each map node, a method is used to calculate the mutual information and combinatory contribution degree is used to determine the number of selected features.

In [18] an implementation based on the MapReduce programming model of Naive Bayes is proposed. During the map phase all counts needed for calculating the conditional probabilities are emitted, and during the reduce phase they are aggregated.

A parallel implementation of the SVM algorithm for scalable spam filtering using MapReduce is proposed in [19]. By distributing, processing and optimizing the subsets of the training data across multiple participating nodes, the distributed SVM reduces the training time significantly. Merging of the results is actually a union of the individually computed support vectors. The cost of the parallelization is that because not all training data is available on all nodes, the performance can degrade. However, if the data is properly distributed on the nodes in regards to stratification per class, this problem can be mitigated.

A method for reducing the dataset to a small but representative subset is proposed in [20]. The idea is to use the representative subset for faster machine learning because the dataset size will be significantly reduces. The speedup is being calculated against a cluster with one node. However, if the dataset is too large, or the computation takes a lot of time the authors suggest to use more than one for estimating the speedup. By doing this one can calculate the speed of the current configuration versus the cluster with some smaller number of nodes.

In [21] an approach based on MapReduce for distributed column subset selection is proposed. In this approach each node has access to a random subset of features. This approach has a limitation that the datataset has to be manually splitted and the MapReduce jobs need to be written on lower level. The reason for this is that HBase segments the data horizontally by rows, so either the dataset needs to be transposed or to manually start different jobs and not to rely on a higher level language like Pig Latin or Hive.

Authors in [22] propose a wrapper approach for parallel feature selection. Here features are added to the selected set if after their addition, the performance of the classifier does not degrade. Then in a second phase from the subset obtained in the previous step, features are removed if their discarding does not degrade the classifier performance.

Apache Mahout [23] is an environment for quickly creating scalable performant machine learning applications based on MapReduce. Even though there are plenty of algorithms available in it, at the time of this writing, only two algorithms related to feature selection and dimensionality reduction in Mahout are available: Singular Value Decomposition (SVD) and Stochastic SVD.

## III. Information gain and its applications

Information gain is a synonym for KullbackLeibler divergence and has variety of applications. Very often it is used for ranking individual features as described in [24, 25]. The research discussed in [26] shows how information gain can be used for feature selection in text categorization problems. Authors in [27] propose using the information gain for discretization of continuous valued features into discrete intervals. In like manner, in [28] information gain is analyzed as an unsupervised method for discretization of continuous features. Likewise, in [29] it is applied for improving decision tree performance by prior discretization of continuous-valued attributes. In fact these papers have inspired many other researchers to propose various other applications based on the information gain and entropy. In [30] the information gain in conjunction with methods based on particle filters is used for exploration, mapping, and localization. Another application of information entropy for extending the rough set based notion of a reduct is proposed in [31]. There it is applied for calculation of minimal subsets of features keeping information about decision labels at a reasonable level.

In the remaining of this section when describing the information gain we use the notation we have also used in [32]. In order to calculate the information gain, first the entropy $H(X)$ of the dataset should be calculated. Let $X$ denote a set of training examples, and each of them $x_i$ is in the form $(x_i^1, x_i^2, ..., x_i^k, y_i)$. Let each column (i.e. feature) be a discrete random variable that takes on values from set $V^j, j = 1..k$. Let the set of possible labels (i.e. classes) is $L$, such as $y_i \in L$. Then the entropy of the dataset $X$ can be calculated with equation (1), where $p(l)$ is the probability of instance $x_i$ to be labeled as $l$ (i.e $y_i = l$) and is defined with equation (2).

$$H(X) = -\sum_{l \in L} p(l) \log p(l) \tag{1}$$

$$p(l) = \frac{|\{x_i \in X | y_i = y\}|}{|X|} \tag{2}$$

The information gain of the $j$-th feature of the dataset $X$ can be calculated with equation (3), where first part in the sum is the probability of the instance $x_i$ to have value $v$ of the $j$-th feature. The second part in the sum in equation (3) denotes the entropy of the subset of instances of $X$ that have the value $v$ of the $j$-th feature.

$$IG(X, j) = H(X) - \sum_{v \in V^j} \frac{\left|\left\{x_i \in X | x_i^j = v\right\}\right|}{|X|} H(\left\{x_i \in X | x_i^j = v\right\}) \tag{3}$$

As shown by equations (1),(2) and (3), calculation of information gain of all features boils down to counting the number of instances per feature, value and class. After we compute these counts, we can calculate the probabilities and consequently calculate the information gain. In section V we propose parallel implementation for calculating the information gain of each feature $j$ in the dataset $X$.

## IV. Framework for parallelization

Parallelization of algorithms introduces a handful of potential software bugs of usually related to race conditions, communication and synchronization between the different subtasks. Owing to that, writing parallel computer programs is more challenging than writing sequential ones. In Hadoop most of those challenges are already addressed by various mechanisms and services, so when it is used as a platform for implementation of algorithms, the programmer does not have to put much effort for solving those kinds of issues. Before explaining the details we want to point out that the proposed framework uses the principle of data-parallelism. Having that in mind, the same principles could be used in a regular SQL environment. Nevertheless, while many of the limitations and benefits of SQL vs NoSQL are much argued in the research community, the scalability properties of NoSQL databases are undisputed.

Given that understanding of the Hadoop ecosystem is essential for understanding our parallelization framework, first in the next subsection IV-A we review its services. Then in the following subsections we describe the several phases of the proposed framework. We have applied similar logic in [32], but the solution was not a generic one and was custom for the task at hand. On Fig. 1 is shown a general overview of the data flow during these phases. For data partitioning we propose using HBase tables which are pre-splitted for optimal data distribution, where as for parallel processing and writing MapReduce jobs we suggest using Pig Latin with appropriate user-defined functions.

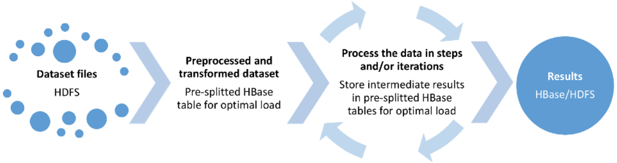

Dataset files
HDFS
Preprocessed and transformed dataset
Pre-splitted HBase table for optimal load
Process the data in steps and/or iterations
Store intermediate results in pre-splitted HBase tables for optimal load
Results
HBase/HDFS

Fig. 1. Data flow phases for processing HBase tables with Pig Latin

### A. Hadoop

The MapReduce [1, 33] paradigm is essential to the distributed computation and storage that Hadoop achieves. It consists of two phases: map and reduce. The first phase, map, splits the data into subsets. The reduce phase, aggregates the result from the output that the map phase produces. Procedures that can be performed in the map phase are: filtering, sorting, projecting and reading the data. The map phase returns an intermediate result consisted of keys and values. The reduce procedures use this data and perform aggregation. Hadoop delegates the data from the map phase to the reduce procedures. The MapReduce simplicity makes it very efficient for large-scale implementations on thousands of nodes.

Hadoop with its different services provides all the logistics and monitoring for the processes like scheduling, distribution, communication and data transfer, and also provides redundancy and fault tolerance. Many services or subsystems exist in Hadoop, but the most notable are: YARN (MapReduce2), HDFS and HBase [34][35].

YARN (Yet Another Resource Negotiator) [36] takes care of job scheduling, monitoring and resource management. Two separate daemons are responsible for these tasks: a global ResourceManager and per-application ApplicationMaster. The ResourceManager deploys resources among all the applications and the per-application ApplicationMaster negotiates for resources with the ResourceManager and works with the NodeManager to execute tasks and perform monitoring. YARN does the resource allocation and the distribution of MapReduce jobs to the apropriate nodes.

Hadoop Distributed File System (HDFS) [37] provides scalable, fault-tolerant, distributed storage system that works closely with MapReduce. It was designed to span large clusters of commodity servers. An HDFS cluster is consisted of a NameNode and DataNodes. The NameNode is responsible for the cluster metadata and DataNodes are responsible for data storage. The data is usually split into large blocks (typically 128 megabytes), independently replicated across multiple DataNodes.

HBase is an open source, non-relational, distributed database modeled after Google's BigTable. It runs on top of HDFS (Hadoop Distributed Filesystem), providing BigTable-like capabilities for Hadoop [38, 39, 40]. HBase is a NoSQL (Not Only SQL) database in which the tables are designed by analyzing usage patterns. This allows simplicity of design, horizontal scaling, and finer availability control. The data structures in NoSQL databases, such as HBase, allow faster executions of some operations than the execution of similar operations in relational databases. This mostly depends on the problem that must be solved. Tables in HBase can be used as the input and output for MapReduce jobs run in Hadoop. According to Eric Brewers CAP theorem, HBase is a CP type system (i.e. Consistent and Partition tolerant) [41].

The MapReduce programming model is very popular due to its simplicity. The extreme simplicity of MapReduce leads to much low-level coding that needs to be done for some operations that are much simpler when using relational databases. This increases development time, introduces bugs and may obstruct optimizations [42]. A group at Yahoo motivated by these repeatable tasks on daily basis, has developed a scripting language called Pig Latin. Pig is a high-level dataflow system that is a compromise between SQL and MapReduce. Pig offers constructs for data manipulation similar to SQL, which can be integrated in an explicit dataflow. Pig programs are compiled into sequences of MapReduce jobs, and executed in the Hadoop MapReduce environment [43].

### B. Loading data into HDFS

This is the first and most simple phase. This phase should be performed once or multiple times, depending on how the dataset is structured. The most common formats for datasets are:

- CSV (comma separated values). This format is usually used to store dense datasets.
- ARFF (Attribute-Relation File Format). Also used to store dense datasets.
- EAV (Entity Attribute Value). Used to store sparse matrices that have a lot of zeros and some non-zero

elements.

If the dataset is only one file then it will be copied from the Linux File System to HDFS using a simple command. This means that for such cases during this step we cannot have parallelism. However if the dataset is dispersed into multiple files, then all of them can be copied simultaneously to HDFS. Be that as it may, this step usually is very fast compared to the following steps for machine learning, so its parallelization may not be necessary at all.

### *C. Facilitating data parallelism with HBase*

After the previous step IV-B is finished the dataset files reside on HDFS. As it is extensively described in [37], each file in HDFS is replicated across several nodes for reliability. A typical file in HDFS is gigabytes to terabytes in size, splitted in blocks of 128 MB by default. If the files are too small than that could degrade the performance of the system and limit the level of parallelism. Map tasks usually process a block of input at a time. If the file is very small and there are a lot of them, then each map task processes very little input, and there are a lot more map tasks, each of which imposes extra bookkeeping overhead. Ideally the dataset will be one large file dispersed on multiple blocks on HDFS so when loaded, transformed and stored into HBase, greater parallelism can be achieved. Be that as it may, datasets are not always so large that HDFS can distribute them on all nodes and get optimal parallelization. One way to mitigate this is by splitting the dataset in multiple smaller files and store them in one folder, so later the Pig script can read from all files in the folder instead of a specific file. Nevertheless, this step again is usually very fast especially compared to the steps that comprise the actual algorithm, so we do not recommend to spend too much time on optimizing the file sizes for better parallelism.

Even though we can achieve parallelism while processing files stored on HDFS, the control of degree of parallelism is difficult, more involved and at very low-level. For instance, we if we have a large file then HDFS will automatically partition it and distribute it on different nodes. Be that as it may, we do not have control of how HDFS will do this, on how many partitions it will store it, where are they going to be distributed, etc. On the opposite side, if we have a small dataset file, it will not be partitioned at all. To have a better control on this one would need to manually split the file in the desired number of chunks and then let HDFS distribute them. Moreover, this process has to be repeated again and again if we have continuous stream of data.

On the other hand, HBase offers many other services built on top of HDFS, among which is a much better control of the degree of parallelism. This is due to the fact that the data in HBase is stored in a structured manner, while having various mechanisms that simplify random reads and writes from rows and columns. Namely, HBase tables are divided into potentially many regions, while one or more regions are serviced by a region server. The tables can be horizontally and vertically segmented while they are physically stored in HBase. Because many machine learning applications access the data by rows, in this paper we will continue to discuss only horizontal segmentation. As HBase was designed with very large tables in mind, a common use case is the following. A table at creation has only one region, which is serviced by one region server (a physical node in the Hadoop cluster). When this table is loaded with data it gets bigger and at some point it will become too big, so HBase will split its region into two regions. Then the new region will be assigned to the same region server or can be moved to another region server. The default splitting threshold is 10 GB. There are numerous reasons why HBase was designed that way, and we will not go into details about that. From parallelization perspective, this can pose a challenge, because for the automatic splits there are no guarantees that every region will contain equal amount of data, when are the splits going to occur exactly, are the regions going to be served by different region servers (nodes) etc. Further more, if one is using Hadoop for research purposes only then the dataset may not be that large, thus never overcoming the threshold for splitting. To overcome this challenge we can pre-split the tables on creation. This in turn means that the table can be configured at creation time to be stored on as many-regions as needed. Usually the number of regions is a multiple of the number of HBase region servers. The logic for having more region servers than acutal nodes is because the nodes are multi-core machines, so different threads on the same node can service different regions.

Before loading the dataset in HBase, we need to define the table structure and create it. Column names and data types are provided when storing data in each row, so at creation time we need to only specify a table name and a column family. There are some advanced configuration features that can be specified, but they are not topic of this discussion. Be that as it may, there is one very important decision that we need to make before loading data in the table. Because HBase tables, unlike SQL tables, cannot have secondary indexes, the primary key (row key) needs to be designed according to the usage patterns of the table. There are many considerations when designing the row key and they are very important for production use of HBase tables. However, for scientific use and for parallelizing machine learning algorithms, we need a simple design that allows uniform data distribution across nodes. In most scientific datasets the data instances (i.e. rows) do not have ids for their instances, or if they do they are not used for the actual machine learning. Nevertheless, in order to store a row in a HBase table, it needs a row key. For flat flies like CSV or ARFF the row key can be the line number of the instance. However, sequential row keys are very bad choice for HBase tables because the inserts will always be on the last region, therefore having no parallelism during the load, a problem called Region Server hotspotting. There are multiple ways of overcoming this problem, and one of them is a technique called salting [38]. With this technique each sequential id is salted with a prefix. The prefix is usually the modulo number of the original sequential id and the number of regions. Even though, this is very important topic, the step of loading the datasets in HBase is not the focus point of this paper.

Once the dataset files are loaded into HDFS we need to transform them if needed and store them in HBase. If we have totally $M$ rows in the dataset, and $R$ regions, then we would like to distribute the data uniformly so each region gets $M/R$ rows. This in turn means that we need to specify $R-1$ split points when creating the table. If we use sequential ids for the row key (like the line number in the file), than these split points would be: $M/R, 2M/R, 3M/R, ..., (R-1)M/R$. If we use a more sophisticated row key design, then the split points should reflect that design. For instance, if we take the modulo number of the id and the number of regions, then each region would get almost the same number of rows. This design of the row key allows fast random reads and writes, and additionally it facilitates addition of new data to the table at a later time without needing to redesign the table for equally dispersed load across regions. The following example shows how a table can be pre-splitted on creation. The row key design is described with the function in listing 1. It returns a tuple in which the first element is the padded modulo number and the second part is the padded sequential id. The numbers are padded with zeros so that they are lexicographically sorted.

```
(pad(seq_id % num_regions), pad(seq_id))
```

Listing 1. Row key design

Once the HBase table that should contain the dataset is created with appropriate split points for even data distribution across the cluster, we can start loading the data. One can write pure MapReduce jobs in Java or Python. If we choose that path, then we need to write a separate map and reduce function for each task. However, by using the scripting language Pig Latin [42, 43] we can write scripts from a higher-level perspective. These Pig scripts generate MapReduce tasks in the background so the programming effort is simplified and the development time is greatly reduced. The downside of using Pig is that when Pig scripts are compiled into MapReduce jobs, there is some overhead. Additionally one may write a more optimal implementation of map and reduce functions manually than the ones generated by the Pig compiler. Nevertheless, these are corner cases and for longer running MapReduce tasks the overhead is insignificant in the range of up to couple of minutes in the worst case. When loading the data usually only a map phase is required. It reads the data from the HDFS files and stores it in HBase tables. In most cases when loading the data there is no grouping of keys, so a reduce phase is not needed. During this step we can add various methods for data preprocessing like discretization, transformation and other methods that rely only on the value in one row of the dataset.

### D. Processing HBase tables

After the dataset is loaded in a HBase table we can continue implementing machine learning algorithms. In general, this phase can be comprised of several substeps or iterations of data processing, depending on the nature of the algorithm that is being implemented. For each of the intermediate steps that we need to store some data we need a HBase table that will also be pre-split at creation time, similar to what we described in the previous subsection IV-C. What happens in the background when a particular HBase table is being processed, is very peculiar. Pig will determine the number of regions of the table and it will start that number of map tasks. Then each map task will process the data of a particular table region on the node where the data resides, therefore leveraging data locality. Note that, in order to benefit from the principle of data locality, each node in the cluster should run HDFS, HBase and YARN (MapReduce) services. The number of reduce tasks is by default one, but this can be also manually specified and does not depend of the table structures. If we specify more than one reduce tasks, then this will solicit a merge phase which will combine the intermediate results from each reduce task. Usually, for smaller datasets specifying more than one reduce task does not improve performance, on the contrary, it can degrade performance. However, being able to specify the number of reduce tasks provides flexibility that can improve the performance for larger datasets or some specific problems.

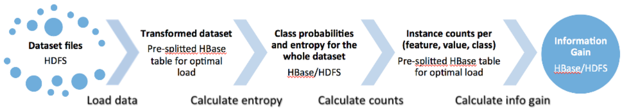

Dataset files
HDFS
Load data
Transformed dataset
Pre-splitted HBase table for optimal load
Calculate entropy
Class probabilities and entropy for the whole dataset
HBase/HDFS
Calculate counts
Instance counts per (feature, value, class)
Pre-splitted HBase table for optimal load
Calculate info gain
Information Gain
HBase/HDFS

Fig. 2. Data flow during the parallel computation of information gain

During this phase, depending on the type of algorithm that is being parallelized, many tables can be used by multiple MapReduce jobs. In the following section V we illustrate this by when we parallelize the computation of information gain.

## V. Parallel computation of information gain

In this section we illustrate how we can use the framework proposed in section IV for computing the information gain of a dataset. Fig. 2 shows the data flow during the steps of the framework. The first steps of loading the data to HDFS and subsequently into HBase tables corresponds to what we explained in subsections IV-B and IV-C. Then, after the dataset is loaded in a HBase table, the processing takes place in three phases explained in the following subsections and they correspond to the specific properties of the current algorithm. This is an illustration how the framework step described in subsection IV-D can contain multiple phases consisted of many different MapReduce tasks that read and store data from several HBase tables.

### A. Calculating entropy of a dataset

As explained in section III, the definition of information gain requires calculation of the entropy of the whole dataset. In order to calculate it, first we need to count the number of instances per class, and afterwards to sum the class probabilities. Notably this is a very simple step and does not require parallelization because its complexity is $O(N)$, where $N$ is the number of instances in the dataset. Moreover, if we are interested in only sorting the features without having the actual information gain for each of them, then we can eliminate the entropy from equation (3). Be that as it may, for other applications we need the actual information gain. If we decide to parallelize this step, despite of its simplicity, then we need two MapReduce jobs. The Pig Latin script shown in listing 2 performs this. Each parameter starting with *$* can be passed to Pig script when it is started. In line 2 we specify that we only want to read the cell where the class is stored, denoted as $\$label$. The first MapReduce job that calculates the counts per class and the class probabilities corresponds to the code from line 2 to line 16. Then from line 17 till the end of the script the second MapReduce job calculates the entropy of the dataset. Notwithstanding, the peculiar thing that is demonstrated here is how easy MapReduce jobs can be combined into a flow. In like manner, one can combine many MapReduce jobs in one flow without any need of manual synchronization between them.

```
register '$udf_path' as paddingUDFs;
pfdata_tmp = LOAD '$table_dataset' USING
    org.apache.pig.backend.hadoop.
hbase.HBaseStorage('r:$label',
       '-loadKey=true'
) AS (rowkey:tuple(prefix_padded:chararray,
    id_padded:chararray),
  class:int);
pfdata_class_group = GROUP pfdata_tmp BY
    class;
pfdata_class = FOREACH pfdata_class_group
    GENERATE
  flatten(group) as class,
  COUNT(pfdata_tmp.class) as count;
pfdata_class_prob = FOREACH pfdata_class
    GENERATE
  class,
  count,
  (count/$num_instances) as prob:double,
  ((-count/$num_instances)*
  UDFs.log2(count/$num_instances)) as
      entropy:double;
total_entropy_group = GROUP pfdata_class_prob
    ALL;
total_entropy = FOREACH total_entropy_group
    GENERATE
  SUM(pfdata_class_prob.entropy) as
      entropy:double;
STORE total_entropy INTO
    '$hdfs_export_entropy' USING
    PigStorage('\t');
```

Listing 2. Pig script for calculating entropy of a dataset

### B. Counting instances per feature index, feature value and class

After the entropy is calculated, the definition of information gain, as presented with (3), requires counts of instances per feature index, feature value and class. This step is the most computationally expensive step in the algorithm. The source code of this step is shown in listing 3 and it is based on the pseudo code we have reported in [32]. Parameters that are passed to the script are the table names, number of features, index of the class value, number of padding digits, etc. First we need to load the dataset from a HBase table (lines 3 through 6). A row of the dataset is represented by the row key, and the dictionary $r$ in which keys are the column names and values are the actual values. This representation allows us to store only the non-zero values of a dataset. Then we need to expand each row of the dataset (denoted as dictionary $r$) to tuples: *(feature index, feature value, class, 1)*. This is performed in lines 7 through 8 with the user-defined function *decode_sparse_row*. If the dataset has $M$ rows (instances) and $N$ columns (features), then from each row we will generate $N$ tuples because now also the zero-valued cells are also included. To summarize, when the whole dataset is processed $M \times N$ tuples will be generated. These tuples are afterwards grouped by the key *(feature index, feature value, class)* in lines 9 through 12, and finally the count is stored in another table (lines 13 through 15). All of the code in listing 3 is compiled in one MapReduce job. The number of generated map tasks will be equal to the number of regions of the input table (denoted by *$table_dataset* in the script), and the number of reduce tasks is set by the parameter *$parallel*.

```
1  register '$udf_path' using jython as UDFs;
2  set default_parallel $parallel;
3  pfdata_tmp = LOAD '$table_dataset' USING
       org.apache.pig.backend.hadoop.hbase.
4    HBaseStorage('r:*', '-loadKey=true') AS
5      (rowkey:tuple(prefix_padded:chararray,
          id_padded:chararray),
6      r:map[]);
7  pfdata_short = FOREACH pfdata_tmp GENERATE
8    FLATTEN(UDFs.decode_sparse_row(r,
        $num_features,
9    $num_features_digits, '$feature_data_type',
        '$label'));
10 feature_value_class_counts_group = GROUP
       pfdata_short BY (feature_index,
       feature_value, class);
11 feature_value_class_counts = FOREACH
       feature_value_class_counts_group GENERATE
12   group as rowkey,
13   SUM(pfdata_short.instance_count) as
        instance_count;
14 STORE feature_value_class_counts INTO
       '$table_feature_index_tmp' USING
15     org.apache.pig.backend.hadoop.hbase.
16     HBaseStorage('r:instance_count');
```

Listing 3. Counting number of instances per feature index, feature value and class with Pig Latin

### C. Calculating the information gain

Having the counts calculated in the previous step V-B, this step only calculates the probabilities and entropies in (3) and stores this result in HBase or HDFS. Nevertheless, it is usually the second longest running step from this list. The code for this is shown in listing 4. First in the lines 3 through 8 it reads the tuples *(feature index, feature value, class, instance_count)* which were calculated in the previous step, and now are stored in the table *$table_feature_index_tmp*. This table was properly pre-splitted on creation so the Pig script will be compiled in one MapReduce job with multiple map tasks. In particular, if the number of features is $N$ and the desired number of regions of the table is $R$, then we specify $R-1$ split points, and in that way each region will contain the tuples for $N/R$ features. We acknowledge that this might not be ideal distribution because some features might have significantly more distict values then others, but nevertheless, it provides decent parallelism. Given that this step is not as computationally intensive as the previous, we did not consider it necessary to further optimize this table. Then when the MapReduce job is compiled, it will have $R$ map tasks and each of them will work with the data of the appropriate table region.

```
1  register '$udf_path' using jython as UDFs;
2  set default_parallel $parallel;
3  feature_value_class_counts_tmp = LOAD
       '$table_feature_index_tmp' USING
       org.apache.pig.backend.hadoop.hbase.
4  HBaseStorage('r:instanceCount',
5          '-loadKey=true'
6  ) AS (
7    id:tuple(feature_index:chararray,
        feature_value:int, class:int),
8    instanceCount:double);
9
```

```
10 feature_value_class_counts = FOREACH
       feature_value_class_counts_tmp GENERATE
11 flatten(id) as (feature_index, feature_value,
       class),
12 instanceCount;
13 feature_index_group = GROUP
       feature_value_class_counts BY
14      (feature_index);
15 feature_index_info_gain = FOREACH
       feature_index_group GENERATE
16    flatten(group) as feature_index_padded,
17    flatten(UDFs.
18    calc_feature_info_gain(($entropy), group,
          feature_value_class_counts,
          ($num_instances))) as info_gain:double;
19 STORE feature_index_info_gain INTO
       '$table_feature_index_info_gain' USING
       org.apache.pig.backend.hadoop.
20 hbase.HBaseStorage('r:ig');
```

Listing 4. Calculating information gain with Pig Latin

The most peculiar part in the script in listing 4 is at line 13. Here all tuples are grouped by feature index. When the Pig Script is translated into a MapReduce job, the during the map phase the feature index is emitted as a key, and during the reduce phase all tuples that are for the same key (in this case the feature index) are grouped together on the same node. The Python UDF *calc_feature_info_gain* utilizes this because for each feature it has the count of instances of all its values per class. Having that it is easy to compute the information gain by (3). Finally, the results can be stored in a HDFS file or in a HBase table. In this script we store them in the HBase table $\$table\_feature\_index\_info\_gain$, performed in the last line.

## VI. Experiments

With intention to monitor various aspects of the parallel implementation, a relatively large dataset was essential. Furthermore, we did not want to focus on significant preprocessing like discretization or transformation of values so we can easily compare our results with other research. The FedCSIS AAIA'14 data mining competition dataset [44] has exactly those properties. It is a sparse matrix with 50000 instances and 11852 numeric features, most of which are have the value 0 or 1. There are about 0.9% non-zero values in it. It represents a multi-label problem that has 3 binary labels, that can be merged with the powerset technique as used in [45] into one one-label multi-class problem that has 8 ($2^3$) possible classes.

We have tested the same dataset on three completely different Hadoop clusters. Each of them was running the same version of Apache Hadoop 2.3.0 (integrated in Cloudera CDH 5.3.0). This is an extension to what we did in [32], where we analyzed the speedup only on one on-premises cluster. Additionally in this paper we analyze the effect of the number of reduce tasks, while in [32] we have used only one reduce task. Finally, the most important difference is that we have evaluated the scalability of the approach by replicating the dataset 80 times. We have performed this by replicating the dataset horizontally so from each instance there are 80 exact copies. This in turn results in a dataset that has 4 million instances and almost 12 thousand features. It should be noted that the computational complexity of the algorithm depends only on the dataset size and not on its sparsity or feature types.

Keeping in mind that our goal is to evaluate the execution time and speedup based on the cluster size, the expansion of the dataset serves this purpose.

The first cluster (denoted by *Amazon32* in the remaining of the paper) was deployed on Amazon AWS. It contained a total of 32 nodes, each of them a m1.xlarge instance with 15GB RAM and 8 compute units (4 cores with 2 compute units each). From the 32 nodes, 8 were hosting HBase Region Servers and HDFS Data Nodes, 3 were specifically dedicated to HDFS Data Nodes and 19 were running only YARN. We acknowledge that this configuration may not be optimal for the current task, but we were given access to this cluster without the ability to modify its configuration. Therefore we have decided to run tests using up to 8 nodes at a time, because when using more it would be difficult to estimate the speedup.

The second cluster (denoted by *FCSE24* in the remaining of the paper) was deployed on-premises at the Faculty of Computer Science and Engineering (FCSE) at the Ss.Cyril and Methodius University, Skopje, Macedonia. It had a total of 24 nodes, each of them an Intel Xeon Processor E5640 with 12M Cache, 2.66 GHz, 24 GB RAM, 4 cores and 8 threads. From them 21 were configured to run the following services: HBase Region Servers, HDFS DataNodes and YARN MapReduce NodeManagers. The remaining nodes were used for other Hadoop and Cloudera management services.

The third cluster (denoted by *FCSE65* in the remaining of the paper) was also deployed on-premises and it was an extended version of the second, containing a total of 65 nodes, of which 54 were running the following services: HBase Region Servers, HDFS DataNodes and YARN MapReduce NodeManagers. A variant of this cluster with 59 instead of 54 active nodes was also used for the experiments presented in [32].

During our tests none of these clusters was executing other tasks. On all of them we ran tests with different table structures in order to simulate clusters with smaller sizes. By pre-splitting the HBase tables to a specific number of regions we were able to force Pig Latin to start the desired number of map tasks for each job. For all these configurations we are computing the speedup of the parallelization against a cluster with one node. We are simulating the one-node cluster by configuring the tables to have only one region, thus all MapReduce jobs that read from those tables have only one map task. We have tested using different number of reduce task by setting a configuration property in the Pig scripts. The remaining of this section is divided in two, VI-A containing summary information for all steps that are fast and did not benefit significantly from the parallelization, and VI-B containing detailed information about the step described in V-B, which was the most computationally expensive. Table I shows the information gain of the top 50 features which can be used for verification of the correctness of our implementation. In the following subsections we describe the results from our experiments.

### A. Computationally cheap steps

The dataset was stored in two files: one containing the data in EAV (entity attribute value) format, and one containing the labels. The EAV format greatly reduces the file sizes to 72 MB compared to 1.1 GB when stored in full format as CSVs.

TABLE I. TOP 50 FEATURES ORDERED BY INFORMATION GAIN

| Rank | Feature | InfoGain | Rank | Feature | InfoGain |
|---|---|---|---|---|---|
| 1 | 11701 | 0.07422 | 26 | 7407 | 0.0256033 |
| 2 | 143 | 0.07000 | 27 | 11825 | 0.0249701 |
| 3 | 11832 | 0.06009 | 28 | 4505 | 0.0249698 |
| 4 | 1509 | 0.05154 | 29 | 11100 | 0.0249225 |
| 5 | 5909 | 0.04936 | 30 | 10331 | 0.0247915 |
| 6 | 8635 | 0.04539 | 31 | 7529 | 0.0247519 |
| 7 | 2182 | 0.04012 | 32 | 2274 | 0.0247061 |
| 8 | 865 | 0.03817 | 33 | 10261 | 0.0246147 |
| 9 | 6523 | 0.03817 | 34 | 7592 | 0.0245778 |
| 10 | 5827 | 0.03795 | 35 | 4319 | 0.0245677 |
| 11 | 5188 | 0.03467 | 36 | 1349 | 0.0245448 |
| 12 | 5513 | 0.03296 | 37 | 7405 | 0.0245288 |
| 13 | 6162 | 0.03294 | 38 | 11463 | 0.0245111 |
| 14 | 5967 | 0.03271 | 39 | 11000 | 0.0244753 |
| 15 | 2835 | 0.03223 | 40 | 6779 | 0.0240003 |
| 16 | 139 | 0.0318404 | 41 | 10428 | 0.0236240 |
| 17 | 9306 | 0.0318030 | 42 | 460 | 0.0235250 |
| 18 | 1772 | 0.0296594 | 43 | 7291 | 0.0233440 |
| 19 | 3257 | 0.0283169 | 44 | 8853 | 0.0232071 |
| 20 | 9848 | 0.0283169 | 45 | 2883 | 0.0232064 |
| 21 | 675 | 0.0282140 | 46 | 5925 | 0.0231852 |
| 22 | 73 | 0.0273487 | 47 | 8114 | 0.0225087 |
| 23 | 7275 | 0.0266788 | 48 | 5330 | 0.0223354 |
| 24 | 7419 | 0.0266100 | 49 | 1156 | 0.0219374 |
| 25 | 1244 | 0.0262854 | 50 | 2701 | 0.0218273 |

The effect is that copying them to HDFS is very fast (about a second). The step described in subsection IV-C was actually two MapReduce jobs. The first is for loading the labels which took 58 to 70 seconds, and the second for loading the data which took 130 to 145 seconds on the on-premises and 175 to 195 seconds on the Amazon cluster. Calculating the entropy of the dataset, described in section V-A, took 118 to 152 seconds on both clusters. The step described in subsection V-B is analyzed in more detail in the following subsection VI-B. After it completed and stored the results in a pre-splitted table, calculating the information gain of each feature, described in subsection V-C, took 69 to 97 seconds on both clusters. The final step, the export of the list of information gain of all features, took 46 to 70 seconds. All of the MapReduce tasks had an overhead of up to 60 seconds for compilation of the Pig script, generating JAR files, distributing them on the cluster and negotiating resources.

When preparing the 80 times replicated dataset we stored it in a slightly different format so we can later process the data and the labels at the same time. Namely, each line of the enlarged file contains pairs of the column indexes and values of all non-zero features. This representation takes 3 GB, whereas if we stored it in pure EAV format we would need about 5.5 GB (80 × 72) owing to the redundancy of line numbers. This does not have effect of any of the other steps except of how is it stored in HDFS. This file when copied on HDFS was automatically fragmented on 24 nodes (not counting the nodes for replication). On Fig. 3 is shown the data load time depending on the cluster configuration. It should be noted that even though there are 54 active nodes in the cluster in some cases we have intentionally created tables with more table regions (108, 162 and 216) aiming to leverage the multiple cores on each node.

Important to realize is that the 24 HDFS nodes on which the file is dispersed is an upper bond to the maximum number of map tasks when processing the file from HDFS and storing it in HBase tables. As a result, even though some tables have more than 24 regions during this phase it does not have an effect of the parallelism. Nevertheless, in the next steps when the data source is an HBase table, its number of regions dictates the number of generated map tasks. Another important thing to notice is that when using less nodes than 24 for the HBase tables, the number of map tasks is still 24 because this is dictated by the data source (HDFS file) and not by the destination (HBase table). From Fig. 3 it is evident that the load time is not reduced when more than 24 nodes HBase table regions are used. Also we see that when using less then 24 table regions the bottle neck is during the writes to the HBase tables. Finally, we want to emphasize the HBase table with only one region (the right-most case on Fig. 3). Even tough it was configured to have only one region by not specifying any split points for it at creation time, during the load it got larger than some configurable threshold, so HBase automatically splitted in two regions. Nevertheless, those two regions are on the same node.

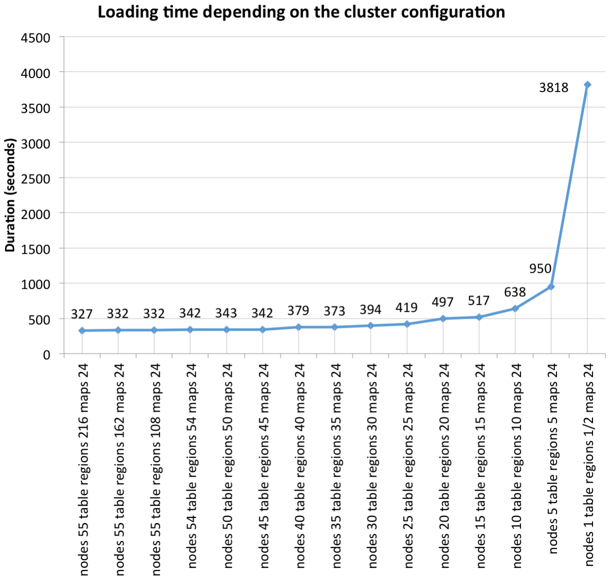


Fig. 3. Data load time for the 80 times replicated AAIA'14 dataset depending on cluster configuration

### B. Computationally expensive step - Calculating counts

The step described in subsection V-B was the most complicated and the speedup for it varied significantly depending on the cluster size and configuration. The remaining of this subsection describes details of the impact of the parallelization of this step and all listed speedups and durations are only for it.

First, we conducted experiments using the original AAIA'14 dataset on the FSCSE65 cluster. These results were published in [32], so here we are only reviewing them. These experiments were using only one reduce task, the default in Pig Latin. Also here we used more map tasks than actual nodes because each node is a multi-core machine. The results confirmed that indeed using more map tasks is beneficial, which is intuitively logical. Nevertheless, when we further increase the number of map tasks, the performance gradually degrades. The explanation for this is that as the number of map tasks gets larger, the operating system on the nodes needs to spend more time on task switching, swapping, while

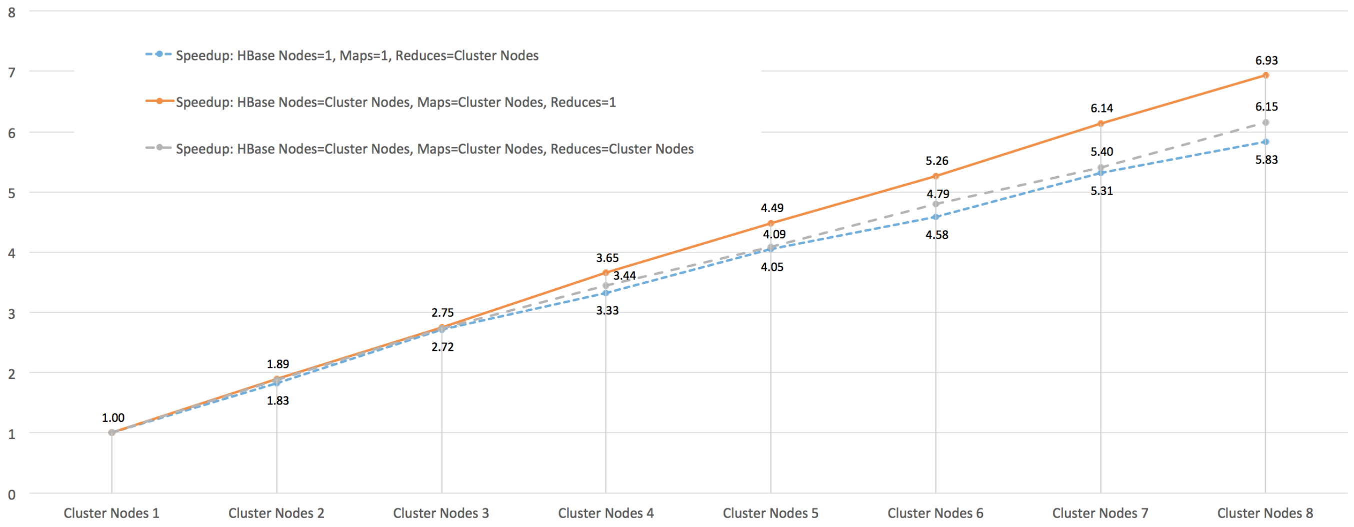


Fig. 4. Speedup depending on the number of active nodes, map and reduce tasks on the Amazon32 cluster

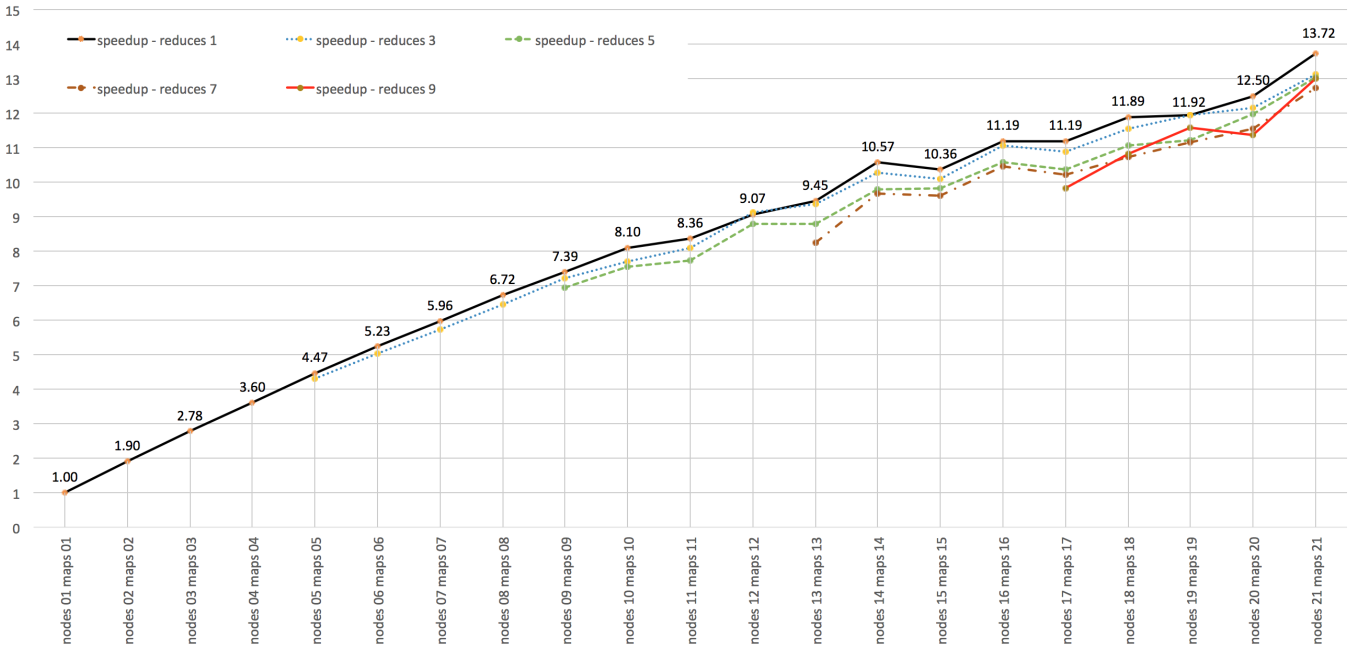


Fig. 5. Speedup depending on the number of active nodes, map and reduce tasks on the FCSE24 cluster

also needing to run many Hadoop and other services in the background. The total duration of this step on the one-node cluster was 3656 seconds, while the quickest solution obtained when using 59 nodes and 177 map tasks took 129 seconds on this cluster and the corresponding speedup was 28.34.

Then we continued our experiments on the Amazon32 cluster, trying to determine the impact of the number of nodes, maps and reduces. We have tried three options when trying to utilize the nodes of the cluster: use as much as possible nodes to run map tasks and have only one reduce task; use as much as possible nodes to run both map and reduce tasks; and use only one node for one map task and use all available nodes for reduce tasks. The speedup compared to the one-node cluster depending on the available nodes using these three options are shown on Fig. 4. It indicates that for this dataset it is best to have only one reduce phase, but use as many nodes as possible for the map tasks. This, in fact, makes sense because the work is performed during the map phase and during the reduce phase these results are only grouped together. Having more than the default of one reduce task actually increases the duration because the partial results in each reduce task need to be merged together. The total duration of this step on the one-node cluster was 4732 seconds, while the quickest solution with speedup of 6.83 took 693 seconds.

Aiming to confirm these findings we continued testing on the FCSE24 cluster, using the same approach. Additionally

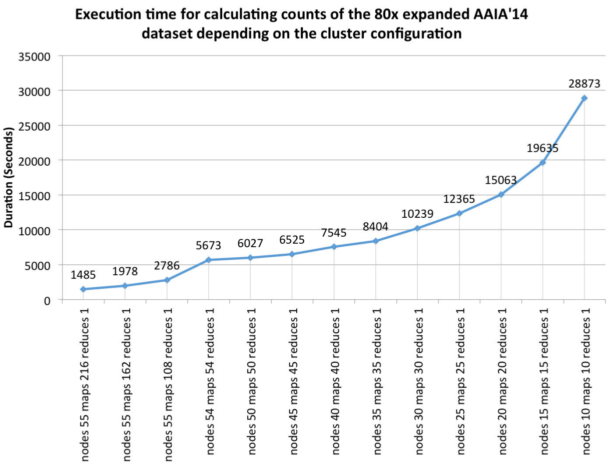


Fig. 6. Execution time for calculating counts of the 80x replicated AAIA'14 dataset depending on the cluster configuration

we have tried using 1,3,5,7 or 9 reduce tasks, depending on the number of available nodes. Our intent was to confirm that using only one reduce task (the default value in Pig Latin) will be more appropriate for a dataset of this size. The charts shown on Fig. 5 indeed confirm this assumption. The greatest speedup was always achieved when using only one reduce task, regardless of the number of available nodes. The total duration of this step on the one-node cluster was 3637 seconds, while the quickest solution with speedup of 13.72 took 265 seconds.

Finally, we have analyzed the execution time on the FCSE65 cluster using the 80 times replicated dataset. We have started experimenting using 54 nodes and gradually reducing the number of nodes by 5. When using 54 nodes we have also tried used 2, 3 and 4 times more table regions than actual nodes. Fig. 6 shows the execution times depending on the various configurations. In all cases the number of reduces was 1. Owing to the fact that this dataset is quite large, executing this step on smaller clusters took a significant amount of time. Additionally because HBase splitted the table on the one node cluster to two regions, using that execution time for calculating speedup would have been inconsistent with the previous setups. Therefore, for this experiment on Fig. 6 we are reporting the execution time and not the speedup. By performing this experiment we have confirmed that the proposed parallel implementation is scalable to large datasets for which the processing with a sequential implementation would be quite difficult if not impossible.

## VII. Conclusion and future work

In this paper we have reviewed the applications of the metric information gain for ranking individual features, discretization of continuous valued features, improving decision tree performance, localization, rough sets, etc. In a Big Data setting those tasks become a significant challenge, and therefore the need for its parallelization. In this paper we have proposed a parallel implementation of it. In oder to facilitate this, we have proposed a generic framework for data parallelization and then all steps from the algorithm for computation of information gain were parallelized using it. The benefits from using the scripting language Pig Latin were evident by the code listings which allowed fast development of MapReduce jobs. We have also demonstrated how can we manually set the degree of parallelism by pre-splitting the HBase tables so they have optimal number of regions and even data distribution across regions. The experiments confirmed that for this type of algorithm it is best to use only one reduce task. We have also validated that the multi-core nodes are providing increased performance when they execute more map tasks simultaneously. By deploying the implementation on Amazon AWS and on-premises clusters we have demonstrated the portability of the approach. The correctness of the implementation was verified by comparing the ranked features with the results we obtained from WEKA. Not to neglect were also the findings related to the scalability of the approach to an even larger dataset with millions of instances and dozens of thousands of features.

In our future work we plan to utilize the proposed implementation for other task In that manner, we also need to propose valid data transformation and normalization techniques, so we can generalize the approach and make it available for datasets that contain non-discretized continuous or nominal features. Additionally, we aim to apply the current parallelization for building decision trees. Finally, we plan to parallelize other more advanced feature selection algorithms using a similar framework.

## Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss.Cyril and Methodius University, Skopje, Macedonia.

## References

[1] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, ser. OSDI'04. Berkeley, CA, USA: USENIX Association, 2004, pp. 10–10. [Online]. Available: http://dl.acm.org/citation.cfm?id=1251254.1251264

[2] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," in *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation - Volume 7*, ser. OSDI '06. Berkeley, CA, USA: USENIX Association, 2006, pp. 15–15. [Online]. Available: http://dl.acm.org/citation.cfm?id=1267308.1267323

[3] "Hadoop wiki: List of institutions that are using hadoop for educational or production uses, howpublished = https://wiki.apache.org/hadoop/poweredby, note = Accessed: 2015-01-29."

[4] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0

[5] T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill Science/Engineering/Math, 3 1997. ISBN 9780070428072. [Online]. Available: http://amazon.com/o/ASIN/0070428077/

[6] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in

*Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. ISBN 1-55860-612-2 pp. 258–267. [Online]. Available: http://dl.acm.org/citation.cfm?id=645528.657649

[7] R. O. Duda, *Pattern classification*, 2nd ed. New York: Wiley, 2001. ISBN 0471056693

[8] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, ser. AAAI'91. AAAI Press, 1991. ISBN 0-262-51059-6 pp. 547–552. [Online]. Available: http://dl.acm.org/citation.cfm?id=1865756.1865761

[9] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 245 – 271, 1997. doi: http://dx.doi.org/10.1016/S0004-3702(97)00063-5 Relevance. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370297000635

[10] P. Langley, *Elements of machine learning*. San Francisco, Calif: Morgan Kaufmann, 1996. ISBN 1558603018

[11] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine Learning: Proceedings of the Eleventh International Conference*. Morgan Kaufmann, 1994, pp. 121–129.

[12] B. Raman and T. R. Ioerger, "Instance based filter for feature selection," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 1–23, 2002.

[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968

[14] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets / Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Standford University*, 2nd ed. Cambridge: Cambridge University Press, 2014. ISBN 9781107077232 1107077230 1316147312 9781316147313

[15] C. Dobre and F. Xhafa, "Parallel programming paradigms and frameworks in big data era," *International Journal of Parallel Programming*, vol. 42, no. 5, pp. 710–738, 2014. doi: 10.1007/s10766-013-0272-7. [Online]. Available: http://dx.doi.org/10.1007/s10766-013-0272-7

[16] S. Singh, J. Kubica, S. Larsen, and D. Sorokina, "Parallel large scale feature selection for logistic regression." in *SDM*. SIAM, 2009, pp. 1172–1183.

[17] Z. Sun and Z. Li, "Data intensive parallel feature selection method study," in *Neural Networks (IJCNN), 2014 International Joint Conference on*, July 2014. doi: 10.1109/IJCNN.2014.6889409 pp. 2256–2262.

[18] L. Zhou, H. Wang, and W. Wang, "Parallel implementation of classification algorithms based on cloud computing environment," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 10, no. 5, pp. 1087–1092, 2012.

[19] G. Caruana, M. Li, and M. Qi, "A mapreduce based parallel svm for large scale spam filtering," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, vol. 4, July 2011. doi: 10.1109/FSKD.2011.6020074 pp. 2659–2662.

[20] I. Triguero, D. Peralta, J. Bacardit, S. García, and F. Herrera, "Mrpr: A mapreduce solution for prototype reduction in big data classification," *Neurocomputing*, vol. 150, pp. 331–345, 2015.

[21] A. K. Farahat, A. Elgohary, A. Ghodsi, and M. S. Kamel, "Distributed column subset selection on mapreduce," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 171–180.

[22] A. Guillén, A. Sorjamaa, Y. Miche, A. Lendasse, and I. Rojas, "Efficient parallel feature selection for steganography problems," in *Bio-Inspired Systems: Computational and Ambient Intelligence*, ser. Lecture Notes in Computer Science, J. Cabestany, F. Sandoval, A. Prieto, and J. Corchado, Eds. Springer Berlin Heidelberg, 2009, vol. 5517, pp. 1224–1231. ISBN 978-3-642-02477-1. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-02478-8_153

[23] S. Owen, R. Anil, T. Dunning, and E. Friedman, *Mahout in Action*. Greenwich, CT, USA: Manning Publications Co., 2011. ISBN 1935182684, 9781935182689

[24] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. Hoboken, NJ: Wiley, 2006. ISBN 9780471241959 0471241954 9780471241959

[25] C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, "Feature selection via maximizing global information gain for text classification," *Knowledge-Based Systems*, vol. 54, no. 0, pp. 298 – 309, 2013. doi: http://dx.doi.org/10.1016/j.knosys.2013.09.019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705113003067

[26] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 155–165, Jan. 2006. doi: 10.1016/j.ipm.2004.08.006. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2004.08.006

[27] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry, France, August 28 - September 3, 1993*, 1993, pp. 1022–1029.

[28] J. Dougherty, R. Kohavi, M. Sahami *et al.*, "Supervised and unsupervised discretization of continuous features," in *Machine learning: proceedings of the twelfth international conference*, vol. 12, 1995, pp. 194–202.

[29] U. M. Fayyad and K. B. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine Learning*, vol. 8, pp. 87–102, 1992. doi: 10.1007/BF00994007. [Online]. Available: http://dx.doi.org/10.1007/BF00994007

[30] C. Stachniss, G. Grisetti, and W. Burgard, "Information gain-based exploration using rao-blackwellized particle filters," in *Robotics: Science and Systems*, vol. 2, 2005, pp. 65–72.

[31] D. Ślezak, "Approximate entropy reducts," *Fundam. Inf.*, vol. 53, no. 3-4, pp. 365–390, Aug. 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=2371245.2371255

[32] E. Zdravevski, P. Lameski, A. Kulakov, B. Jakimovski, S. Filiposka, and D. Trajanov, "Feature ranking based on information gain for large classification problems with mapreduce," in *Proceedings of the 9th IEEE International Conference on Big Data Science and Engineering*. IEEE Computer Society Conference Publishing, August 2015,

in print August 2015.

[33] D. Miner, *MapReduce design patterns.* Sebastopol, CA: O'Reilly, 2013. ISBN 9781449327170

[34] A. Holmes, *Hadoop in practice.* Shelter Island, NY: Manning, 2012. ISBN 9781617290237 1617290238

[35] T. White, *Hadoop: the definitive guide,* 3rd ed. Beijing: O'Reilly, 2012. ISBN 9781449311520

[36] "Apache hadoop nextgen mapreduce (yarn)," http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html, accessed: 2015-01-29.

[37] "Hdfs architecture guide," http://hadoop.apache.org/docs/r1.2.1/hdfs design.html, accessed: 2015-01-29.

[38] L. George, *HBase the definitive guide.* Sebastopol, CA: O'Reilly, 2011. ISBN 97814493157711449315771. [Online]. Available: http://public.eblib.com/choice/publicfullrecord.aspx?p=769368

[39] Y. Jiang, *HBase administration cookbook master HBase configuration and administration for optimum database performance.* Birmingham: Packt Publishing, 2012. ISBN 9781849517157 18495171501849 517142 9781849517140. [Online]. Available: http://site.ebrary. com/id/10598980

[40] N. Dimiduk and A. Khurana, *HBase in action.* Shelter Island, NY: Manning, 2013. ISBN 16172905219781617290527

[41] E. A. Brewer, "Towards robust distributed systems (abstract)," in *Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing,* ser. PODC '00. New York, NY, USA: ACM, 2000. doi: 10.1145/343477.343502. ISBN 1-58113-183-6 pp. 7–. [Online]. Available: http://doi.acm.org/10.1145/343477.343502

[42] A. F. Gates, O. Natkovich, S. Chopra, P. Kamath, S. M. Narayanamurthy, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava, "Building a high-level dataflow system on top of map-reduce: The pig experience," *Proc. VLDB Endow.,* vol. 2, no. 2, pp. 1414–1425, Aug. 2009. doi: 10.14778/1687553.1687568. [Online]. Available: http://dx.doi.org/10.14778/1687553.1687568

[43] A. Gates, *Programming Pig.* Sebastopol: O'Reilly Media, 2011. ISBN 9781449317690 1449317693 9781449317683 1449317685. [Online]. Available: http://public.eblib.com/choice/publicfullrecord .aspx?p=801461

[44] A. Janusz, A. Krasuski, S. Stawicki, M. Rosiak, D. Slezak, and H. S. Nguyen, "Key risk factors for polish state fire service: A data mining competition at knowledge pit," in *Computer Science and Information Systems (FedCSIS),* 2014 Federated Conference on, Sept 2014. doi: 10.15439/2014F507 pp. 345–354.

[45] E. Zdravevski, P. Lameski, A. Kulakov, and D. Gjorgjevikj, "Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets," in *Computer Science and Information Systems (FedCSIS),* 2014 Federated Conference on, Sept 2014. doi: 10.15439/2014F500 pp. 387–394.

[46] A. H. Team, "Apache HBase reference guide," http://hbase.apache .org/book.html, accessed: 2015-03-29.

DOI: 10.15439/2015F386
ACSIS, Vol. 5

# Comparison Of Language Models Trained On Written Texts And Speech Transcripts In The Context Of Automatic Speech Recognition

Sebastian Dziadzio [1], Aleksandra Nabożny[1], Aleksander Smywiński-Pohl[1,2,3], Bartosz Ziółko[1,2]
[1] AGH University of Science and Technology,
Faculty of Computer Science, Electronics and Telecommunications, Krakow, Poland
[2] Techmo, Krakow, Poland, techmo.pl,
[3] Jagiellonian University, Department of Computational Linguistics, Krakow, Poland
dziadzio@student.agh.edu.pl, aleksander.pohl@uj.edu.pl, bziolko@agh.edu.pl

*Abstract*—**We investigate whether language models used in automatic speech recognition (ASR) should be trained on speech transcripts rather than on written texts. By calculating log-likelihood statistic for part-of-speech (POS) n-grams, we show that there are significant differences between written texts and speech transcripts. We also test the performance of language models trained on speech transcripts and written texts in ASR and show that using the former results in greater word error reduction rates (WERR), even if the model is trained on much smaller corpora. For our experiments we used the manually labeled one million subcorpus of the National Corpus of Polish and an HTK acoustic model.**

*Index Terms*—**automatic speech recognition, morphosyntactic language model, written and spoken language comparison**

## I. Introduction

STATISTICAL language models (LM) are employed in various natural language processing applications, such as machine translation, information retrieval, ASR [21], or part-of-speech tagging [20]. Generally, they describe relations between words (or other tokens), thus enabling to choose most probable sequences. This proves to be especially useful in speech recognition, where acoustical models usually produce a number of hypotheses, and re-ranking them according to a language model can substantially improve recognition rates [20],[4],[6].

Despite extensive research into alternative techniques, n-gram models remain a technology of choice for most modern ASR systems. They are based on Markov assumption, which states that probability of a certain word is dependent only on its n-1 predecessors. It should be noted that efficiency of n-gram models is heavily language dependent. They correspond well to grammatical structure of positional languages (such as English), but in case of Polish and other highly inflected languages, words order is not a key indicator of relations between them [8]. The main difficulty in language modelling and learning problems in general is the curse of dimensionality. Higher-order models are usually more accurate, but with more dimensions the volume of space increases so fast that available data quickly become insufficient [2]. This problem is amplified in case of Polish due to complex inflectional rules resulting in a variety of word-forms.

Several techniques were proposed to account for long-span word dependencies and address the data sparsity problem. One of them are part-of-speech (POS) n-grams, which cluster words into categories based on grammatical classes [12], [14]. Such models are easy to build and allow the use of higher order n-grams, since there are far fewer grammatical categories than words. Furthermore, they can be trained on much smaller corpora, which is especially important for under-resourced languages. Written texts are usually easier to obtain than speech transcripts and consequently language models are commonly trained on the former [5] [18].

## II. Motivation

There has been a lot of studies in the humanities and social sciences dealing with the comparison of speech and text. It is known that there are fundamental dissimilarities between oral and written language in terms of grammatical structures, sentence lengths, choice of words etc. [3]. Whether those differences can be captured by means of statistical analysis, remains an open question.

The main motivation behind our study was to investigate whether LM based on written texts are an appropriate source of information about spoken language for automatic speech recognition. We conducted a comparative analysis of two corpora. One of them consisted of speech transcripts, while the other contained only written texts. We were looking for general features allowing to distinguish between two channels of communication (speech vs. text) rather than stylistic differences resulting from distinct language domains. That is why traditional methods of corpus comparison based on word frequencies were not applicable [15]. We therefore decided to compare POS n-grams in order to find grammatical patterns typical of either spoken or written language. Our initial hypothesis holds that there are statistically significant differences between those two n-gram sets. If this assumption is correct, it would imply that training LM solely on speech transcripts could lead to greater WERR in ASR systems.

978-83-60810-66-8/$25.00©2015, IEEE

## III. Related Work

The idea of comparing speech and text corpora in terms of POS tags was motivated by previous research concerning the use of morphosyntactic n-grams in speech recognition of Polish. Until recently, there was little interest in using POS tags in ASR. In [22] a POS tagger was tested as a possible improvement in speech recognition of Polish. The results were negative, because the tagger frequently produced ambiguous output. This issue was later addressed in [11] by reducing model specificity (only grammatical classes were taken into account). It was concluded that simplified POS tags can be very useful for building statistical models of Polish.

In [12] an optimal set of grammatical categories was experimentally derived. Thirteen trigram language models were built, each employing both grammatical classes and one selected grammatical category. Then they were compared to a model based only on grammatical classes (hereinafter called POS-only model) in terms of WERR. Only three categories (gender, number, and case) offered significant improvements over the POS-only model. Surprisingly, combining those categories resulted in a model performing insignificantly better than the POS-only model. For this reason, our research is mostly based on the POS-only model, although we also take into account three aforementioned categories.

## IV. Data Preparation

The National Corpus of Polish (NKJP) is divided into two parts: manually annotated 1-million corpus (1MC) and automatically annotated 1-billion corpus (1BC). Texts are labeled on several lavels: word and sentence boundaries, morphosyntactic tags, named entities, and syntactic groups. Annotation in 1MC is conducted very strictly, as each element was labeled by two independent researchers and then corrected by a super-annotator in case of a tie. The corpus includes diverse materials: classic literature, daily newspapers, scientific journals, and a variety of short-lived and Internet texts. Most importantly, it also includes speech transcripts from parliament proceedings, real-life conversations, radio, and television [13]. The proportion of speech transcripts to text data in 1MC is 109 919 (speech) vs. 1 091 981 (text) tokens.

Each segment in NKJP belongs to one of 35 grammatical classes. They are far more detailed than traditional parts of speech (for example there are 14 distinctive verb classes and 4 adjective classes). Obtaining information about grammatical classes was straightforward and required parsing XML label files. Unfortunately each paragraph is described by several label files stored in a separate directory, so they had to be processed individually. Although rather inconvenient, this design prompted us to take advantage of parallel processing, which will later be useful in case of 1-billion corpus.

Extracting grammatical categories was a more demanding task, mainly because category tags take a form of a single, colon-delimited string. For example, the word objęcia has a following tagging: ger:sg:gen:n:perf:aff. The first element is the grammatical class (POS) tag, followed by a set of grammatical category tags. This notation is further complicated by the fact that each grammatical class has a different set of categories. For example, adjectives have gender, number, case, and degree, while verbs are described by their number, person, and aspect. As it has already been said, only gender, number, and case were taken into account, as they play primary role in agreement relation.

It should be noted that we ignored all non-lexical backchannels and other noise in the transcripts. We also discarded all utterances containing incomprehensible words, as we wanted to focus on grammatical properties of the spoken language.

## V. Statistical Comparison

Selecting appropriate statistical tools was yet another challenge. We considered three methods: the Spearman's coefficient, χ2-test and log-likelihood statistic. We concluded that the first method is not applicable to POS n-grams because of its tendency to overestimate differences for rare units. We also rejected the χ2-test because its null hypothesis is that compared corpora comprise words drawn randomly from a larger population. Since words in texts are obviously not random, the null hypothesis is defeated for almost all common words [9]. It is especially problematic for POS n-grams, where there are typically several very common units (which can be expected to give high χ2 values) and a lot of rare units (for which the χ2 test is not applicable). We decided to use the third method, as it is applicable to corpora of different sizes and has been reported to work well with POS n-grams [15]. Given the frequency lists, we build a contingency table for each POS n-gram:

Table I.
Example contingency table.

| | **Corpus A** | **Corpus B** |
|---|---|---|
| Count of unit: | $n_A$ | $n_B$ |
| Count of other units: | $N_A-n_A$ | $N_B-n_B$ |
| Total: | $N_A$ | $N_B$ |

Values $n_A$ and $n_B$ are called observed values (O). We then calculate expected values (E) according to the formula:

$$E_i=\frac{N_i\sum_i O_i}{\sum_i N_i} \qquad (10$$

Using the data from Table 1, we obtain $E_A=\frac{N_A(n_A+n_B)}{N_A+N_B}$ and $E_B=\frac{N_B(n_A+n_B)}{N_A+N_B}$. The log-likelihood value is then calculated according to the following formula:

$$2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right) \qquad (2)$$

In our case this equals:

$$2n_A \ln\left(\frac{n_A}{E_A}\right) + 2n_B \ln\left(\frac{n_B}{E_B}\right) \tag{3}$$

The higher this value, the more significant is the difference between two frequency scores. LL of 3.8 or higher is significant at the 5% level. For the purpose of comparison, we used five corpora of written texts and five corpora of speech transcripts (full corpus, two half-corpora and two smaller samples). We then performed a round robin comparison: for each pair of corpora we calculated the number of units for which the LL value was greater than 3.8. Averaged results are presented below. S-S and T-T denote intra-corpus comparisons (speech and text, respectively). S-T denotes a comparison between speech and text corpora.

TABLE II.
AVERAGE NUMBER OF N-GRAMS WITH DIFFERENCES IN FREQUENCY SIGNIFICANT AT 5% LEVEL. VALUES IN BRACKETS ARE STANDARD DEVIATIONS.

| n | S-T | S-S | T-T |
|---|---|---|---|
| 1 | 30.3 (2.0) | 14.1 (5.1) | 17.2 (3.6) |
| 2 | 418.8 (42.6) | 127.2 (28.2) | 182.5 (49.0) |
| 3 | 2281.4 (482.7) | 1205.1 (215.8) | 1628.4 (274.3) |

The log-likelihood analysis reveals large differences in frequencies of POS-tags. The LL scores were significant at 5% level for more than 30 unigrams (out of 35). This number is much lower in case of intra-corpus comparisons. The same holds true for higher-order n-grams (bigrams and trigrams). As stated before, we used five corpora for speech and text (resulting in 10 intra-corpus comparisons and 25 inter-corpus comparisons), so observed differences are not an effect of differing corpus sizes. Qualitative analysis of POS tags with highest LL score could reveal usage patterns characteristic for written and spoken language.

Another test involved calculating the percentage of common n-grams in the set of k most popular units:

$$\frac{|K_1 \cap K_2|}{|K_1 \cup K_2|} \cdot 100 \tag{4}$$

In the above formula, K1 and K2 denote sets of k most popular n-grams in compared corpora. We considered unigrams, bigrams, and trigrams. We decided to set k in relation to the total number of units (5%, 10%, and 20% of all units). Table 3 presents calculated values. “S-T” denotes a comparison of full speech corpus vs. full text corpus. “S-S” and “T-T” denote a comparison between two halves of the same corpora (the split was made by randomly assigning each paragraph into one of two subcorpora).

The test reveals significant differences in POS n-gram distributions. The values in the first column (speech vs. text) are not only lower, but also decreasing with the model complexity. The values in the second and third column (speech vs. speech and text vs. text) are much higher and stay the same as the order of n-grams increases. This shows that grammatical patterns typical for spoken or written language can be captured with morphosyntactic n-gram models.

TABLE III.
PERCENTAGES OF COMMON UNITS AMONG K MOST POPULAR N-GRAMS.

Unigrams

| k | S-T | S-S | T-T |
|---|---|---|---|
| 2 | 100.0 | 100.0 | 100.0 |
| 5 | 100.0 | 100.0 | 100.0 |
| 10 | 90.0 | 100.0 | 100.0 |

Unigrams with categories

| k | S-T | S-S | T-T |
|---|---|---|---|
| 20 | 85.0 | 100.0 | 100.0 |
| 40 | 87.5 | 95.0 | 98.0 |
| 80 | 85.0 | 97.5 | 98.8 |

Bigrams

| k | S-T | S-S | T-T |
|---|---|---|---|
| 35 | 78.6 | 94.3 | 100.0 |
| 70 | 77.1 | 95.7 | 100.0 |
| 140 | 74.3 | 93.6 | 98.6 |

Bigrams with categories

| k | S-T | S-S | T-T |
|---|---|---|---|
| 400 | 70.6 | 88.8 | 97.8 |
| 800 | 72.8 | 87.4 | 95.8 |
| 1600 | 70.5 | 85.9 | 94.6 |

Trigrams

| k | S-T | S-S | T-T |
|---|---|---|---|
| 250 | 64.6 | 89.2 | 97.2 |
| 500 | 63.9 | 90.2 | 96.4 |
| 1000 | 64.6 | 89.2 | 95.8 |

## VI. PERFORMANCE IN ASR

The results of statistical analysis indicated that language models trained on speech transcripts or written texts would have different properties and therefore give different results when applied to ASR. In order to test this hypothesis, we have built several language models and employed them in rescoring of the hypotheses produced by HTK (without any LM or grammar) for several hundred Polish sentences. For tagging we used Concraft-pl, a conditional random field tagger for Polish which had proved to be particularly effective in ASR applications [17],[12]. The rescoring was done by combining the probabilities of the acoustic and morphosyntactic model

$$P(h_i)=P(h_i)_{LM}^{\alpha}\cdot P(h_i)_{AM}^{1-\alpha} \qquad (5)$$

where

$P(h_i)$ – the probability of the i-th hypothesis,

$P(h_i)_{LM}$ – the probability of the i-th hypothesis according to the language model,

$P(h_i)_{AM}$ – the probability of the i-th hypothesis according to the acoustic model,

$\alpha$ – the weight of the LM component.

The models were tested on several audio corpora. The first one (K1) includes 107 sentences spoken by one male voice, without any added noise, but recorded in an office with working computers. It consists of political speeches and spoken fragments of political song lyrics. The second corpus (K2) includes 23 samples spoken by a young female professional speaker. The third corpus (K3) consists of 221 short utterances recorded during various tests of speech/speaker recognition systems at AGH University of Science and Technology with addition of recordings from meetings of the Department Council. This corpus includes many various voices (one speaker says no more than six sentences) and recording devices, often with a natural random noise due to bad acoustic conditions (reverberation, voices in the background, traffic from outside etc.) We also used some recordings from LUNA, a corpus of telephone conversations from a call center of Warsaw public transport [10]. 192 samples of various female voices (K4) and 226 of male voices (K5) were used. These are informal utterances with many questions. The corpus is full of grammar mistakes, very common in natural conversations. The last test corpus (K6) consists of 86 recordings randomly chosen from Polish Global Phone corpus [16]. It is a corpus of speech dictated from an everyday journal.

The union of the corpora was divided into two subsets: a tuning set containing 15% randomly chosen sentences, used to estimate the alpha parameter, and a testing set, containing the remaining sentences. The text and the speech corpora were used to build two language models (LMs): one containing only POS tags (POS-only) and the other containing POS tags together with gender, number and case tags (POS-gnc). In each case a trigram model was built, smoothed using Witten-Bell method [19], due to their small size.

The comparison of speech and text based LMs was conducted by measuring the Word Error Rate Reduction (WERR) obtained with a given model. The results of the test are given in Table 4. LMs with Speech prefix are based on the Speech sub-corpus of 1MC, with Text prefix – on the Text sub-corpus, and with Text-sample, on a text sub-corpus of the same size as the Speech sub-corpus. The best result is obtained for the LM based on the speech corpus using POS, gender, number and case tags. The difference between the best result and the second result (Text-sample-POS-gnc) is statistically significant (paired Student's t-test, n=724, P < 0.028). Interestingly, although the Speech-POS-only LM performs better than the Text-POS-only LM, the difference is not statistically significant.

Table 4.
Performance of different LMs in ASR.

| LM | WERR [percentage points] |
|---|---|
| Speech-POS-gnc | 29.5 |
| Text-sample-POS-gnc | 28.0 |
| Text-POS-gnc | 27.8 |
| Speech-POS-only | 27.1 |
| Text-POS-only | 26.5 |
| Text-sample-POS-only | 25.9 |

## VII. Conclusions

Building language models based on POS n-grams is a promising technique in ASR of highly inflected languages. Benefits include simple structure, substantial dimensionality reductions, and noticeable improvements in performance of ASR systems [12]. Our analysis shows that it is possible to discriminate between speech and text data using only POS n-grams. It implies that morphosyntactic models trained on written texts do not accurately reflect the grammatical structure of spoken language. This hypothesis was confirmed by the ASR experiments. The Speech-POS-gnc model outperformed all text-based models, even those trained on ten times more data. The experiment also show that grammatical categories (gender, number, and case) carry important information about the structure of inflectional languages. Including them improved recognition rates in all cases.

## VIII. Acknowledgements

This work was supported by LIDER/37/69/L-3/11/NCBR/2012 and DOB-BIO6/22/133/2014 grants from the Polish National Center for Research and Development.

Narodowe Centrum
Badań i Rozwoju

## References

[1] Bardoel, T. “Comparing n–gram frequency distributions”. Tilburg University School of Humanities. Tilburg center for Cognition and Communication. 2012.

[2] Bengio, Yoshua, Ducharme, Réjean, Vincent, Pascal, Jauvin, Christian. “A neural probabilistic language model”. Journal of Machine Learning Research. vol. 3. pp. 1137-1155. 2003.

[3] Biber, Douglas. “Variation across speech and writing”. Cambridge University Press. 1991.

[4] Chelba Ciprian, Bikel Dan, Shugrina Maria, Nguyen Patrick, Kumar Shankar. “Large scale language modelling in automatic speech recognition.”. Google Research. 2012.

[5] Hirsimaki, T., Pylkkonen, J., Kurimo, M., “Importance of high-order n-gram models in morph-based speech recognition”. IEEE Trans.

Speech and Language Processing. 17(4):724-32. 2009. http://dx.doi.org/10.1109/TASL.2008.2012323

[6] Janicki, A., Wawer, D., “Automatic Speech Recognition of Polish in a Computer Game Interface”, Proceedings of the Federated Conference on Computer Science and Information System 2011, pp. 711–716. 2011.

[7] Jurafsky, D., Martin, J. H. “Speech and language processing. 2nd edition”. Prentice-Hall. Inc. New Jersey. 2008.

[8] Karpov, A., Ronzhin, A., Markov, K., Kipyatkova, I., Vazhenina, D. “Large vocabulary Russian speech recognition using syntactico-statistical language modelling”. Speech Communication 56 (2014) 213-228. 2014. http://dx.doi.org/ 10.1016/j.specom.2013.07.004

[9] Kilgarriff, Adam. “Comparing Corpora”. International Journal of Corpus Linguistics. 6:1. 97-133. 2001.

[10] Marciniak, M. “Anotowany korpus dialogów telefonicznych.”. Akademicka Oficyna wydawnicza EXIT. 2011.

[11] Pohl, A., Ziółko, B. “Using part of speech n-grams for improving automatic speech recognition of Polish”. 9th International Conference on Machine Learning and Data Mining MLDM. 2013. http://dx.doi.org/10.1007/978-3-642-39712-7_38

DOI: 10.15439/2015F182
ACSIS, Vol. 5

# Towards the automatic motion recovery using single-view image sequences acquired from bike

Jakub Kolecki
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: kolecki@agh.edu.pl

***Abstract*—This paper describes the design, implementation and results of the image-based ego-motion estimation algorithm. As a source data the images captured from the bike platform are used. The device is supposed to be a part of a mobile mapping system prototype. Firstly the feature detection and matching is carried out providing the set of characteristic points in all images in the sequence. The 5-point solution based on the Gröbner basis is used to solve for essential matrices and to reject outliers. Least-square relative pose model fitting is accomplished using quaternion-based bundle adjustment. In the next step the modified Horn formula is used to recover bike trajectory up to the absolute orientation. Within this step the scene structure recovery is provided in the form of a point cloud. Finally ground control information is used to obtain data geo-referencing and the accuracy analysis. Obtained results provide satisfying robustness and accuracy. However some improvements and development scenarios are suggested.**

## I. Introduction

CURRENTLY imaging sensors are extensively used as components of mobile mapping systems (MMSs), mobile robots and unmanned aerial vehicles. Each camera is a source of usually large number of images, captured with the specified frequency. Acquired image sequences may be processed to provide automatically extracted mapping information using algorithms refered as a dense point cloud generation or structure from motion (SFM). Additionally images may be utilized to estimate motion trajectory of the vehicles. Such application is often called the visual odometry. The real-time trajectory estimation is applied in the navigation. The visual navigation can take place autonomously or together with the inertial/GNSS sensors, completing a multi-sensor navigation system. Parallel navigation and mapping are sometimes combined together in the SLAM procedure.

Basically mapping applications do not require real time computation of a trajectory. The accurate trajectory computation is conducted in the post processing and is a crucial step in the mobile mapping workflow as it greatly influences the accuracy of final products. Processing of the image sequences can be divided into two main steps:

- feature detection and matching
- ego-motion estimation, based on the detected features

In the machine vision feature detectors try to imitate humane vision to search for some characteristic points (keypoints), that are suitable to be traced in subsequent images. The feature correspondence is tested using descriptors and detectors that try to simulate the mental process [1]. Corners are a typical example of features suitable for tracing. The result of a feature detection is a list of keypoints' IDs and their coordinates provided in the image 2D coordinate frame. Evaluation of particular feature detectors are not within the scope of this work, but generally a set of automatically measured keypoints has a large number (sometimes over 50%) of outliers i.e. missmatched features. The first approach to deal with outliers is to prevent false matches using the external information about image geometry. This information can come from positioning sensors such as GPS/INS systems [2], [3]. If the orientation of two images acquired with calibrated camera is approximately known, location of corresponding features is held down to the neighborhood of epipolar lines associated with those points (Fig. A1). However more robust approach to prevent false matches is to use multi-view camera configuration providing multi-view image sequences [4], [5], [6]. Commonly two cameras are applied. As a result of a system calibration the accurate orientation of the second camera in the coordinate frame of the first camera is known. This enables the accurate epipolar line location for a certain keypoint (Appendix A).

In the case of single-view sequences, the keypoint matching cannot really benefit from the epipolar constraint (Appendix A). If no GPS or IMU are available, only the approximate motion characteristic is known, constraining corresponding keypoint searching to region of interests (ROI) rather then lines. As a result a significant number of outliers may occur. Approaches to the outlier rejection are based on the epipolar constriant imposed on the fundamental matrix ($F$) or the essential matrix ($E$) (Appendix A). Snavely, Seitz and Szeliski [7] propose the estimation of the $F$ matrix using the 8-point algorithm inside the RANSAC [8]. As a result a set of 8 image points that best fit the fundamental matrix model is found in each image pair. At the same time outliers can be detected and rejected. Bartelsen and Mayer [9] prefer to use the essential matrix instead of the fundamental matrix. In contrast to the fundamental matrix, the essential matrix estimation requires the knowledge about the camera calibration but remains robust to the critical configurations met in the

Researchers conducted within founds of Department of Mining Surveying and Environmental Engineering (AGH University of Science and Technology) no. 11.11.150.949

978-83-60810-66-8/$25.00©2015, IEEE

planar scenes. It also requires smaller number of coresponding points. The 5-point algorithm developed by Nister [10] and the locally optimized RANSAC [11] are proposed as a solution method. Finally the $E$ or $F$ model fitting can by carried out based on inliers only, using least-square approaches.

Applying the epipolar constraint cannot eliminate badly matched points that are located near the corresponding epipolar lines. A 3D information is necessary to detect the remaining, relatively small number of outliers. To solve the problem, the ray intersection (Appendix B) is carried out for each image model to calculate the spatial coordinates of the tie points (Fig. A.1). Sequential orientation of the subsequent images [2] using for example the DLT approach [12] inside the RANSAC procedure or formation of image triplets [9] allows to complete the rejection of outliers. Overlapping triplets can be linked to recover orientation of every image in the sequence. However the recovered orientation suffers from the drift effect. Besides it can be determined only up to the absolute scale, rotation and translation. Aforementioned 7 parameters, if needed, can be estimated based on ground control information such as the GPS coordinates of the image projection centers [9] or coordinates of the control points. Finally to estimate the orientation of the images more accurately the least square bundle adjustment can be applied. For the real time scenario the good solution to the drift reduction is the detection of loop closures [13].

A terrestrial mobile mapping can be carried out from almost every vehicle. Cars are used commonly in case of commercial systems. However data acquisition using cars is restricted to streets and their surroundings only. To overcome those limitations systems designed for smaller vehicles are developed, among which bikes seem to fill the gap between hand-held systems, mobile robots and cars. Bikes can access many more location than cars and move faster then pedestrians and mobile robots. Probably the most famous bike system for mobile data acquisition was developed by Google in the Street View project. Besides, students from Stuttgart University designed the prototype of the bike mobile mapping system with the laser scanner and two-antenna GNSS/INS unit [14].

Similarly to other mapping or visual odometry systems, bike systems can be a source of image sequences. Automatic processing of image data acquired from bike can be used to determine trajcetory and finally to reconstruct the geometry of objects. However it should be noticed that the bike movement is different to the car movement. When cycling it is more difficult to keep constant speed and direction than in case of driving. The turn rate of bike is small when compared to car. In case of single-view sequences with no information about approximate image geo-referencing, motion estimation from image data is supposed to be much more challenging task than in case of the multi-view systems and the smoother movement.

Works addressed in this paper aim to provide the solution to the problem of the automatic orientation of a single-view image sequences. It is beyond the scope of this study to examine the hardware potential as well as the achieved processing time. However the study offers some important insight into the analytical approaches and their practical aspects, providing at the same a kind of overview of the existing solutions. The following sections describe the consecutive steps of the algorithm, starting from keypoint detection and matching proceeding to the relative orientation and finally to the absolute orientation. The fifth section describes the experiments. Then the results and discussion are provided.

## II. Feature detection and matching

As this step of the algorithm is still under the development, only the outline of the matching strategy is provided. In the urban environment corner points are likely to occur in almost every image and can be detected using one of many available detectors. Proposed algorithm uses Kovesi's implementation [15] of Noble's version [16] of Harris feature detector [17]. After completing the detection, features are matched using the monogenic phase approach [18], [19] with the set of parameters proposed by Kovesi [15]. Assume a certain corner is detected and matched in the initial pair in the sequence. Matching algorithm searches for the corresponding point in the third and subsequently in the next images. At the same time new corners appear in consecutive images. As the approximate image-to-image distance and the scene depth are known, some simple geometric constrains like row and column limits can be imposed on a searching area greatly reducing the computation time.

Corner points are generally detected as the maxima in the image that is the result of applying Harris operator. The absolute maximum value (AMV) constraint can be set to limit the number of keypoints. However applying this simple constraint leads to nonuniform distribution of detected points. This is the result of variations in the scene content. Some parts of the scene like trees, windows, cars etc. are "rich" in keypoints, while others like flat walls contain no keypoints at all. To overcome this problem the image is divided into blocks of equal sizes. In the addressed case study the 24 blocks (4 × 6) are used . Besides absolute value of Harris maxima, two other parameters are set to provide more favorable distribution of detected features: maximum number of points in each block (PIB) and minimum distance between points (DBP). Decreasing the AMV and the PIB and at the same time increasing the DBP leads to the more uniform distribution of points. Exemplary values of feature detection parameters, applied in the refereed case study were as follows: AMV = 10 (Kovesi suggested 200 [15]), PIB = 60, DBP = 50.

As the result of matching procedure a list of points and their pixel coordinates is provided, allowing computation of the relative orientation of consecutive images.

## III. Relative orientation

### A. Essential matrix computation

The relative orientation of two images acquired with calibrated camera is encoded in the 3 × 3 essential matrix $E$. Derivation of the essential matrix from the relative orientation parameters - translation and rotation ($t$,$R$) and the camera matrix can be found e.g. in Krauss [20], or in the simpler form in the Appendix A. However the solution of the inverse problem is not so trivial as there are 4 possible solutions that have to be tested for cheirality [21]. The essential matrix as well as the fundamental matrix satisfies the well known complanarity constraint [20], [21]:

$$x'^{\top} E x = 0 \tag{1}$$

where $x$ and $x'$ are the column vectors of homogenous image coordinates. The rank deficiency of essential matrix implies the following constraint:

$$\det(E) = 0 \tag{2}$$

The $E$ matrix has two non-zero singular values that are equal. This constraint can be expressed in the algebraic form as [22]:

$$2EE^{\top}E - \mathrm{tr}(EE^{\top})E = 0 \tag{3}$$

It is advantageous to compute the essential matrix using one of few available close-form solutions using minimal, i.e. 5, number of corresponding points [10], [23], [24], [25]. Using the close-form solution requires no prior approximation and can be easily tested for outliers using the RANSAC procedure. The algorithm developed by Stéwenius, Engels and Nistér [25] was adopted within proposed solution because of its relatively simple implementation. Using equation (1) and finding its four-dimensional null-space, the essential matrix can be parametrized with three unknowns $x$, $y$, $z$:

$$E = xE_1 + yE_2 + zE_3 + E_4 \tag{4}$$

Inserting equation (4) into (2) and (3) produces the system of 10 3rd degree polynomial equations in three unknowns. This system is solved using the Gröbner basis. Up to 10 solutions for $E$ exist but only the real ones are of the further interest.

### B. Detection of outliers

In the proposed approach the essential matrix is estimated inside the RANSAC procedure. This enables detection of outliers. Assume that for the subsequent image pairs in the sequence $N$ samples are chosen, each consisting of 5 point pairs. As there are up to 10 real solutions for $E$, in the worst case there can be $10N$ possible solutions. According to the typical RANSAC each point in the image pair is classified as inlier or outlier according to the specified threshold value. In the addressed solution a kind of locally optimized RANSAC [11] is used. All points in each sample get a score that is inversely proportional to the distance from the model value. If the distance is higher than the threshold, the score is zero and the point is classified as outlier. The sample with the highest total score wins.

The easiest way to score each point is to calculate distance to the epipolar line using (1) (see Appendix A). However it may happen that due to mismatching, a point that is projected near the epipolar line lies behind the camera. To avoid treating such points as inliers it was decided to recover the rotation matrix ($R$) and the translation vector ($t$) of the second image from each real $E$ [21] and to calculate coordinates of keypoints in three dimensional coordinate frame of the first image using intersection of rays (Fig. A1, Appendix B). Now only the keypoints with negative $Z$ coordinates are going to be tested further. Given $R$, $t$ and estimated 3D coordinates of keypoints, the projections to images are found so that the 2D euclidean distances to the measured locations can be computed.

Assume three consecutive image pairs: [$k$,$k$+1], [$k$+1,$k$+2], [$k$+2,$k$+3]. A certain feature is matched correctly in pair no. 1. Subsequently this feature is matched incorrectly in pair no. 2. As a result it is classified as outlier. Nevertheless this keypoint may not be rejected because it could happen that incorrectly matched feature in image $k$+2 is correctly matched with the feature in image $k$+3. As a consequence this keypoint is recognized as two separate keypoints and gets separate id's in images forming pairs 1 and 3. It won't be used in relative orientation of pair no. 2.

### C. Estimation by least square fitting

The sample with the best score provides good estimations of $R$ and $t$ but it does not take into account all inliers. To utilize all available information, the least square adjustment can be carried out using all points classified as inliers. The image coordinates of keypoints are treated as observations and explicitly related to the parameters in the form of well known colinearity equations (see e.g. [20]). As a consequence the system of nonlinear equation is formed. The elements of $R$ are not treated as parameters directly. To avoid possible singularities resulting from parametrization in terms of Euler angles [26], the entries of the $R$ matrix are expressed as functions of the elements [27] of a quaternion. Both quaternion and $t$ are assumed to have unit norms that leads to the additional constrains imposed on the parameters. Finally the 3D coordinates of all tie points complete the set unknowns. The $R$, $t$ and 3D coordinates of tie points resulting from the RANSAC should be accurate enough to linearize the equation system and subsequently solve it in only one iteration.

## IV. Sequence orientation

### A. Model-to-model transformation

As a result of the relative orientation the $R$ and $t$ are provided for each image pair in the sequence. Such relatively oriented image pair is called a model. Assume image $k$ forming the model with image $k$+1.The consecutive model is formed by images $k$+1 and $k$+2. Common points are now used to stitch both models. In this way the orientation of a short, 3 image, sequence is recovered. Subsequently the algorithm proceeds to the transformation of the third model based on the reference points that appear in the previously created block. Model stitching is carried out further, until all images are oriented.

Each model is oriented according to the Horn algorithm [26] and involves estimation of 7 parameters: 3 for the rotation, 3 for the translation and finally the scale. The Horn approach consists of the following steps. At first the centroids in both point sets are calculated. Coordinates of all points are reduced to respective centroids. Secondly the rotation that maximizes the dot product of vectors pointing from centroids to corresponding points is found. The rotation is parametrized in terms of four elements of the unit quaternion. Once the quaternion is known, it is possible to align vectors in both frames to make them nearly parallel. Vectors won't never be exactly parallel due to outliers. Finally the scale is recovered using translated and rotated vectors. Three different approaches to scale computation are proposed depending on which point set is assumed to have better accuracy. The approach of Horn deals with minimal 3-point case as well as with greater number of points. It fulfills the condition of the least sum of squared residuals. Finally no approximation of the parameters is needed.

It should be mentioned that in addition to the terrain points (keypoints), adjacent models have one additional common point, namely the projection center of the common image - $k$+1 in the later example. This point lies far away from the rest of points and certainly has a worse reliability. Applying a standard Horn solution would cause that even a small errors in 3D coordinates of the tie points can result in large residual of projection center locations incorporating relatively large errors to the estimated motion. Therefore during the minimization of the dot product the utilization of the model frame coordinates is preferred to the usage of coordinates reduced to their centroids. In fact such modification means that the translation is simply calculated, not estimated, hence the estimation of remaining four parameters is more reliable i.e. less sensitive to the influence of erroneous tie point locations.

During the model-to-model transformation the sparse point cloud of keypoints is being formed. Besides points used inside the Horn algorithm, each stitched model incorporates a set of new points. If this points appear also in the next model, they are used as the reference. However some points exist that appear only in one model. The correctness of the location of such points cannot be fully checked. As a result some erroneous points in the sparse point cloud appear.

### B. *Absolute orientation*

Until now the image sequence was oriented up to the scale, absolute rotation and absolute translation. If the missing parameters are to be recovered, the external information need to be utilized. Basically there are two approaches to provide the external orientation to images: direct measurement and geo-referencing through ground control points (GCPs) referred as the indirect approach. To measure the external orientation directly one can use GPS and inertial sensors. If the GPS is used alone, the estimation of absolute orientation parameters takes place using the coordinates of projection the centers recovered in the previous step and the reference trajectory line recorded by a receiver [9]. Geo-referencing through control points usually requires the manual measurement of terrain features, the coordinates of which are known from other survey. In case the terrain coordinates of control points are known from a geodetic survey the indirect approach is assumed to be more accurate. In the presented study the second approach was utilized as no GPS measurements were available. After completing the absolute orientation it is possible to smooth the results and increase accuracy by performing the bundle adjustment. In such a case the loop closures, if only present, can be taken into account for the further accuracy increase.

## V. Experiments

### A. *Preparatory works*

The accuracy of the motion recovery from single view sequences strongly depends on the imaging geometry. In case of corridor sequences, when camera looks forwards or backwards, the intersection angles between correspondent rays are narrow, leading to the large errors of tie point locations. As a result the recovered camera orientation tends to drift quickly. In contrast to the corridor sequences a sequences with camera looking perpendicular to the moving direction (aside-looking sequences) should allow to achieve a better accuracy. In the following tests only the motion recovery from the aside looking sequences is covered, however the algorithm is supposed to deal with the geometry of any kind.

To test the proposed approach the decision was made to acquire the image sequence of the test-field area located at the AGH University Campus (Fig. 1). This test field is equipped with the number of natural GCPs, that are to be used to evaluate the accuracy. GCPs are located mostly at the building façades. It was decided to use the wide angle camera to be able to capture the façades from top to bottom. In addition to the large overlap, even in case of wide baseline, the wide angle lens provides increased accuracy of depth component of the tie point location in space. This is of the fundamental importance for the process of model stitching as the drift is supposed to accumulate slower. Besides, the obtained sparse point cloud would have better accuracy than in case of using the narrow angle lenses. In addition to better accuracy the wide angle lens performs better when imaging in motion. It guarantees a large depth of field allowing imaging with small aperture and short exposure time. Taking all the above into consideration the Nikon D5200 camera with the Sigma 10-20 mm f/3.5 rectilinear lens was chosen as the imaging sensor. In addition to the acquisition of a high resolution 24 megapixel (4000 × 6000) images the sensor of the camera allows HD video recording. The focal length was set to 12 mm providing the horizontal viewing angle about 90 °. The principal distance was fixed by blocking the focusing ring. The camera was calibrated to determine the interior orientation parameters and the distortion.

### B. *Data acquisition*

Initially the tests involving acquisition of HD videos were made, but because of low quality of extracted frames it was decided to switch the camera to the time-lapse mode, choosing the highest possible frequency of 1 Hz. However it came out quickly that capturing images with the 1 Hz frequency makes the camera buffer stuck - the shutter is not released until the last image is saved. Lowering the frequency would either lengthen the imaging base, possibly leading to the problems with feature matching or force decreasing the cycling speed

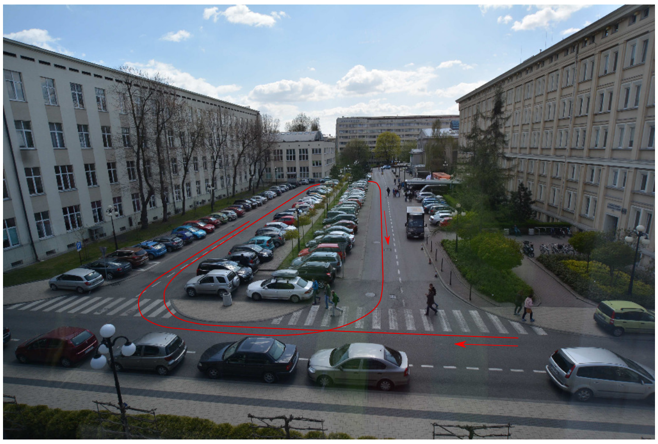


Fig. 1. Planned trajectory line imposed on the image of the test-field

resulting in extension of the overall acquisition time. The reasonable solution to avoid the above mentioned effects was to switch to the lower resolution of 13.488 megapixel.

The camera was fastened to the bike using a specially constructed device consisting of the 3 DOF head, allowing sequence acquisition from freely selectable viewing angle. The camera was inclined to look slightly upwards and perpendicular to the cycling direction. The test sequence was acquired in the aperture priority mode. The aperture value was set to 5 resulting in the exposure time between 1/2000 and 1/1000 second. The planned trajectory line is shown in the Fig 1. The test sequence was to have the shape of a loop. With the aim of comparison a part of the loop was to be cycled twice. The decision had to be made which side the camera should look at. Choosing the right direction provides convergent image configurations within all the turns and a good overlap. However in the case of the test-field the test were carried out in (Fig. 1) it was better to look left as to capture the façades lying closely to the trajectory line, providing advantageous distribution of keypoins. The disadvantages of such configurations are the occurrence of the divergent images within the turns leading to decrease in the overlap and occurrence of the narrow angles of intersecting rays.

After applying all the above mentioned settings the sequence of 195 images was acquired.

### C. Data processing

After collecting the data, the keypoint detection and matching algorithm was tested. The rough motion characteristic was known allowing to restrict the location of possible matches to the ROIs of a fixed size. The imposed constraint was supposed to reduce the number of possible outliers. The keypoint matching is followed by the relative pose estimation of consecutive image pairs. The least square relative orientation was tested but due to a very long computation time it was not applied in the final solution. Four image pairs, each located within the turns were not oriented properly due to the very high outlier rate and improper keypoint distribution. In this case the problem was fixed by adding some tie points manually.

Having the relative pose of the subsequent models estimated, the sequence formation was carried out. The first model in the sequence was chosen as a starting model. As the relative pose estimation constrains the base vector to equal 1, all the linear quantities calculated within this stage such as translations, residuals, errors are expressed in the unit of the length of the first base. The threshold parameter of the RANSAC procedure, i.e. the linear residual of the tie point, was set to 0.1.

For now the orientation of all of the images in the sequence was estimated up to the absolute quantities (translation, rotation, scale). To solve for the missing parameters and provide the accuracy analysis the 26 natural control points were used. Each control point was measured in the selected model (image pair). The accuracy assessment was provided by the residuals of the control point coordinates. Finally the sparse point cloud provided in the global coordinate frame was examined visually to look for previously undetected mismatches. The extent of the inconsistencies observed as a result of cycling the same part of the loop twice were to be analysed deeply.

## VI. Results

Despite applying the ROI-restricted matching a large number of outliers was observed in almost all image pairs (Fig. 2). During the RANSAC-based estimation of the relative pose it came out that the number of outliers considerably exceeds 50%. Besides the limitations of the monogenic phase matcher the reason of such a high outlier rate could be simply the content of the scene. For instance a number of corners appearing on the similar windows' frames are hard to be matched correctly. In addition the epiploar lines are nearly parallel to the horizontal edges of windows' elements so that even solving for relative pose cannot eliminate some outliers. It can also be noticed that there are quite a lot of trees in front of the façades (Fig. 1, Fig. 2, Fig. 3). As a results a number of a false keypoints is detected at the intersections of branches and twigs. Some of them are also incorrectly matched. The similar happens for keypoints detected in the reflections appearing in the window panes (Fig. 2). Also a lot of corner points detected at the grainy structure of the asphalt are matched incorrectly. Using the 5-point algorithm inside the RANSAC allows to eliminate most of the outliers. Fig. 2 and Fig. 3 are provided as an example.

It was decided to examine the influence of the drift (accumulation of the errors within the sequence formation stage) on the accuracy of the absolute orientation. The results are provided in the table 1. At the beginning the sequence of 10 images was oriented using four GCPs. The centimeter-level errors were obtained. Afterwards the number of images was increased until the appearance of a next group of available control points. Finally the sequence of 158 images was oriented. As no GCPs were measured in the further images, the last row of the table 1 represents the accuracy of the absolute orientation of the whole sequence. Performed analysis shows that generally

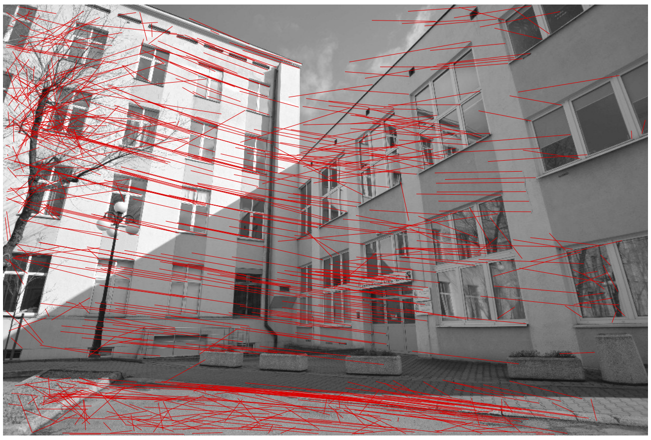


Fig. 2. One of the images from the southern part of the sequence and the vectors showing the displacement of the keypoints to the the next image. Results before applying the 5-point relative pose estimation algorithm inside RANSAC.

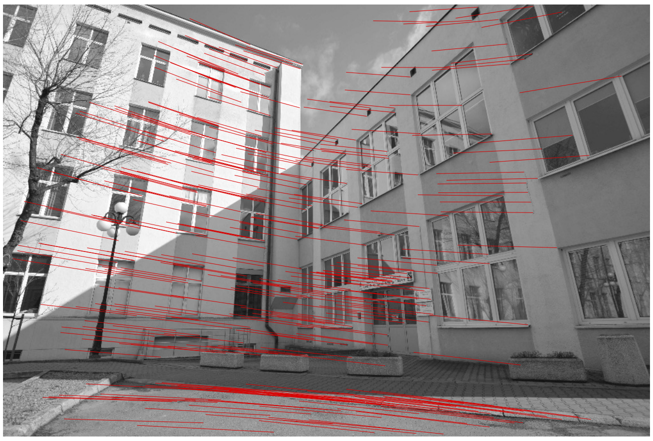


Fig. 3. One of the images from the southern part of the sequence and the vectors showing the displacement of the keypoints to the the next image after automatic rejection of outliers

while increasing the length of the sequence errors tend to increase, however not in the regular way. The worst results were obtained for the $Y$ coordinate and the best for the $Z$.

Fig. 4 shows the sparse cloud of tie points obtained as a result of the sequence formation. The colour of points changes from blue to red as to show the inconsistencies in the point cloud resulting from the orientation drift. The first matched keypoint in the first image pair is coloured in blue. The last matched keypoint in the last model is coloured in red. The black spots represent the location of projection centres. The black line represents the trajectory. The total length of the trajectory is 232.58 m. The arrows show the cycling direction

The façades are clearly visible in the cloud as well as the kerbs and the trees. During the data acquisition there was not as many cars parked as it can be seen in the Fig. 1. Few of them can also be visible in the cloud. There are also quite a lot of points that seem to be located inside the buildings. This tie points may represent mismatched keypoints located only in two images, occupying consistent epipolar lines. Such points pass the RANSAC testing, carried out within the model formation, but cannot be tested in the model stitching procedure. As a result of capturing certain parts of the scene twice, the inconsistencies in the resultant point cloud appear. To show them in details two parts of the cloud bounded by green rectangles are shown in the greater scale in the Fig. 5 and Fig. 6.

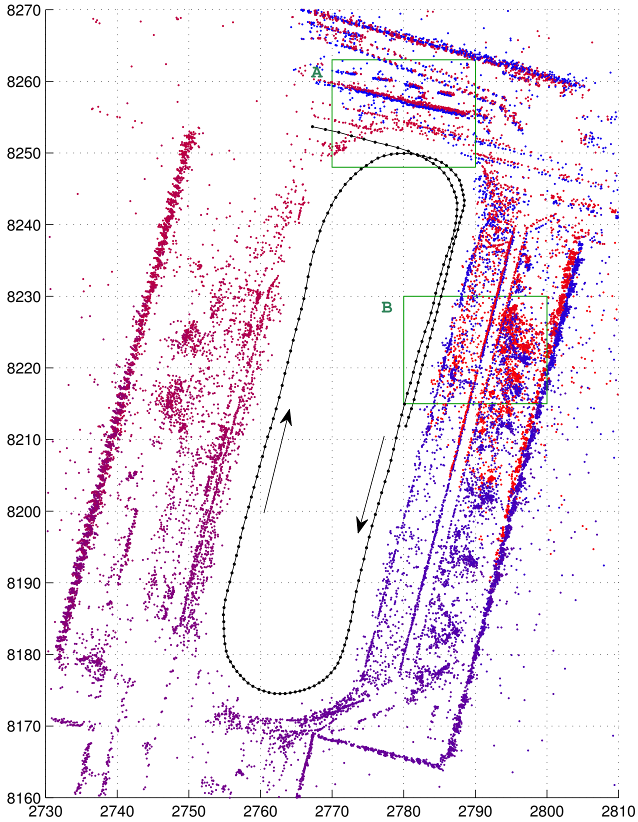


Fig. 4. The trajectory line and the point cloud. The arrows point the cycling direction. Green rectangles A and B mark the areas that will be referenced further. Orientation: north, units: meters.

The thickest strip of points in the Fig. 5 represents the front edge of the hedge, part of which is visible at the bottom of the Fig. 1. Points forming four segments parallel to the hedge are likely to be located at the crowns of the trees and the items used to shape them in the espalier-like form - see the very bottom of the Fig.1. In front of the hedge there are some points at the pavement. Some of them form a linear features that may represent kerbsides. Two linear features that lie behind the trees represent the railings of the ramp that belongs to the building the image in the Fig. 1 was captured from. Looking at the points located at the hedge it can be noticed that the red points are shifted with respect to the blue ones. The shift is about 40 cm in the south-north direction. When looking at the trees and railings and finally at the façade (Fig. 4) this shift seem to decrease.

TABLE I
ACCURACY OF THE ABSOLUTE ORIENTATION OF THE SEQUENCES OF A DIFFERENT LENGTH. THE LAST COLUMN PROVIDES THE RMS ERRORS OF THE 3D CONTROL POINT LOCATION.

| **Num. of images** | **Distance [m]** | **Num. of points** | $RMSE_X$ **[mm]** | $RMSE_Y$ **[mm]** | $RMSE_Z$ **[mm]** | $RMSE_P$ **[mm]** |
|---|---|---|---|---|---|---|
| 10 | 15.31 | 4 | 6 | 16 | 12 | 21 |
| 34 | 42.94 | 7 | 63 | 33 | 30 | 77 |
| 55 | 75.82 | 9 | 87 | 71 | 60 | 127 |
| 73 | 95.48 | 12 | 99 | 111 | 60 | 160 |
| 89 | 107.46 | 16 | 212 | 224 | 83 | 320 |
| 107 | 126.83 | 16 | 290 | 265 | 103 | 406 |
| 141 | 177.39 | 23 | 272 | 403 | 126 | 502 |
| 158 | 193.56 | 26 | 267 | 405 | 240 | 542 |

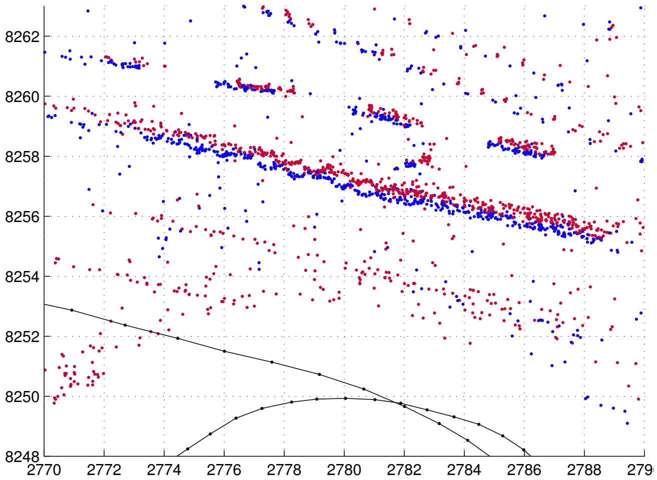


Fig. 5. Inconsistency of the point cloud within the area A

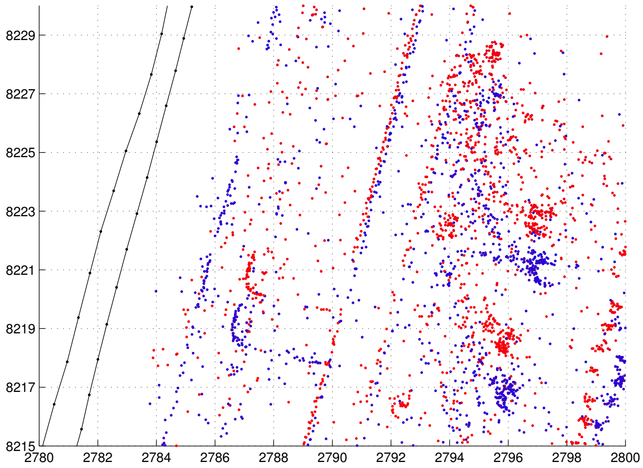


Fig. 6. Inconsistency of the point cloud within the area B

In the Fig. 6 the points located at the car body can be observed as well as linear features representing kerbs. There are groups of points representing trees that grow in the front of the building the façade of which can be seen in the bottom right corner of the figure. Looking at the points representing the car one can notice a considerable inconsistency in the point cloud. The red points are shifted about 2 meters with respect to blue points. The reason why the shift increased to such a high value can be the unfavourable imaging geometry (divergent camera axes, decreased overlap) at the north-east turn. Finally it can be found quite unexpected that no drift in the heading component of the angular orientation can be noticed - the blue and red linear features visible in Fig. 5 and Fig. 6 stay almost exactly parallel.

## VII. DISCUSSION

In this paper the new solution to the automatic orientation of single-view image sequences is proposed and the results of the tests conducted based on the data acquired from bike are presented. The solution assumes the calibrated camera case to achieve more robust performance in outlier detection and better accuracy. Modifications to the model-to-model stitching procedure are proposed as to achieve better reliability of the sequence formation, which result in more robust trajectory estimation.

The conducted tests allow the examination of certain steps of the proposed solution. The first step i.e. the feature detection and matching seem to perform quite well for images with a similar angular orientation. However it tends to fail for images captured within turns so that even the manual point measurement was to be carried out to fix the problem. To improve the keypoint matching firstly a more robust feature descriptors and matcher can be applied. Secondly the matching should be integrated with the relative pose solution. The $E$ matrix is quite accurately estimated using the proposed method, even in the presence of outliers, so that the equations of epipolar lines can be used to impose a stronger constraints for the feature re-matching (Appendix A). Afterwards features can be re-matched after solving for the essential matrix. Then the refined essential matrix is to be estimated and the solution can proceed in the iterative manner. Having the robustness improved one can think about improving the accuracy by using the sub-pixel corner measurement. Additionally the motion recovered within the process of model stitching can be smoothed using the bundle adjustment, however this approach may take quite a lot of computation time.

Assuming no real time application it is also better to select a different starting model for the sequence formation. Probably choosing the model near to the middle of the sequence would reduce the error as the drift is to accumulate on distances of about the half distance of the sequence.

The improvement of the orientation procedure is going to be followed by integration of other sensors like GPS or IMU. It would demand changing the imaging sensor from SLR to the industrial camera. Besides providing the time synchronization interface the industrial camera allows imaging at the higher frequency and at the same moment allowing real-time image data processing. Generally at this stage of the research the obtained results can be found satisfactory and consist a good starting point for developing a bike MMS equipped with the visual orientation unit. There exists a field to improve the robustness, accuracy and operational performance of the solution by improving both algorithms, implementation and a hardware.

## Appendix

### A. Essential matrix the and epipolar constraint for a calibrated camera

Assume that the two images of approximately the same scene were taken with calibrated camera (Fig. A.1). Assume this two images form photogrammetric model. Relative orientation of the second image with respect to the first image can be parametrized by the orthonormal rotation matrix ($R$) of the second camera frame and translation vector ($t$) of the second camera projection center (O'). The translation vector simply equals the base vector ($b$). The relative orientation can be estimated up to scale factor. Usually $b$ is assumed to be the unit vector.

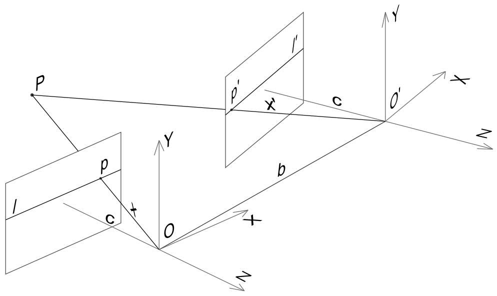


Fig. A.1. Two metric images forming the model

Point $P$ is located in the scene and projected into images to form points $p$ and $p'$. Assume two vectors $x$ and $x'$ that originate in projection centers $O$ and $O'$ and point at points $p$ and $p'$. Cooridinates of those vectors are given in the reference frame of respective cameras so that the third coordinate is equal to the principal distance of camera with the minus sing and for metric images is assumed to be the same, i.e.:

$$x = \begin{bmatrix} \xi \\ \eta \\ -c \end{bmatrix}, x' = \begin{bmatrix} \xi' \\ \eta' \\ -c \end{bmatrix} \tag{A.1}$$

Now the coplanarity constraint reads as follows:

$$x^\top (b \times Rx') = 0 \tag{A.2}$$

Coordinates of $b$ fill the elements of skew-symmetric matrix $B$:

$$B = \begin{bmatrix} 0 & -b_z & b_y \\ b_z & 0 & -b_x \\ -b_y & b_x & 0 \end{bmatrix} \tag{A.3}$$

so that:

$$x^\top BRx' = 0 \tag{A.4}$$

and consequently:

$$x'^\top Ex = 0 \tag{A.5}$$

where $E$ is the essential matrix. The projection center $O$ and point $P$ define a ray that is projected into second image as the line $l'$ (Fig. A1). The equation of this line is obtained by inserting the coordinates of $x$ into equation (A.5). In the similar way the equation of the epipolar line $l$ can be derived.

### B. Intersection

Assuming the $R$ and $t$ are known, it is now possible to estimate the coordinates of point $P$ in the model coordinate frame, i.e. the coordinate frame of the first camera (Fig. A.1). If points $p$ and $p'$ represent two correctly matched keypoints, theoretically both rays should meet in point $P$. However due to measurement errors rays are not going to intersect. Knowing the $x$ and $x'$ vectors (A.1) the location of point $P$ can be determined by least square solution. Collinearity equations for two corresponding keypoints are as follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_P = \lambda_1 \begin{bmatrix} \xi \\ \eta \\ -c \end{bmatrix} \tag{B.1}$$

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_P = \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix} + \lambda_2 R \begin{bmatrix} \xi' \\ \eta' \\ -c \end{bmatrix} \tag{B.2}$$

where $\lambda_1$ and $\lambda_2$ are unknown scale coefficients. After elimination of $\lambda_1$ and $\lambda_2$ from (B.1) and (B.2) followed by term's rearrangement the observed 2D coordinates of the keypoint can be written using explicitly elements of the relative orientation as the functions of unknowns:

$$\begin{bmatrix} \xi \\ \eta \\ \xi' \\ \eta' \end{bmatrix} = \begin{bmatrix} -c\frac{X_P}{Z_P} \\ -c\frac{Y_P}{Z_P} \\ -c\frac{R_{1,1}(X_P-b_x)+R_{2,1}(Y_P-b_y)+R_{3,1}(Z_P-b_z)}{R_{1,3}(X_P-b_x)+R_{2,3}(Y_P-b_y)+R_{3,3}(Z_P-b_z)} \\ -c\frac{R_{1,2}(X_P-b_x)+R_{2,2}(Y_P-b_y)+R_{3,2}(Z_P-b_z)}{R_{1,3}(X_P-b_x)+R_{2,3}(Y_P-b_y)+R_{3,3}(Z_P-b_z)} \end{bmatrix} \tag{B.3}$$

This system of equations can be rewritten in the linear form. The solution provides coordinates of point $P$. In case of erroneous measurements in $x$ and $x'$ the projection of estimated $P$ point into images won't coincide with points $p$ and $p'$ so that the residual vectors will appear.

## REFERENCES

[1] A. Śluzek, M. Paradowski,"Is Visual Similarity Sufficient for Semantic Object Recognition?", *Computer Science and Information Systems (FedCSIS) Federal Conference on. IEEE,* Wrocław, 2012, pp. 167-173.

[2] R. J. Handley, J. P. Abbott, C. R. Surawy, "Continuous Visual Navigation - An Evolution of Scene Matching", *Proceedings of the 1998 National Technical Meeting of The Institute of Navigation,* Long Beach, CA, 1998, pp. 217-224.

[3] C. V. Tao, M. A. Chapman, B. A. Chaplin, "Automated Processing of Mobile Mapping Image Sequences", *ISPRS Journal of Photogrammetry and Remote Sensing,* 55, 2001, pp. 330-346, DOI: http://dx.doi.org/10.1016/S0924-2716(01)00026-0

[4] D. Griessbach, D. Baumbach., S. Zuev,"Vision Aided Intertial Navigation," EUROCow, Castelldefels, 2010.

[5] F. Fraundorfer, D. Scaramuzza, "Visual Odometry - Part II: Matching, Robustness, Optimization and Applications", *IEEE Robotics and Automation Magazine*, June, 2012, pp. 78-90, DOI: http://dx.doi.org/10.1109/MRA.2012.2182810

[6] Y. Xu, F. Chen, "Real-time and Robust Visual Navigation Localization Algorithm based on ORB", *Applied Mechanics and Materials,* Vol. 241-244, 2012, pp. 478-482, DOI: http://dx.doi.org/10.4028/www.scientific.net/AMM.241-244.478

[7] N. Snavely, S. M. Seitz, R. Szeliski, "Modeling the World from Internet Photo Collections", *International Journal of Computer Vision*, 80(2), 2007, pp. 189-210, DOI: http://dx.doi.org/10.1007/s11263-007-0107-3

[8] M. A. Fischler, R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM,* 24(6), 1981, pp. 381-395.

[9] J. Bartelsen, H. Mayer, "Orientation of Image Sequences Acquired from UAVS and with GPS Cameras", EUROCow, Castelldefels, 2010.

[10] D. Nister, "An efficient solution to the five-point relative pose problem", IEEE PAMI, 26(6), 2004, pp.756-770, DOI: http://dx.doi.org/10.1109/TPAMI.2004.17

[11] O. Chum, J. Matas, J. Kittler, "Locally Optimized RANSAC," *Pattern Recognition - DAGM,* Springer Verlag, Berlin, 2003, pp. 249-256, DOI: http://dx.doi.org/10.1007/978-3-540-45243-0_31

[12] Y.I. Abdel-Aziz, H.M.Karara, "Direct linear transformation from comparator coordinates into object-space coordinates in close-range photogrammetry," *Proceedings of the ASP/UI Symposium on Close-Range Photogrammetry,* Falls Church, VA, 1971 pp. 1-18.

[13] K. L. Ho, P. Newmann, "Detecting Loop Closure with Scene Sequences," *International Journal of Computer Vision*, 74(3), 2007, pp. 261-286, DOI: http://dx.doi.org/10.1007/s11263-006-0020-1

[14] FARO, http://blog-uk.faro.com/2013/08/mobile-mapping-system-do-it-yourself/

[15] MATLAB and Octave Functions for Computer Vision and Image Processing, http://www.csse.uwa.edu.au/ pk/Research/MatlabFns/index.html

[16] A. Noble, "Descriptions of Image Surfaces", PhD thesis, Department of Engineering Science, Oxford University, 1989, p. 45.

[17] C.G. Harris and M.J. Stephens, "A combined corner and edge detector", *Proceedings Fourth Alvey Vision Conference,* Manchester, 1988, pp 147-151.

[18] M. Felsberg and G. Sommer, "A New Extension of Linear Signal", *Processing for Estimating Local Properties and Detecting Features,* DAGM Symposium, Kiel, 2000.

[19] M. Felsberg and G.Sommer, "The Monogenic Signal", *IEEE Transactions on Signal Processing,* 49(12), 2001, pp. 3136-3144, DOI: http://dx.doi.org/10.1109/78.969520

[20] K. Kraus, "Photogrammetry - Geometry from Images and Laser Scans", Walter de Gruyter, Berlin, 2007

[21] R. Hartley, A. Zisserman, "Multiple View Geometry in Computer Vision," Cambridge University Press, 2003, DOI: http://dx.doi.org/10.1017/CBO9780511811685

[22] J. Philip, "A Non-Iterative Algorithm for Determining all Essential Matrices Corresponding to Five Point Pairs", *Photogrammetric Record,* 15(88), 1996, pp. 589-599.

[23] Z. Kukelova, M. Bujnak, T. Pajdla, "Polynomial eigenvalue solutions to the 5-pt and 6-pt relative pose problems", BMVC 2008.

[24] D. Batra, B. Nabbe, M. Hebert, "An alternative formulation for five point relative pose problem", *IEEE Workshop on Motion and Video Computing,* 2007, DOI: http://dx.doi.org/10.5244/C.22.56

[25] H. Stewénius, C. Engels, and D. Nister, "Recent developments on direct relative orientation", *ISPRS Journal of Photogrammetry and Remote Sensing,* 60, 2006, pp. 284-294, DOI: http://dx.doi.org/10.1016/j.isprsjprs.2006.03.005

[26] B. Wrobel, D. Klemm, "Über die Vermeidung singulärer Fälle bei der Berechnung allgemeiner räumlicher Drehungen," *International Archives of Photogrammetry and Rmote Sensing,* 25, 1984, pp. 1153-1163.

[27] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions." JOSA A, 4.4, 1987, pp. 629-642, DOI: http://dx.doi.org/10.1364/JOSAA.4.000629

DOI: 10.15439/2015F232
ACSIS, Vol. 5

# Fast Artificial Landmark Detection for Indoor Mobile Robots

Dmitriy Kartashov, Arthur Huletski
The Academic University
Saint-Petersburg, Russia
{dmakart, hatless.fox}@gmail.com

Kirill Krinkin
St.Petersburg State Electrotechnical University ”LETI”
St.Petersburg, Russia
kirill.krinkin@fruct.org

***Abstract*—Nowadays the big challenge in simultaneous localization and mapping (SLAM) of mobile robots is the creation of efficient and robust algorithms. Significant Number of SLAM algorithms rely on unique features or or use artificial landmarks received from camera images. Feature points and landmarks extraction from images have two significant drawbacks: CPU consumption and weak robustness depending on environment conditions. In this paper we consider performance issues for landmark detection, introduce a new artificial landmark design and fast algorithm for detecting and tracking them in arbitrary images. Also we provide results of performance optimization for different hardware platforms.**

## I. Introduction

LANDMARKS are generally defined as passive objects in the environment that provide a high degree of localization accuracy when they are within the robot's field of view [1]. Artificial landmarks may carry additional information about the environment and may be used to assist a robot in localization and navigation. Although this approach requires environment preprocessing, it makes developers free to choose a type of landmark and information it holds. Furthermore, a developer could use any design for the landmark that can be sensed in small (with respect to an entire environment) location, but since artificial landmarks are supposed to simplify extraction of location features, appropriate landmark should satisfy the following requirements:

- can be reliably detected in a given environment. Detection should be robust for bad lighting conditions, glares, wide spectrum detection angles;
- can be identified;
- can be easily created and fixed in a given environment.

It this paper we analyze existing approaches for landmarking and introduce new kind of color landmark for fast and robust detection. The paper is organized as follows. In the next section quick response codes [2] and detection methods for them are discussed. In Section III. new design for color landmark is introduced and detection algorithm is suggested. In Section IV. tracking approach is introduced. In Section V. algorithm performance and detection rate are discussed. Finally, in Section VI. a summary of this paper and some issues for further development are listed.

## II. Artificial landmarks

One possible option for artificial landmark is a visual printed landmark, e.g. QR codes. This kind of landmarks has several advantages comparing to RFID and other technologies: it doesn't consume power, requires only camera which most mobile robots are already equipped, cheap and easy to produce.

There are several ways to detect QR codes in an image. In [3] the Viola-Jones framework based on Haar features is used to detect QR finder patterns (FIPs). Found FIPs are aggregated into a graph by their size and distance between them. The graph is searched for 3-cycles that satisfy the orientation criterion and represent QR codes. Unfortunately, the QR code plane must be almost orthogonal to the camera axis to reach the claimed 90% detection rate. In addition, this algorithm gives no information about the QR code position in the space.

Another approach that is described in [4] utilizes a special line parametrization called PClines which is a variant of the Hough transform. This parametrization uses a parallel coordinate system and allows faster accumulation than the basic Hough-transform method. The image is searched for the specific parallel lines pattern that is declared as QR code. The algorithm can handle various orientations of QR codes in images, tolerant to uneven illumination and allows real-time processing but is unable to detect QRs from far away or in blurred images. It's also unclear whether the algorithm can detect several QRs on the same image.

Some other algorithms (e.g. [5], [6]) assume that there is only one QR code in the image, so they are not suitable for our problem as the robot may see several landmarks simultaneously.

In order to estimate performance and resource requirements a variant of the QR detection algorithm described in [3] have been implemented. It uses cascade classifier to find both the finder and alignment patterns in the detected QR code. The last is necessary since 4 points are needed in order to compute position of the landmark in the space, but the existing algorithm gives only 3 points. Experiments have shown that the usage of bare QR codes as landmarks has following significant drawbacks:

- it's rather difficult to detect and extract bare QR code when it is far enough (more than 1 meter, which is quite often condition for indoor mobile robots);

978-83-60810-66-8/$25.00©2015, IEEE

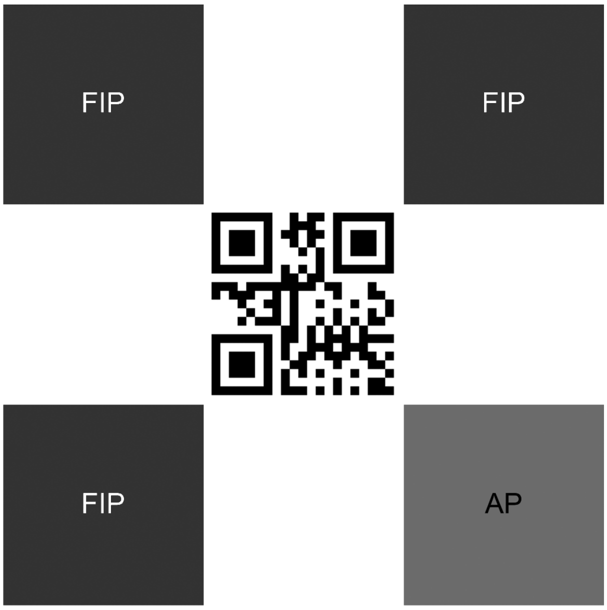


Fig. 1. Landmark design

- detection quality turns out to be very sensitive to the angle between camera and QR code plane.

So it was decided to create a new type of landmark that can be easily and reliably detected (i.e. chance of detection doesn't depend much on the camera position) and allows to carry extra information (like position or point identifier in the environment, or even instructions for mobile robots).

Various types of such visual landmarks are discussed in [7]. Besides good discussion of existent landmark types, this work introduces a landmark design based on QR code. According to it QR code is placed into blue rectangle that has three colored circles around it. Landmark of proposed design is scalable, special algorithm is introduced to extract inner area from outer circle for further QR code detection. Provided evaluation claims that landmark with diameter of outer circle equal to 20 cm can be reliably detected from 2 meters while its horizontal angle lies in range $[-60°; 60°]$.

The aim of our work is to propose a landmark that can be detected in broader horizontal angle range and has smaller size at the same time. The novel design was inspired by QR code detection algorithm described in [3] and [8]. The last one introduces artificial striped landmark. Two colors are used for stripes, each stripe has neighbours of different color.

## III. Landmark design and detection

In this section the new artificial landmark design and detection algorithm are described.

### A. Landmark layout

A general idea for a new landmark is based on QR code layout extended with color markup. The suggested layout consists of 3 blue squares called finder patterns (FIP) and one red square – alignment pattern (AP) – on the white background (in fact, any 2 easily distinguishable colors may be used). They are located on the landmark like FIPs and APs in QR codes (see Fig. 1). The significant feature of such layout is that the landmark looks like 4 light squares on a dark background in the saturation channel of the HSV color space in daylight (Fig. 3b). This allows to run some edge detection algorithm on the saturation channel to find contours in the image (Fig. 3c) and some of the detected contours are chosen to be FIP or AP candidates for the landmark.

### B. Detection algorithm

The full algorithm pipeline is shown in Fig. 2. Below we discuss each stage of the algorithm.

To select FIP candidates consider bounding rectangles of the detected contours in the RGB color space. For each candidate all pixel values within the bounding rectangles are accumulated separately for each channel. Then the following conditions should be checked:

$$\sum blue > a \cdot \sum red \quad \text{and} \quad \sum blue > b \cdot \sum green$$

If these conditions are satisfied then the contour is pushed into the list of FIP candidates. If the similar conditions are met for the red channel then the contour is stored as an AP candidate. The coefficients $a$ and $b$ in the formula above are called a color ratio and determine significance of the color component within the contour bounding rectangle: higher coefficient values discard more candidate contours at this stage. On the other hand, the farther from the camera the landmark is the lower the coefficient values should be to detect a FIP. The coefficient values from the range $[1.2; 1.6]$ are reasonable to use in practice. The output of this stage is shown in Fig. 3d.

Note that the found contour is not checked to be rectangular: due to optical or perspective distortions the specific shape of a FIP might be hardly recognizable. So there is no constraints

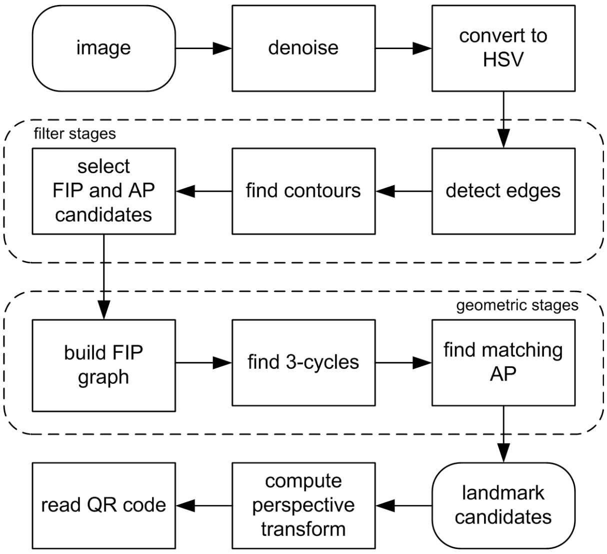
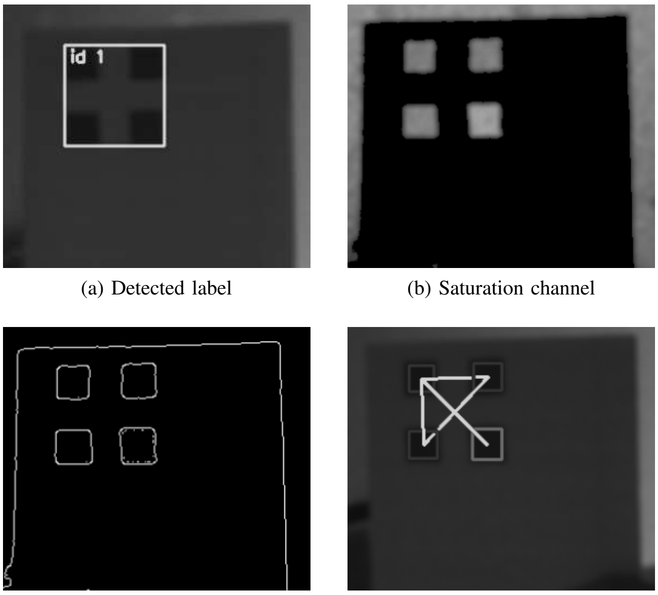


Fig. 2. Detection algorithm pipeline



(a) Detected label

(b) Saturation channel

(c) Edge detector output

(d) Detected FIPs and APs

Fig. 3. Detector output

on the landmark shape at all. In our case, discarding contours by color is more robust than by geometric features, though this method is very sensitive to the external lightning and FIP selector parameters should be adjusted depending on the environment conditions.

After obtaining the lists of FIPs and APs quadruples (3 FIPs and 1 AP) are formed from candidates using landmark geometric features. At first, the graph is constructed in the following way: vertices are FIPs and an edge connects FIPs if they satisfy two types of constraints:

1) the minimum and maximum distance between FIPs:

$$\begin{cases} D_{min} \cdot width < |X_{FIP1} - X_{FIP2}| \\ D_{min} \cdot height < |Y_{FIP1} - Y_{FIP2}| \\ |X_{FIP1} - X_{FIP2}| < D_{max} \cdot width \\ |Y_{FIP1} - Y_{FIP2}| < D_{max} \cdot height \end{cases} \tag{1}$$

where $width$ and $height$ are the size of a FIP pair; $X$, $Y$ – FIP position; $D_{min}$, $D_{max}$ – constraint coefficients;

2) the difference in height and width:

$$\begin{cases} W_{min} < width_{FIP1}/width_{FIP2} < W_{max} \\ H_{min} < height_{FIP1}/height_{FIP2} < H_{max} \end{cases} \tag{2}$$

where $width$ and $height$ – FIP size; $H$, $W$ – size ratio coefficients.

Parameters in these constraints may vary and their actual values depend on the layout of the landmark. In our case a distance constraint is from 1 to 3 FIP sizes and a FIP size ratio is from the range $[0.5; 1.5]$. The resulting graph is searched for 3-cycles that are considered to be landmark candidates.

Finally, the last component of the landmark is added – AP. From all of the AP candidates we choose one that meets the constraints on the location (the top-left corner or center must be inside the bounding rect of the three selected FIPs) and which size is closest to the selected FIPs. Note that the constraint on the top-left corner doesn't allow detecting upside down landmarks. If an appropriate AP is found then the landmark candidate is accepted.

The output on each stage of detection is shown on the Fig. 3. Note that a small saturation threshold is used on Fig. 3b. It amplifies the difference in intensity of light and dark regions on an image and slightly improves the edge detection quality. Fig. 4 shows detection with weak constraints – it results in many false positive FIP detections, but the landmark still can be accurately detected.

There are several parameters that can be adjusted in the detection algorithm, but the exact set of parameters depends on the filtering that is applied to the input image:

- geometric constraints – the distance between FIPs and APs and their size ratio;
- color constraints – the color ratio used in FIP detection;
- filter parameters – for example, Canny's threshold, saturation threshold, blur kernel size, etc.

Given the location of FIPs and APs the perspective transformation can be computed in order to get orthogonal projection of the landmark and its position in the space. A QR code located in the center of the landmark can be extracted using that orthogonal projection and can be decoded using any QR code decoding algorithm (e.g. [3], [4]). It's clear that it still can't be done from far away, but the robot can always drive up to the landmark if it knows where the landmark is. An example of QR code extraction is shown in Fig. 5. Some postprocessing has been applied to the extracted QR code to get the image in Fig. 5b and most bar-code readers can decode the resulting QR, though the additional rectification may be applied.

## IV. Landmark tracking

Landmark tracking is the next step for increasing detection robustness and quality. A robot can change the physical posi-

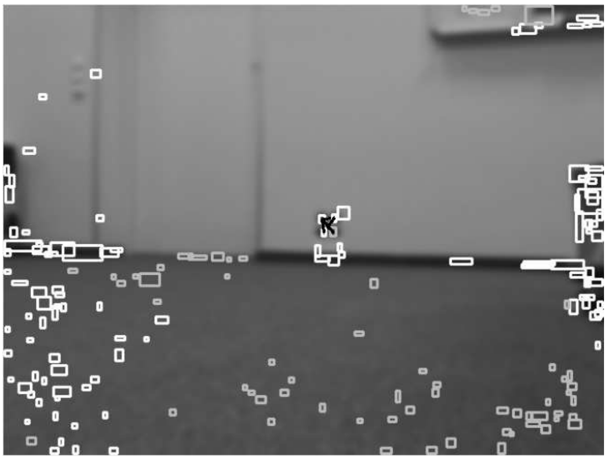
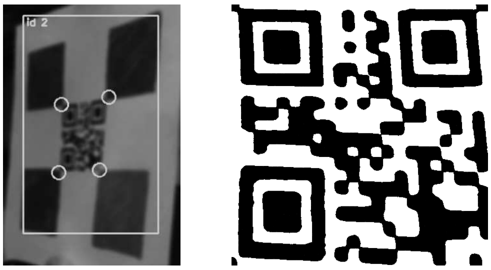


Fig. 4. Detection with weak constraints. The landmark in the center of the image still can be detected



(a) Detected label (b) Extracted QR code

Fig. 5. QR code extraction

tion of observation relatively smooth so the detected landmarks in the camera video stream will be relatively close on the two consecutive frames and therefore the detected landmarks may be tracked in a sequence of images. This can be done in the following way:

1) Initially there is an empty object (landmark) pool. Lifetime is assigned to every landmark and allows to keep the landmark in the pool for some time even if it isn't detected in several frames. The landmark is kept in the pool while its lifetime is less than the fixed maximum value.
2) Processing the next frame obtains a new set of landmarks. There are 3 different cases possible:
    a) If the object pool is empty then all new landmarks are stored in the pool and assigned ids.
    b) If the new set of landmarks is empty and the pool is not then the lifetime of all landmarks is increased and obsolete landmarks are removed.
    c) If both the pool and new landmark set are non-empty then it's required to solve the assignment problem where the pool landmarks are agents and the new landmarks are tasks (or vice versa), and the cost is the distance between a pool landmark and new one. So the total distance between old and new landmarks is minimized.
3) The obtained solution is checked to meet certain restrictions:
    - mapped landmarks are of similar sizes;
    - distance between them doesn't exceed the maximum value.

    If the mapped landmarks meet these restrictions then the position of the pool landmark is updated and its lifetime is reseted. Otherwise the lifetime of the pool landmark is increased and the new landmark is added to the pool.
4) All new landmarks that don't map to the pool landmarks are added to the pool.

Landmark tracking example is shown in Fig. 6. This algorithm is pretty standard and has 3 parameters that can be adjusted:

1) the landmark maximum lifetime depends on camera frame rate, but total time of 1 second looks appropriate;
2) the maximum distance between landmarks;
3) the landmark size ratio.

Specific settings of these parameters depend on robot physical features: maximum speed, camera frame rate, etc.

## V. Evaluation

In this section the detection algorithm performance and robustness as well as some algorithm drawbacks are discussed.

### A. *Performance*

The detection algorithm is supposed to be used on simple mobile robots with limited resources, so the testing environment based on popular low cost, credit-card sized computer Raspberry Pi [12] and Robotic Operation System (ROS) has been created.

ROS (Robot Operating System) [11] is a framework for robot software development that provides various system services such as hardware abstraction, low-level device control, implementation of commonly used functionality and message-passing between processes. Set of ROS processes is represented in a graph architecture where processing takes place in nodes that may receive and post sensor, control, planning and other messages. Simplified structure of the developed robot is shown in Fig. 7. The detector node works as a service that is polled by the main controller from time to time. Note that

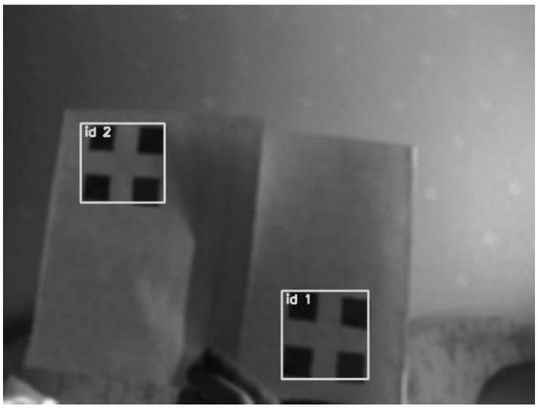


Fig. 6. Tracking example

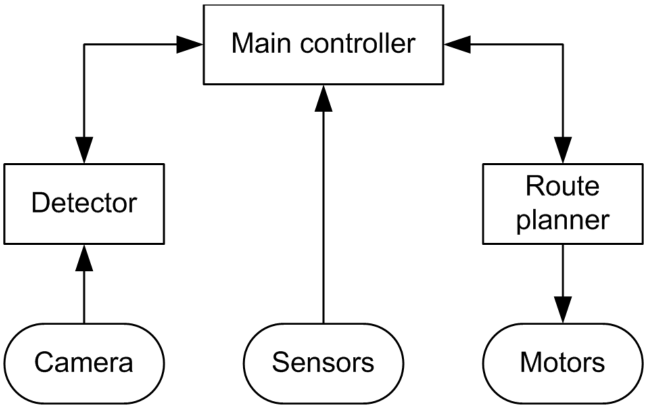


Fig. 7. Simplified robot structure

TABLE I
DETECTION ALGORITHM PERFORMANCE (FILTERS)

| | 640 × 480 (0.3 MP) | | | | 1280 × 720 (0.9 MP) | | | | 1920 × 1080 (2.1 MP) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Blur | HSV | Canny | Total | Blur | HSV | Canny | Total | Blur | HSV | Canny | Total |
| Intel Core i5 | 3.462 | 5.136 | 2.167 | 12.784 | 10.262 | 15.656 | 7.216 | 43.730 | 22.961 | 34.997 | 17.879 | 197.032 |
| GeForce 650M | 1.893 | 0.418 | 1.776 | 10.046<br>+2.258 | 4.845 | 0.798 | 4.058 | 23.855<br>+5.714 | 6.143 | 1.326 | 5.932 | 45.910<br>+11.976 |
| HD Graphics 4000 | 2.153 | 0.501 | 1.874 | 25.661<br>+17.791 | 7.098 | 0.895 | 4.777 | 72.584<br>+55.751 | 10.876 | 1.643 | 8.480 | 191.327<br>+128.117 |
| Raspberry Pi CPU | 310.192 | 66.863 | 69.340 | 509.667 | 1123.2 | 191.653 | 213.537 | 1639.81 | 2737 | 441 | 493 | 3955 |
| Raspberry Pi GPU | 13.430 | 2.600 | 11.921 | 116.330<br>+8.744 | 28.901 | 6.093 | 29.974 | 258.378<br>+46.417 | – | – | – | – |

the ROS itself is not a real-time OS and message passing introduces some overhead.

In the initial implementation of the detection algorithm all image processing operations have been performed by OpenCV framework [10]. The performance of this algorithm implementation on the Raspberry Pi that utilizes 700 MHz CPU is around 1-2 FPS and when the detector is integrated in the whole robot-control system it handles only 0.5 frames per second. Profiling has shown that the most computationally expensive parts of the algorithm are denoising, HSV color space conversion and tresholding and edge detection. All these operations can be easily performed by GPU, so it was decided to transfer some parts of the algorithm to GPU.

There are several options for GPU computations:

- GPU-specific tools and technologies (e.g. nVidia CUDA) allow achieving high performance but only on the limited set of devices;
- OpenGL shaders are cross-platform but it's hard to perform non-image processing on shaders;
- using GPU assembler it's possible to achieve the maximum performance on the specific device but the development process is very effortful.

The second option has been chosen because it allows parallel development and testing on desktop computer and mobile robot. Currently only Gaussian blur, saturation extraction and Canny edge detector are ported to OpenGL. Algorithm testing has been performed on the following set of hardware:

- OpenCV implementation:
  - Intel Core i5 3210M, 2.5 GHz;
  - ARM1176JZF-S, 700 MHz (Raspberry Pi CPU);
- OpenGL implementation:
  - Intel HD Graphics 4000 (integrated);
  - nVidia GeForce 650M (discrete);
  - VideoCore IV (Raspberry Pi GPU).

Testing results for filter stages are shown in the Table I and for geometric stages – in the Table II. All timings are in milliseconds. In the Table I in column "Total" for GPU implementations the image transferring overhead is indicated.

The results show that for desktop systems the image transferring overhead is greater than the speed gain in filters especially for the integrated GPU, but as expected, on the big images GPUs are still more efficient than CPU.

On Raspberry Pi the usage of GPU gives 4.5x gain in overall performance that increase to 6x gain when increasing the image size. Unfortunately, the algorithm cannot process big images due to RPi limitations. Thereby the detector with GPU optimization can process 8-10 frames per second which is sufficient for simple mobile robots, but further optimizations may be investigated.

### B. Robustness

To estimate detection rate we use two landmarks with 9 and 15 cm side and determine the maximum angle between landmark and camera planes at which 95% landmark detection rate can be reached. Results are shown in Fig. 8. Detector parameters have been tuned for every measurement so the graph shows the maximum reachable angle for reliable detection. In practice, static detector parameters adjusted to the environmental conditions allow to detect 9 cm landmark in $[-60°; 60°]$ horizontal angle range from 1.5 meters and 15 cm landmark in $[-75°; 75°]$ horizontal angle range from 2.5 meters.

The maximum distance at which landmark can be detected is determined by camera resolution. We used 0.3 MP webcam for testing and 9 cm landmark is indistinguishable in the image from approximately 2.5 meters and 15 cm landmark – from 4.5 meter. Thereby the landmark size is determined by the used camera and working environment features. In relatively small rooms with good lightning a 10-12 cm landmark is sufficient.

TABLE II
TOTAL DETECTION ALGORITHM PERFORMANCE (640X480)

| | Filter | Geometry | Total | FPS |
|---|---|---|---|---|
| Intel Core i5 | 13.546 | 0.057 | 13.633 | 73 |
| GeForce 650M | 10.623 | 0.042 | 10.686 | 93 |
| HD Graphics 4000 | 26.545 | 0.043 | 26.616 | 37 |
| Raspberry Pi CPU | 516.773 | 0.685 | 517.627 | 2 |
| Raspberry Pi GPU | 115.427 | 0.705 | 116.43 | 9 |

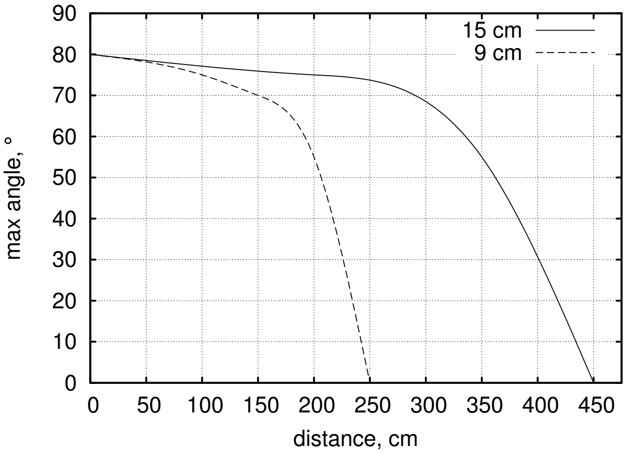


Fig. 8. Maximum angle for reliable detection

### C. Drawbacks

The proposed landmark detection algorithm suffers from image noise like any other algorithm based on edge detection. Various denoising algorithms can be applied (Gaussian blur in our case), but there is a trade-off between overall performance and image quality. For example, large blur kernels can affect edge detection quality, but good denoising algorithms like non-local means [9] run extremely slow.

Another disadvantage of the algorithm is that it relies on color features that depend on external lighting. The landmark looks like light squares on a dark background in daylight and completely different in dim light. So detection algorithm parameters should be adjusted depending on the ambient conditions.

## VI. Conclusion

The proposed landmark fits good landmark criteria: it is easy to create, set up, detect and identify. Although identification often requires approaching the landmark to read QR code, the fact of landmark(s) presence narrows amount of potential positions during localization process.

Described design and detection algorithm have various opportunities for enhancement:

- adaptive filters that can adjust their own parameters depending on environmental conditions;
- video stream for image stabilization, noise reduction and label tracking;
- experiments with colored FIPs and APs width can be performed to increase reliable angle range.

All programs and algorithm implementations are published as Open Sources Software and can be accessed by the following link: http://github.com/OSLL/landmark-detection.

## Acknowledgment

Authors would like to thank JetBrains company for provided support and materials for working on this research. The paper has been prepared within the scope of project part of the state plan of the Board of Education of Russia (task # 2.136.2014/K).

## References

[1] R. Siegwart, I. R. Nourbakhsh, D. Scaramuzza, Introduction to Autonomous Mobile Robots. *MIT Press*, 2011, p. 453.

[2] International Standard ISO/IEC 18004:2000 Information technology – Automatic identification and data capture techniques – Bar code symbology – QR Code, 2000.

[3] Luiz F. F. Belussi, Nina S. T. Hirata, “Fast Component-Based QR Code Detection in Arbitrarily Acquired Images”, *Journal of Mathematical Imaging and Vision*, vol.45, no.3, Mar. 2013, pp. 277-292.

[4] Gabriel Klimek, Zoltan Vamossy, “QR Code Detection Using Parallel Lines”, *Computational Intelligence and Informatics (CINTI), 2013 IEEE 14th International Symposium*, Nov. 2013, pp. 477-481.

[5] Yunhua Gu, Weixiang Zhang, “QR Code Recognition Based On Image Processing”, *International Conference on Information Science and Technology*, Mar. 2011, pp. 734-736.

[6] Yue Liu, Mingjun Liu, “Automatic Recognition Algorithm of Quick Response Code Based on Embedded System”, *In. proc. of the Sixth International Conference on Intelligent Systems Design and Applications*, Oct. 2006, pp. 783-788.

[7] Hao Wu, Guohui Tian, Peng Duan, Sen Sang, “The Design of a Novel Artificial Label for Robot Navigation”, *Proceedings of 2013 Chineseintelligent Automation Conference*, pp. 479-486.

[8] Kuk-Jin Yoon, Gi-Jeong Jang, Sung-Ho Kim, In-So Kweon, “Fast Landmark Tracking and Localization Algorithm for the Mobile Robot Self-Localization”, *IFAC Workshop on Mobile Robot Technology*, 2001, pp. 190-195.

[9] Antoni Buades, Bartomeu Coll, Jean-Michel Morel, “A non-local algorithm for image denoising”, *In proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2005, pp. 60-65.

[10] OpenCV website, http://opencv.org

[11] ROS.org — Powering the world’s robots, http://www.ros.org

[12] Raspberry Pi, http://www.raspberrypi.org

DOI: 10.15439/2015F383
ACSIS, Vol. 5

# Logical Structure Recognition of Diagram Images

Jerzy Sas
Wroclaw University of Technology
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
Email: jerzy.sas@pwr.edu.pl

Urszula Markowska-Kaczmar
Wroclaw of Technology,
Wyb.Wyspianskiego 27, 50-370 Wroclaw, Poland
Email: urszula.markowska-kaczmar@pwr.edu.pl

***Abstract*—This document presents a method of a logical links structure recognition between elements on diagrams. The applied approach intuitively mimics a human way of recognition that relies on merging already found connectors into more complex ones. This procedure is modeled by our method where simple and obvious connectors and gradually extended to more complex structures. Each iteration may lead to modification of connectors set obtained so far. The modifications are managed by a rules set describing logical and graphical constraints that should be satisfied by the connectors structure. If the extension leads to violation of constraints defined by the rules then the modification is not carried out. In this way, the recognized diagram structure is consistent with the assumed principles. The method was experimentally validated using the set of diagrams from three domains. In conclusions, method's advantages and drawbacks are discussed.**

## I. Introduction

In the last years there is a big interest in similar content image retrieval. There is plenty of research in this area. A deep survey is presented in [1]. The similar content image retrieval can be very helpful in automatic image annotation, in story illustration, copy detection, web image search and art image analysis.

It can be also helpful in searching similar text documents if they contain images. Usually, images illustrate the content of document. They contain information included in document in condensed form and there is less problems with unambiguous comparison of expressed idea. This explains a strong interest in its application to similar patents search in order to speed up the procedure of patenting and to protect intellectual property rights. For automatic querying it is necessary to convert the information in the images into a high-level description. Usually, in the case of technical documents, images represent engineering drawings, diagrams, algorithms, operations and processes shown as charts. Images of this kind present the structure consisting of certain elements and connections between them. Automatic recognition of the connections structure is the first step towards further automatic analysis or even machine understanding of images. The problem is important and challenging, because the documents have highly-complex structure, tabular and graphical information is embedded and they contain conflicting technical jargon. Processing embedded images can aid to solve the problem.

Typically, to apply such approach it is necessary to find images in the whole document. Then, images are processed in order to classify them to various classes: charts, diagrams, schemes, flowcharts, plots and photos. Next, a method dedicated to a given class of images is applied in order to recognize particular elements and their interconnections.

It is worth noting that such image interpretation allows to write the content of an image in the electronic form, which facilitates its storage and further processing and comparison.

In our research we focus on connectivity of elements in diagrams and flowcharts. The methods dedicated to this kind of graphics have to find: a) types of elements shown (various depicted shapes), b) segments of lines not belonging to found shapes, and c) connections created by these segments. Because diagrams usually contain texts embedded in diagram elements, it is also necessary to detect text areas, recognize it by applying OCR techniques and finally assign recognized texts to graphical diagram elements.

The aim of the research described in this paper was to find a method that is able to retrieve logical links between elements depicted in the diagram. This logical structure is then expressed in an XML file. It is a difficult task especially when we consider that one connection may exist between more than two elements and the line segments constituting connectors can intersect.

The paper consists of six sections. The next section describes related works. Section III formulates problem to solve. The subsequent section presents the developed method. Section V experimentally validates our approach. Finally, some conclusions and recommendations related to further works are presented.

## II. Related Works

Early survey of works in this area is described in [2]. The author writes that diagram recognition faces many challenges, including the great diversity in diagrammatic notations, and the presence of noise and ambiguity during the recognition process. Despite the flow of time from the year of this publication, all mentioned above features characterizing diagram interpretation constantly cause problems in chart recognition now.

The paper [3] reviews research from the last decade. The authors present the whole process of chart recognition: chart segmentation, chart classification, chart interpretation and discuss existing solutions.

Relatively many works are devoted to online flowcharts recognition. In [4], the analysis to label each stroke of the flowchart and to group the strokes depending on the symbol they belong to is presented. The same area of research is

978-83-60810-66-8/$25.00©2015, IEEE

represented in the paper [5]. In this paper the search for a suitable interpretation of the input is formulated as a combinatorial optimization task containing the max-sum problem. The recognition pipeline consists of two main stages. First, groups of strokes possibly representing symbols of a sketch (symbol candidates) are segmented and relations between them are detected. Second, a combination of symbol candidates best fitting the input is chosen by solving the optimization problem. The work [6] also concerns online charts but is focused on hand-drawn electric circuit diagram recognition using 2D dynamic programming. The paper [7] presents another approach to hand drawn organizational diagrams that is based on Bayesian conditional random fields (BCRFs) that jointly analyzes all drawing elements in order to incorporate contextual cues. The classification of each object affects the classification of its neighbors. BCRFs allow flexible and correlated features. The online recognition of diagram is mainly applied in order to automatically check student tests.

Currently there is a great interest in flowchart recognition in the context of patent search. The paper [8] describes measures for assessing the effectiveness of flowchart recognition methods in the context of patent-related use cases. A survey of approaches can be found in [9]. A system for semi-automatic chart ground truth generation is introduced in the paper [10]. Using the system, the user is able to extract multiple levels of ground truth data.

Some works are devoted to chart recognition in documents. They apply various classification methods. In [11] spiking neural networks are used. The paper [12] presents a system for recognizing a large class of engineering drawings characterized by alternating instances of symbols and connection lines. The class of considered images includes domains such as: flowcharts, logic and electrical circuits, and chemical plant diagrams. The output of the system includes a list identifying the symbol types and interconnections. It may be used for design simulation or as a compact portable representation of the drawing. The method consists of two steps. First, domain independent rules are used to segment symbols from connection lines in the drawing image that has been thinned, vectorized, and preprocessed in routine ways. Then a drawing understanding subsystem works together with a set of domain-specific matchers to classify symbols and correct errors automatically. They also proposed an interface to correct residual errors interactively.

Another important problem in diagram recognition and its automatic interpretation is recognition of texts appearing on diagrams. Separation of textual and graphical layers simplifies the further diagram structure analysis by reducing a number of involved graphical elements. It also makes it possible to attach textual information attributes to detected graphical elements of the diagram. In our approach we used the text separation method described in [13]. The method consists of three stages. In the first stage the text region candidates are elicited based on connected components analysis and some simple geometrical properties of connected components clusters. At the second stage, pattern recognition methods are applied to the set of candidates to discriminate between true text areas and other "false" candidates. Finally, OCR is applied to candidate regions and the final text region set is refined based on the analysis of the contents of the OCR-recognized strings.

## III. Problem formulation

In the further part of this article by a *diagram* we will mean a drawing that shows a set of entities - *diagram elements* (DEs) and connections between them. In the literature, the term "diagram" is used interchangeably with "chart", but diagram seems to be more general. We will be considering diagram images created with appropriate software or precisely drawn manually using drawing tools (rulers, drafting templates) and then converted into raster images by scanning. Hand-sketched diagrams are out of the scope of this article due to big inaccuracies appearing in this type of drawings. Our aim is to retrieve the logical structure of the diagram, so that it corresponds to intents of the diagram author. Informally, by the logical structure of the diagram we mean here the links between diagram elements. Diagram elements represent various items appearing in the real world modeled by the diagram. Their meaning depends on the domain of application. In the case of program flowcharts they can represent: statements, code blocks, conditions, data sources etc. In a logical circuit diagram they represent gates and functional blocks like: registers, multiplexers, flip-flops, etc. In organizational charts their elements are usually officials or departments. Although the method described here can be applied to any kind of diagrams, we mainly focus on organizational charts and program flowcharts. The element of the diagram is depicted by a simple 2D geometric shape like: rectangle, circle ellipse, rhombus, diamond. There are some attributes assigned to diagram elements. The basic attributes of an element are the kind of 2D shape and the textual description (the text usually inscribed into the element shape). Additional attributes that can be easily retrieved from the diagram image are: the shape interior color, shape line color and the shape border line width. They can be meaningful in certain types of diagrams, in other types of diagrams they may be ignored. The methods used to recognize shapes appearing in diagrams and to evaluate their attributes will be shortly described in the section IV-I.

The link in the diagram represents a logical relation or an association between DEs. Links are graphically represented by polylines or sets of intersecting or connected polylines, which endpoints are in the close vicinity of DEs being connected. Depending on an application domain, various types of links can be distinguished. The simplest links are one-to-one links which represent the association between a pair of DEs. The many-to-many link is the more complex case that associates the larger set of DEs. The links can be undirected or directed. In directed links some polyline endpoints are arrows. The directed link usually indicates the information flow or organizational subordination. If directed links are used the many-to-many association may turn into one-to-many association where only the single endpoint of the link is not arrowed, while all remaining endpoints are arrows. The graphical representation of the link will be further called *connector*. The connector is therefore the set of polylines that intersect each other or constitute T-style junctions. Finding connectors that are known to represent one-to-one links is a simple technical problem. Also in the case where it is known that the connection of polylines belonging to a common connector is indicated by dots (or other graphical marks) placed at connection points, the problem only lies in reliable connection marks recognition. It is however much more complicated if we cannot assume that connections of polylines belonging to a connector are graphically marked by connection marks. In such case the recognition of diagram structure must be based on the trial to "guess" the diagram author intention. In the further part of the paper, we will describe the method that deals with this kind of diagrams.

We are considering here the method which starts with vectorized diagram image on its input. The vector representation of the original (raster) image is obtained by applying the sequence of image processing operations followed by the vectorization procedure that converts a binary image into the set of line segments (vectors). Let us also assume that DEs have been already successfully found. In our approach we used the shapes recognition method based on vector sequence matching to basic shapes defined either by rules or algebraically, as described shortly in section IV-I. The vectors constituting found DEs were identified and extracted from further considerations. Let $E = \{e_1, e_2, ..., e_N\}$ denote the set of found DEs and let $L$ denote the set of line segments (called later *edges*) not assigned to any detected DE. The edge is a pair of 2D points being its ends on the plane $l_j = (p_{0j}, p_{1j})$, $p_i = (x_i, y_i)$. Line segments in $L$ possibly belong to connectors. Formally, a connector is a subset of connected edges from $L$. Our aim at this stage is to gather as many as possible edges into disjoint subsets $C_i \subseteq L$ corresponding to connectors, while leaving as little as possible lines unassigned, i.e. belonging to the unassigned set $U$. The result is the family of subsets $\{C_1, C_2, ..., C_M\}$. The construction of connector sets can be considered as a rule-based process, where rules define some constraints that must be satisfied in order to put a certain subset of edges into a connector, as well as principles that force inclusion of some edges in a single connector. Rules determining principles of reasonable construction of connectors were derived from the analysis of diagram structures appearing on typical diagram images. The analysis has been carried out using the set of diagram images from various domains, that we used for testing of our diagram analysis methods. The rules defining the construction of a connector $C_i$ of $L$ are as follows:

- the elements of $C_i$ are coherent (i.e. for each pair of edges in $C_i$ there is a sequence of other edges in $C_i$, possibly empty, that connects them);
- if there is an edge in $C_i$ that has a vertex not shared with other lines in $C_i$ (i.e. it is the endpoint of a polyline which elements are within $C_i$) then it must be close to one of detected DEs from the set $E$, such a vertex is a *terminal vertex*;
- there are no two vertices of edges in $C_i$ that are terminal vertices and are close to the same DE (the connector consisting just of a single edge cannot link the DE with itself);
- there are no cycles in the graph defined by the set $C_i$, i.e. there are no polylines in $C_i$ that intersect with themselves;
- if there is an edge $l_a$ in $C_i$ that has a common vertex with another edge $l_b$ in $L$ and $l_b$ is not assigned to any other connector $C_j$ then $l_b$ must also belong to $C_i$ (no connectors ending in the middle of polylines);
- if $p$ is an internal vertex shared by two edges belonging to $C_i$ then it cannot be closer than $\epsilon$ to an internal vertex shared by two edges belonging to another connector $C_j$ (connectors cannot touch each other, except of the case that the terminal edge endpoint touches a terminal edge endpoint of another connector);
- each edge is uniquely assigned to one of subsets $C_i$ or to $U$ (edges are not shared by connectors);
- the longest path connecting two terminal vertices in $C_i$ cannot be longer than assumed threshold i.e. 8 (no connectors of very complicated shape);
- the angle between two adjacent edges in $C_i$ which do not share the common vertex with any other edge within the same connector is not smaller than the right angle (no acute angles in the polyline segments of connectors);
- width of each edge within $C_i$ does not differ by more than 30% from the weighted average line width in the connector;
- two elements $e_i, e_j \in E$ cannot be connected along more than one path within the single connector (they can be however connected by many paths, provided that they belong to different connectors).

Because it seems reasonable to merge connectors into more complex ones, as far as it does not lead to violation of rules presented above, then the ultimate aim is to partition the set $L$ into the family of subsets $\{U, C_i : i = 1, ..., M)\}$ so as to minimize the number of subsets $C_i$ with the additional constraint that the set $U$ does not contain any subset of edges that constitutes a valid connector.

For practical purposes related to automatic analysis of the diagrams it is essential not just to find connectors (being the sets of edges) but rather to determine the sets of connected DEs which are connected by individual connectors. Therefore, the final result of the structure recognition procedure is the family of sets of DEs, where each set defines elements connected by the single connector. Additionally, because each connector endpoint can be marked with an arrow, this information should be also retrieved and included in the data structure being the output of the recognition procedure. Each endpoint of the connector is described by a pair $(e, t)$ where $e \in E$ and $t \in \{true, false\}$ indicates whether or not the endpoint of the connector linked to $e$ is an arrow. The elements connected by the single connector $C$ can be specified by the multiset $V_C = \{(e_i, t_i) : e_i \in E, t_i \in \{true, false\}\}$. The expression $t = true$ denotes the arrow appearance and $t = false$

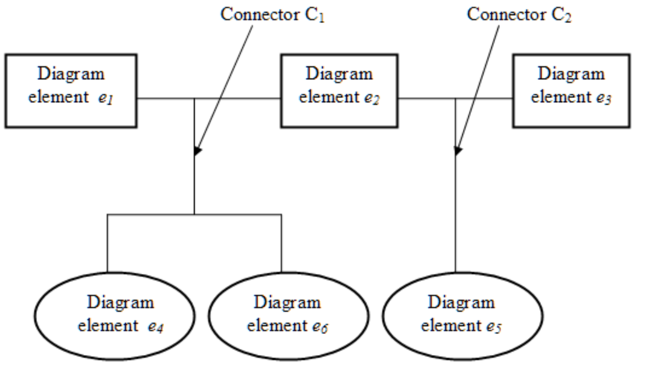


Fig. 1. Exemplary diagram consisting of two connectors sharing a common element $e_2$.

denotes arrow-less connector endpoint. Multisets are applied here instead of simple sets because the diagram element can be connected with itself. In such case there exist two connector endpoints associated with the same element. It leads to the appearance of the pair containing this element twice in $V_C$. Finally, the product of the diagram structure recognition is the set of connectors and the family of corresponding multisets of connected diagram elements:

$$(\widehat{C}, \widehat{U}) = (\{C_i : i = 1, ..., M\}, \{V_{C_i} : i = 1, ..., M\}), \quad (1)$$

where $M$ is the number of detected connectors. A DE may belong to more than one multiset $V_{C_i}$ if it is connected with other DEs by various connectors. The case where the central element is shared by two connectors is shown in Fig.1. The elements of the diagram are connected by two connectors: $C_1$ and $C_2$. In this case, the structure recognition procedure builds the following diagram description:

$$\begin{aligned}(\widehat{C}, \widehat{U}) = (&\{C_1, C_2\},\\ &\{\{(e_1, false), (e_2, false), (e_4, false), (e_6, false)\}, \quad (2)\\ &\{(e_2, false), (e_3, false), (e_5, false)\}\}).\end{aligned}$$

## IV. Description of the method

Unfortunately, the formulation of the diagram structure recognition problem does not lead to an efficient solution, other that "brute force" approach based on exhaustive search of all subsets of $L$, which is obviously infeasible in most practical cases. Therefore, we propose simplified suboptimal solution that leads to construction of a connector set $\{C_1, C_2, ..., C_M\}$, which however does not guarantee that the minimal number of connectors are found. On the other hand however, it applies some intuitive principles that humans typically apply when trying to read a structure on a diagram presented on an image. Experiments described in section V show that diagram structure recognized with the proposed algorithm is close to the human interpretation of test diagram images.

The approach taken consists in the observation that when a human tries to find connectors in a diagram by sight, it intuitively starts with long edges and tries to interpret them as "simple connectors" connecting pairs of DEs. Then a human tries to find "branches" that connect other DEs to previously found simple connectors. Next, one tries to find inter-connectors, i.e. polylines that connect previously found connectors. Finally, we (humans) try to merge already found connectors into more complex ones by finding intersecting lines belonging to various connectors that are candidates for merging. It is a process that starts with simple and obvious connectors and gradually extend them to more complex structures. This process can be modeled as a procedure implemented in a computer. Each stage outlined above is in fact an iterative operation that processes successive items (edges, polylines, simpler connectors - depending on the stage), where each iteration may lead to modification of the connector set obtained so far. The modification is however conditioned on the rules set presented in the previous section. If it leads to violation of constraints defined by the rules then the modification is not carried out. In this way, at each stage of the procedure we have the diagram structure that is consistent with the assumed principles.

Now the procedure will be described in more details. The input to the procedure is the set of diagram elements $E$ and the set of line segments (vectors) obtained from the raster image vectorization procedure (vectorizer). The applied vectorization procedure is based on the algorithm described in [14]. We will not deal here with the details of methods of shape recognition used to obtain the set $E$. They are briefly described in the subsection IV-I. In the proposed algorithm, the set of constants usually applied as thresholds are utilized. Values of these thresholds were estimated experimentally by analyzing a set of typical diagram images from the validation set. The selected validation set is disjoint from the testing set used in order to evaluate the method performance.

The connectors finding method consists of the following steps:

### A. Edge detection

The aim of this step is to create the set of edges $L$ from the set of line segments fetched by the vectorizer. We use the term "edge" to emphasize the difference in relation to the notion of simple line segment which is the direct product of vectorization. The line segments created by the vectorizer should not be used directly in further steps of the procedure. Our experiments showed that there are some troublesome artifacts, especially at intersections of relatively thick lines or at vertices of polylines. They are short line segments of the length of single pixels that connect longer line segments. Such geometrical structures need a kind of smoothing in order to obtain longer and straight line segments, most likely being "true" lines in the original diagram. Here we call such smoothed and merged line segments "edges". An example of an erroneous line segment structure created by the vectorizer is presented in Fig.2. The unwanted artifacts are indicated by blue circles.

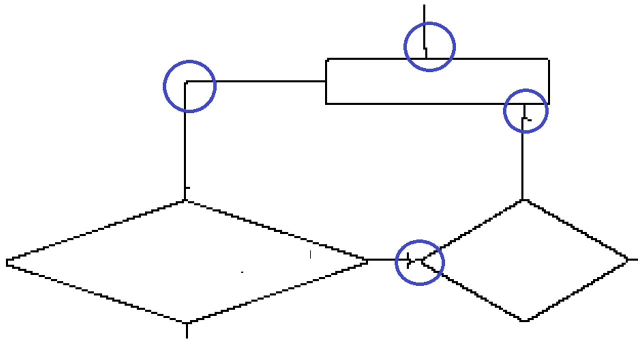


Fig. 2. Examples of inaccuracy artifacts introduced at the stage of image vectorization

The procedure takes the line segment that is not yet assigned to any element in $E$ nor to any edge in $L$ and tries to extend it to a longer edge by building a polyline being a sequence of interconnected line segments. In each iteration the algorithm tries to attach the next line segment that is adjacent to one of end points of the already created polyline. This segment is selected that is most collinear with the straight line approximating already created polyline. The attachment criterion is used to select candidates for extension. It takes into account the angle between the candidate segment and already approximated line and the length of the candidate. Short segments (being probably artifacts of the vectorization procedure) can be connected even if the angle is relatively big. The logical predicate used as the extension criterion is as follows:

$$\angle(e,c) < \alpha_{max} \vee len(l_t) \leq l_{min} \vee len(c) \leq 1.5 * w, \quad (3)$$

where $e$ is the edge (single line segment) approximating the polyline created so far, $c$ is the line segment - the candidate for extension, $l_t$ is the terminal line segment in the polyline that is adjacent to $c$ and $len(\bullet)$ is the length of the line segment. $w$ is the average width of the line segments already attached to the polyline. $\alpha_{max}$ was experimentally set to $10°$.

According to the criterion (3), the angle between connected segments must be small enough or at least one of adjacent segments (the candidate one or the terminal segment) is short enough. This alternative makes it possible to use residual vectors of the length of 1-2 pixels that are artifacts of the vectorization procedure. Such residual vectors often are oriented at big angles with the relation to its (longer) neighbors. The procedure iteratively tries to extend the polyline until no new segments can be attached. After each extension the new linear approximation of the polyline (denoted by $e$ in the formula 3) is evaluated by least square fitting of points on the polyline to the approximating line.

### B. Preparation to processing

The aim of this stage is to identify DEs that are close enough to endpoints of edges created in the previous stage. In this way it is possible to identify edges that are candidates for terminal edges of connectors. The terminal edge of the connector is the edge directly attached to the element connected by a connector. Then for all edge vertices the closest DE is found that is within assumed maximal allowed distance from the vertex. The tolerance is estimated depending of the shape and edge line widths. By analyzing the set of exemplary diagrams we assumed that the tolerance should be evaluated as $min(3 * max(w_e, w_s), 0.25 * s_{BB})$ where $w_e$ is the width of the edge line, $w_s$ is the width of the line of the DE shape and $s_{BB}$ is the smaller of $x$ and $y$ sizes of the bounding box enclosing the element. The angle between the closest element edge and the connector edge is also taken into account to avoid considering as very close an edge that is almost parallel to an edge of DE. As the result of this stage, each edge endpoint is annotated either with the index of the close diagram element or with "dummy" index, denoting that there is no close element to the edge endpoint. Additionally, the "edge structure" is created, that makes it possible to quickly find all edges in $L$ adjacent to a given vertex.

### C. Finding simple connectors

At this stage simple connectors are being found, where two DEs are directly connected by a single edge. It lies in finding edges with two endpoints marked with various diagram elements. Cases where the edge connects the shape with itself and is completely within this shape bounding box are excluded by applying one of constraints defined in Section III.

### D. Finding polyline connectors

This stage consists in finding edge sequences (a polyline consisting of edges) that connect two shapes. The procedure used here starts with a polyline consisting of a single edge that is not yet assigned to any connector. It iteratively tries to extend the polyline by attaching its left/right endpoint neighbors until no further extension is possible or the attached edge is connected to an element. Backtracking is applied in cases where there are many neighboring edges adjacent to the terminal edge in the polyline and extending edge selection in certain iteration leads to the polyline that neither can be further extended nor it terminates with the edge adjacent to any DE.

The procedure is being repeated, each time starting with an edge that is not yet assigned to any connector. If it leads to the polyline connecting two diagram elements then it is assumed to be the polyline connector. The new connector is then created and all edges are labeled as assigned to this connector.

### E. Finding inter-connectors

In the previous stages only such connectors (or fragments of connectors) were recognized which link pairs of DEs. In the next phase, new polyline connectors are tried that connect already found connectors with other connectors or diagram elements. The interconnector appears for instance in the diagram in Fig.1. At the first stage, the simple connector between elements $e_1$ and $e_2$ is found. The polyline connector linking elements $e_4$ and $e_6$ is detected in the second stage. In the current stage the interconnector that links this two simpler connectors will be recognized. The case of "branch"

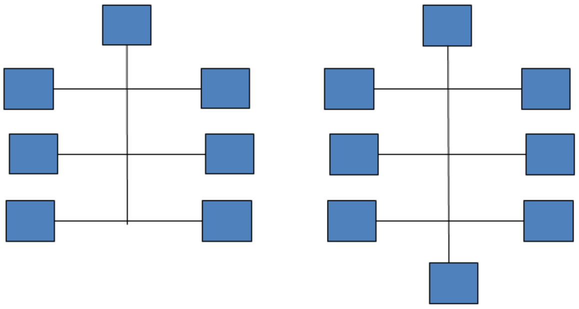


Fig. 3. Pine-like connector structures

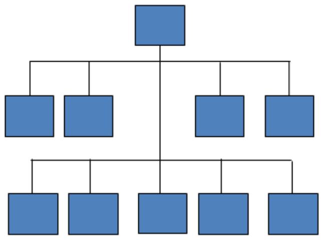


Fig. 4. Comb-like connector structures

that extends simpler connectors appears in Fig.1 in the case of the single connector that links elements $e_2$ and $e_3$. The branch consisting of the vertical edge in the right part of the diagram links the simple connector with the element $e_5$.

The procedure iterates until a single iteration does not result in any extension of the obtained connector sets. The single iteration in turn, consists of subiterations that iterate over all unassigned line segments, where each unassigned segment is tried to be extended into a polyline connecting two earlier detected connectors or a connector and a diagram element. Typically, this stage creates T-connections. T-connection is the connection of edges, where one edge perpendicular to another one touches them in the middle. The connection between the edge linking $e_2$ with $e_3$ and the vertical edge adjacent to $e_5$ is a typical T-connection.

The next two stages are aimed on merging connectors found earlier into more complex ones. Pairs of connectors that are candidates to merging must have intersecting edges. Merging all connectors having intersecting edges in many cases would lead to a structure not intended by the diagram author. Actually, the problem of simpler connector merging seems to be the hardest one in the process of diagram structure recognition. We distinguished two specific graphical configurations of edges that are typically used when composing diagrams and we perform connector merging only if the merged connector conforms to the one of these specific configurations. We called these configurations *pine-like* and *comb-like* structures. The structures are shown in Fig.3 and Fig.4

### *F. Constructing pine-like connectors*

In this step, the specific type of connector is detected which consists of a simple single edge that intersects other connectors. It is assumed that in such case the diagram creator intention was to depict the situation where elements connected by such a connector are connected each to another.

In order to find pine-like connectors, the simple connectors (being just single edges) are tested against intersection with edges of other connectors, let us call it *trunk*. All other simple connectors that intersect the trunk are merged to the group containing the trunk.

### *G. Constructing comb-like connectors*

The comb-like structure is presented in Fig.4. It consists of the vertical trunk connector intersected by one or more "combs". Comb is a connector with the principal horizontal edge to which a series of simple vertical edges (branches) are T-connected. Additionally, it is required that comb teeth are approximately of to the same length in the interval $< 0.1 * l_{pr}, 0.7 * l_{pr} >$ where $l_{pr}$ is the length of the principal comb edge. The comb-like structure must also have trunk - the vertical line being a connector that have a single DE at its top and the trunk length must be at least $2 * l_{t_{min}}$, where $l_{t_{min}}$ is the shortest tooth length of the comb. The procedure of comb-like connector construction consists in finding connectors that satisfy aforementioned conditions. The set of connectors that satisfies it is replaced in the set $\{C_1, C_{@}, ..., C_M\}$ by the product of the merge operation.

### *H. Merging connectors by dot-markers*

Finally, all these pairs of connectors are merged that contain lines that cross one with another, where there is a dot-marker at the intersection of lines belonging to various connectors. In order to consider two line segments as marked for merging, the following conditions must be satisfied:

- intersecting lines are approximately of the same width $(0.5 \leq w_1/w_2 \leq 2.0)$, where $w_1, w_2$ are widths of lines;
- lines are approximately of the same color, the color tolerance is defined for the components of 24-bit RGB color space;

$$\begin{aligned} R_{max} - R_{min} &< 30, \\ G_{max} - G_{min} &< 30, \\ B_{max} - B_{min} &< 30, \end{aligned} \tag{4}$$

- the angle between intersecting lines is approximately the right angle (with the tolerance range from 80° to 100°);
- the four diagonal pixels at the distance $\sqrt(2) * max(w_1, w_2)$ from the lines intersection point are closer in colors to the average lines color than to the background color.

The last condition is responsible for detecting the dot intersection marker at the lines intersection point.

### *I. Diagram elements recognition*

Because this paper is mainly focused on the recognition of connections between diagram elements, we will only briefly describe methods used in order to recognize diagram elements. We assume that diagram elements are: a) polygons, b) circles, ellipses or arches of ellipses and c) shapes being combination of a) and b), e.g. the symbol of a drum often used in flowcharts to denote mass storage. Methods used to recognize polygonal shapes are based on rules that define the mutual geometrical relations between line segments constituting polygon edges. For example, the parallelogram not being just a rectangle is defined as the sequence of 4 edges $(l_0, l_1, l_2, l_3)$ defined by their endpoints $(p_i^{(B)}, e_i^{(E)}), i = 0, ..., 3)$ that satisfy the set of constraints:

- $\mid p_i^{(E)} - p_{i \oplus 1}^{(B)} \mid \leq \epsilon$ for $i = 0, ..., 3$;
- either $p_i^{(E)} = p_{i \oplus 1}^{(B)}$ or $p_i^{(E)}$ is connected with $p_{i \oplus 1}^{(B)}$ by a chain of short line segments that are entirely included inside the bounding box defined by $p_i^{(E)}$ and $p_{i \oplus 1}^{(B)}$;
- $l_0 \parallel l_2$ and $l_1 \parallel l_3$;
- $\angle(l_0, l_1) \leq 90° - \alpha_{toll}$ or $\angle(l_0, l_1) \geq 90° + \alpha_{toll}$,

where $i \oplus 1 = (i+1) \mod 4$ and $\alpha_{toll}$ denotes the tolerance for right angles. The last constraint makes it possible to distinguish between rectangles and other parallelograms.

The procedure of shape recognition starts with a line segment from the vector set created by the vectorization procedure and successively tries to extend it to a sequence of segments, so that the constraints defined for allowed shapes are satisfied. It may happen that certain sequence of line segments satisfies constraints for more that single shape. For example, due to drawing inaccuracies, a quadrilateral found in the diagram may satisfy both constraints for the rectangle, the trapezoid as well as for the rounded vertices rectangle. In such a case, the measure of inaccuracy for all candidate shapes is computed and this shape is finally selected for which the inaccuracy measure is lowest. The procedure is repeated, each time starting with a next line segments that is not already assigned to any shape, until all unassigned line segments are tried.

In the case of ellipses and arches the procedure starts with a candidate edge and tries to extend it to a polyline that best approximates the ellipse fragment. Only axis-aligned ellipses are considered. Let $S = (l_1, l_2, ..., l_n)$ be a sequence of line segments that approximate an ellipse arch. At each stage the procedure tries to extend it with one of line segments that is adjacent to $l_1$ or $l_n$. Such extending segment is selected for which the ellipse approximation error is the lowest. The approximation error is the average distance of pixels constituting the polyline $S$ to the best fitting ellipse. The pixel set used for approximation is created by applying Bresenham line drawing algorithm to all lines in the set $S$. The axis-aligned ellipse is defined by four parameters: coordinates of the ellipse center $(x_c, y_c)$, the length of $x$-axis $d_x$ and the shape factor $a$ - the ratio of $x$ and $y$ axes lengths $a = d_x/d_y$. The best fitting ellipse is found using the method described in [15]. The parameters of the optimal ellipse as assumed to be within the reasonable ranges determined in relation to the image size, e.g. the ellipse $(x_c, y_c)$ center must be within the range defined by the image resolution and both ellipse axes must be not longer than the corresponding image size along $x$ or $y$ axes. If the parameters computed by the optimization procedure are out of these ranges then the approximation is assumed to fail and another extending line is tried. The extension is continued until the closed ellipse is obtained or no more lines can be attached. The final ellipse arch is accepted if it constitutes at least 50% of the complete ellipse. This acceptance threshold may seem to be high, but shapes that we consider here as diagram elements never consist of shorter ellipse fragments. On the other hand, setting too low value of the threshold may lead to false recognition of other shape elements as ellipse fragments.

## V. Experiments

### *A. Evaluation of diagram structure recognition accuracy*

The accuracy of diagram structure recognition can be assessed by the complexity of operations necessary to convert the recognized structure into the correct one (ground truth). This complexity can be measured by the summed cost of elementary operations that can be used in order to turn the recognized structure into the correct one. We focus here on the evaluation of the multiset $\widehat{U}$ as defined in (1). Let us assume that the recognized structure described by $\widehat{U}$ is to be converted into the correct structure $\widehat{U}^*$ by applying the sequence of elementary operations. In the result, the sequence of structures is created: $(\widehat{U} = \widehat{U}_1, \widehat{U}_2, ..., \widehat{U}_K = \widehat{U}^*)$ where $\widehat{U}_k$ is converted into $\widehat{U}_{k+1}$ by applying one of the following elementary operations from the set $\mathfrak{O} = \{o_C, o_S, o_M, o_C, o_E, o_R, o_A, o_D\}$ defined as follows:

- $o_C$ - creating a new connector that links two DEs;
- $o_S$ - splitting the multiset $U_i \in \widehat{U}_k$ into two multisets $U_i^1, U_i^2 \in \widehat{U}_{k+1}$ - it corresponds to dividing the compound connector into two simpler ones;
- $o_M$ - merging two multiset $U_i, U_j \in \widehat{U}_k$ into the single multiset $U_l \in \widehat{U}_{k+1}$ - it corresponds to merging two connectors;
- $o_E$ - adding a branch to a diagram element to a connector, i.e. replacing $U_i, \in \widehat{U}_k$ by the new multiset $U_i' = (U_i \cup (e,t)) \in \widehat{U}_{k+1}$;
- $o_R$ - removing a branch to a diagram element from a connector, i.e. replacing $U_i, \in \widehat{U}_k$ by the new multiset $U_i' = (U_i \setminus (e,t)) \in \widehat{U}_{k+1}$;
- $o_A$ - changing the arrow status of the connector endpoint for a certain element descriptor $(e,t)$ in a certain multiset $U_i, \in \widehat{U}_k$.
- $o_D$ - discarding of the whole result of recognition, i.e. replacing of $\widehat{U}_1$ by the empty set.

The last operation is only allowed as the first one in the conversion sequence and can be applied if the recognition result is extremely different from the ground truth diagram structure. Each operation has its cost. The accuracy of the recognized diagram structure can be assessed by the total cost of the least costly operation sequence that converts $\widehat{U}$ into $\widehat{U}^*$. This concept is similar to the edit distance widely used e.g. in automatic speech recognition accuracy evaluation or spelling errors correction ([16]). The value computed in this way is however the absolute measure of labor amount necessary to make a correction to the recognized structure. Certain value of the edit distance may indicate quite good accuracy in the case of very complex diagram, while it may correspond to a poor accuracy in the case where the diagram consists just of few elements and connector edges. Therefore the relative measure related to the actual diagram complexity seems to be more appropriate. The diagram complexity can be measured by the cost of operations necessary to build the set $\widehat{U}^*$ from very beginning, i.e. from the empty family of multisets, using only elementary operations. Let $p_o$ denote the cost of the operation $o$, and $n_o^{(\widehat{U}_1)}$ and $n_o^{(\emptyset)}$ denote the counts of the operation $o$ that must be applied to obtain the correct diagram description $\widehat{U}^*$ from the actual recognition result $\widehat{U}_1$ and from the empty set correspondingly. The total costs of corrections and building from the very beginning are:

$$P_{corr} = \sum_{o \in \mathfrak{O}} p_o * n_o^{(\widehat{U}_1)},$$

$$P_{build} = p_D + \sum_{o \in \mathfrak{O}} p_o * n_o^{(\emptyset)}. \tag{5}$$

and the final recognition accuracy can be computed as:

$$Q = \frac{P_{build} - P_{corr}}{P_{build}} \in < 0, 1 > . \tag{6}$$

The value of $Q$ is normalized into $< 0, 1 >$ interval. If the diagram is perfectly recognized then $P_{corr} = 0$ and $Q = 1$. On the other hand, if the recognition procedure totally misses then the cost of correction is not higher that discarding all recognition results (operation $o_D$) and building the structure from the very beginning. In this case $P_{corr} = P_{build}$ and in result $Q = 0$.

### B. Hardware environment and efficiency issues

The described algorithm was implemented in C++ language. Tests were carried out using the PC equipped with Intel i7 3610QM CPU and 16GB of RAM. The execution time of the algorithm varies depending on the image contents. For simple diagrams consisting of just a few DEs connected by simple connectors the algorithm is executed in less than a single second. The longest execution time (17.6 sec.) was observed in the case of the complex diagram consisting of 58 DEs. The average processing time (including image vectorization) of a single diagram was 2.4 sec. Currently the algorithm implementation is fully sequential.

### C. Results

The accuracy of diagram structure recognition was tested using three types of diagrams that differ in the complexity of intersecting connectors: a) flowcharts, b) organizational charts and c) digital circuit block diagrams. The test set consisted of 11 digital circuit diagrams, 15 flowcharts and 17 organizational charts. Flowcharts seem to have simplest connector structures, while in the case of digital circuits the intersecting connectors appear very often. Hence, the later type of diagrams is the most difficult to recognize. Because in this article we are dealing merely with the problem of connectors recognition, we selected for the tests only such diagrams, where there were no mistakes in diagram elements recognition. For each recognized diagram structure the counts of operations from the set $\mathfrak{O}$ that are necessary to correct the structure were determined as well as the number of operations necessary to build the diagram from the very beginning. We assumed that in the process of construction of the diagram only the operations of new connector creation ($o_C$) and extension of the existing connector with an additional branch ($o_E$) were used. Finally the recognition quality $Q$ was computed for each diagram. We assumed that the unit cost $p_o$ of each operation $o$ is equal to 1.0. The results are presented in Table I. Columns in the left part of the table include numbers of individual correcting operations from the set $\mathfrak{O}$, summed by types of diagrams. The meaning of operation symbols used in Table I were explained in the previous section. The column containing costs of correction/creation contain average costs for individual types of diagrams. The bottom row presents the results analogous to described for individual types of diagrams, but now they are prepared for the whole set containing all types of diagrams.

The small number of diagrams used in the tests do not give rights to create very general conclusions, although the average result for all diagrams equal to 92% is very promising. From the perspective of diagram classes, the worst average result (88% - which seems to be pretty good) was obtained for the class of digital block diagrams. This set contains the most difficult intersecting connections. Our subjective evaluation of the obtained results is very optimistic. Most errors were caused by low image quality and inaccuracies in drawing. It can be observed that there were no errors consisting in detecting neither "false" (i.e. actually not existing) connectors nor branches. The most typical error consisted in omitting connector branches to some DEs. Detailed analysis of missing branches revealed that in most cases errors of this kind were caused by too wide gaps between a diagram element and a terminal endpoint of the connector branch.

Table II presents the results of diagram structure recognition of the exemplary organizational diagram. Fig. II.1 shows the original image. All detected connectors are drawn in Fig. II.2. The connectors are drawn with thick blue lines. Remaining figures show selected individual connectors detected in this diagram. Some simple connectors were omitted and only more complex ones are presented. It is clear that intersecting connectors were properly separated and complex connectors with "branches" were constructed as intended by the diagram author. In the case of this diagram all connectors were recognized correctly.

## VI. Conclusion

The subjective and objective evaluation of the method gives us good perspective for further development of the method,

TABLE I
STRUCTURE RECOGNITION ACCURACY EVALUATION ON THREE TYPES OF DIAGRAMS

| **Diagram type** | **Correction (number of correcting operations)** | | | | | | **Construction (number of constructing opeartions)** | | | **Q** |
|---|---|---|---|---|---|---|---|---|---|---|
| | $o_C$ | $o_S$ | $o_M$ | $o_E$ | $o_R$ | Correction cost | $o_C$ | $o_E$ | Construction cost | |
| Digital circuits | 3 | 0 | 8 | 0 | 19 | 30 | 191 | 22 | 224 | 0.88 |
| Flowcharts | 2 | 0 | 1 | 0 | 10 | 13 | 178 | 12 | 205 | 0.94 |
| Organizational diagrams | 2 | 0 | 1 | 0 | 12 | 15 | 308 | 14 | 339 | 0.96 |
| Total: | 7 | 0 | 10 | 0 | 41 | 58 | 677 | 48 | 768 | 0.92 |

although we realize that the set of diagrams used in order to asses the quality of structure recognition is quite small and conclusions drawn from described experiments may be disturbed by the random factor. In future, significantly bigger set of diagrams will be manually annotated and used for reliable accuracy estimation.

The main weakness of the proposed method seems that it utilizes the set of constants-thresholds, in most cases defining tolerances of various graphical attributes (lengths, angles, widths, distances etc.). It makes possible to correctly recognize some structures that are drawn inaccurately but, on the other hand, it may also lead to recognition errors. In our experiments we tuned these constants intuitively, so as to obtain best recognition results on the validation image set (which is disjoint from the test set). We cannot however assure that these parameter values are selected optimally. In future works, other method of parameter tuning should be applied that will make it possible to find suboptimal parameter values without the engagement of a human. Image processing methods used at the stage of diagram image binarization and vectorization should also be improved, so that low quality images as well as images with background patters or containing "decorated" shapes can be more reliably analyzed.

Some problems encountered in determining connections between diagram elements are caused merely by low quality of scanned images, while others follow from complicated shapes of connectors that intersect each other. In future, an experiment will be carried out which will compare the average accuracy achieved in a set of scanned images with the accuracy achieved in a set of images directly converted to the raster format, e.g. by saving the image created in a graphic diagram editor.

## ACKNOWLEDGMENT

This work was partially supported by the European Commission under the 7th Framework Programme, Coordination and Support Action, Grant Agreement Number 316097, ENGINE - European research centre of Network intelliGence for INnovation Enhancement.

## REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Computing Surveys*, vol. 40, no. 12, pp. 5:1–5:60, 2012. doi: 10.1145/1348246.1348248. [Online]. Available: http://doi.acm.org/10.1145/

[2] D. Blostein, “General diagram-recognition methodologies,” *Graphics Recognition Methods and Applications*, vol. 1072, pp. 106–122, 1996. doi: 10.1007/3-540-61226-2-10. [Online]. Available: http://dx.doi.org/10.1007/3-540-61226-2_10

[3] Y. Liu, X. Lu, Y. Qin, Z. Tang, and J. Xu, “Review of chart recognition in document images,” in *Proc. SPIE 8654, Visualization and Data Analysis*, vol. 865410, 2013. doi: 10.1117/12.2008467. [Online]. Available: http://dx.doi.org/10.1117/12.2008467

[4] A. Lemaitre, H. Mouch‘ere, J. Camillerapp, and B. Couasnon, “Interest of syntactic knowledge for on-line cognitionr,” in *Proc. of ninth IAPR International Workshop on Graphics Recognition (GREG2011)*, 2011. doi: 10.1007/978-3-642-36824-0 9 pp. 85–98.

[5] M. Bresler, D. Prusa, and V. Hlavac, “Modeling flowchart structure recognition as a max-sum problem,” in *ICDAR, IEEE Computer Society*, 2013. doi: 10.1109/ICDAR.2013.246 pp. 1215–1219. [Online]. Available: http://dblp.uni-trier.de/db/conf/icdar/icdar2013.html/BreslerPH13

[6] G. Feng, C. Viard-Gaudin, and Z. Sun, “On-line hand-drawn electric circuit diagram recognition using 2d dynamic programming,” in *Pattern Recognition*, vol. 42, 2009. doi: 10.1016/j.patcog.2009.01.031 pp. 3215–3223. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00419076

[7] Y. Qi, M. Szummer, and T. P. Minka, “Diagram structure recognition by bayesian conditional random fields,” in *CVPR 2005. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 191 – 196.

[8] *Evaluating Flowchart Recognition for Patent Retrieval*, 2013. doi: 10.1007/s10791-013-9234-3. [Online]. Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/EVIA/08-EVIA2013-LupuM.pdf

[9] A. Hanbury, N. Bhatti, M. Lupu, and R. Mörzinger, “Patent image retrieval: a survey,” in *Proceedings of the 4th worshop on Patent information retrieval*, 2011. doi: 10.1145/2064975.2064979 pp. 494–497.

[10] L. Yan, W. Huang, and C. L. Tan, “Semi-automatic ground truth generation for chart image recognition,” in *Workshop on Document Analysis Systems (DAS)*, 2006. doi: 10.1016/j.patrec.2015.02.001 pp. 324–335.

[11] M. Awadalla and A. Sadek, “Spiking neural network-based control chart pattern recognition,” *Journal of Engineering and Technology Research*, vol. 3, no. 1, pp. 5–15, 2011. doi: 10.1007/s10845-012-0659-0. [Online]. Available: http://www.academicjournals.org/journal/JETR/article-abstract/59E571210699

[12] Y. Yu, A. Samal, and S. C. Seth, “A system for recognizing a large class of engineering drawings,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997. doi: 10.1109/34.608290

[13] J. Sas and A. Zolnierek, “Three-stage method of text region extraction from diagram raster images,” in *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013, Milkow, Poland, 27-29 May 2013*, 2013. doi: 10.1007/978-3-319-00969-8-52 pp. 527–538. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-00969-8_52

[14] A. N. Kolesnikov, V. V. Belekhov, and I. O. Chalenko, “Vectorization of raster images,” *Pattern Recognition and Image Analysis*, vol. 6, no. 4, pp. 786–194, 1995. [Online]. Available: http://cs.joensuu.fi/~koles/dissertation/Kolesnikov_Paper1.pdf

[15] A. Fitzgibbon, M. Pilu, and R. B. Fisher, “Direct least square fitting of ellipses,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 476–480, May 1999. doi: 10.1109/34.765658. [Online]. Available: http://dx.doi.org/10.1109/34.765658

[16] R. A. Wagner and M. J. Fischer, “The string-to-string correction problem,” *J. ACM*, vol. 21, no. 1, pp. 168–173, Jan. 1974. doi: 10.1145/321796.321811. [Online]. Available: http://doi.acm.org/10.1145/321796.321811

TABLE II
EXAMPLE OF RECOGNIZED DIAGRAM STRUCTURE

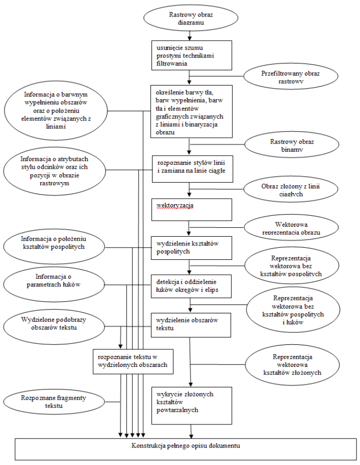

Rastrowy obraz diagramu
usunięcie szumu prostymi technikami filtrowania
Przefiltrowany obraz rastrowy
Informacja o barwnym wypełnieniu obszarów oraz o położeniu elementów związanych z liniami
określenie barwy tła, barw wypełnienia, barw tła i elementów graficznych związanych z liniami i binaryzacja obrazu
Rastrowy obraz binarny
Informacja o atrybutach stylu odcinków oraz ich pozycji w obrazie rastrowym
rozpoznanie stylów linii i zamiana na linie ciągłe
Obraz złożony z linii ciągłych
wektoryzacja
Wektorowa reprezentacja obrazu
Informacja o położeniu kształtów pospolitych
wydzielenie kształtów pospolitych
Reprezentacja wektorowa bez kształtów pospolitych
Informacja o parametrach łuków
detekcja i oddzielenie łuków okręgów i elips
Reprezentacja wektorowa bez kształtów pospolitych i łuków
Wydzielone podobrazy obszarów tekstu
wydzielenie obszarów tekstu
rozpoznanie tekstu w wydzielonych obszarach
Reprezentacja wektorowa kształtów złożonych
Rozpoznane fragmenty tekstu
wykrycie złożonych kształtów powtarzalnych
Konstrukcja pełnego opisu dokumentu

Fig.II.1. Original diagram

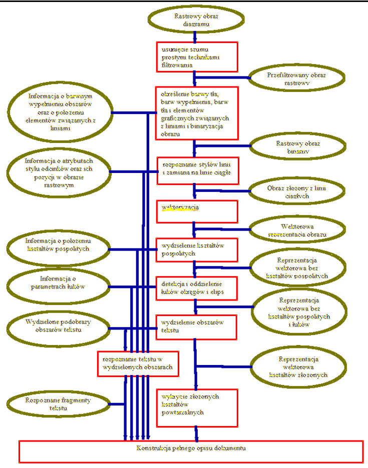


Fig.II.2. All detected connectors

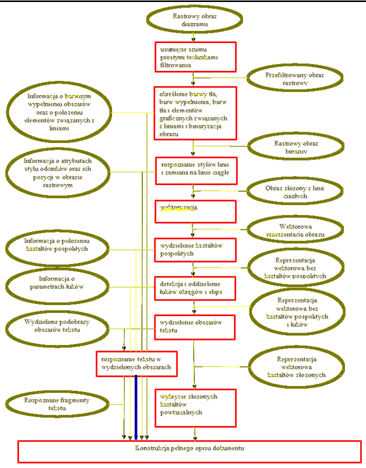


Fig.II.3. Example of simple connector

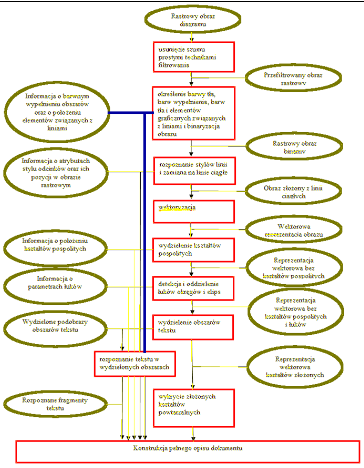


Fig.II.4. Example of compound connector

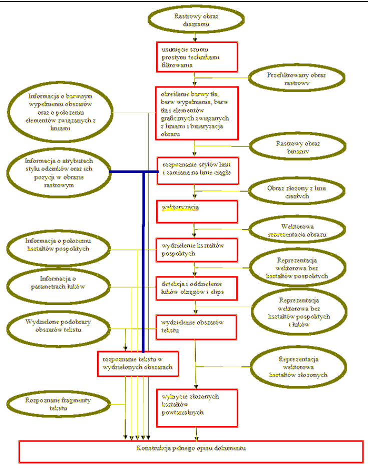


Fig.II.5. Example of compound connector

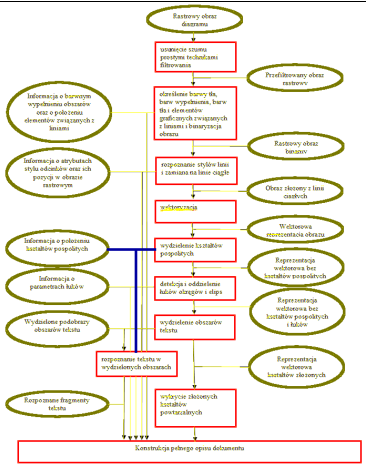


Fig.II.6. Example of compound connector

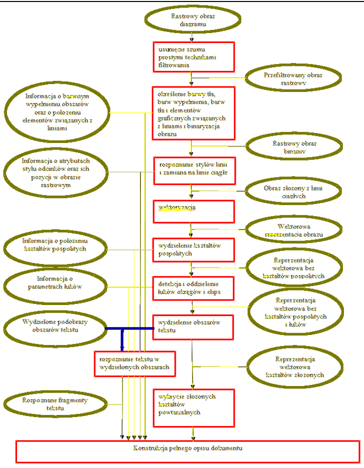


Fig.II.7. Example of compound connector

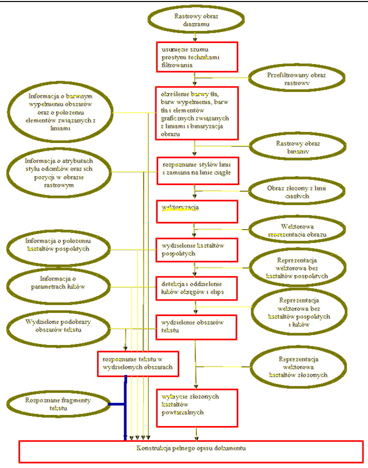


Fig.II.8. Example of compound connector

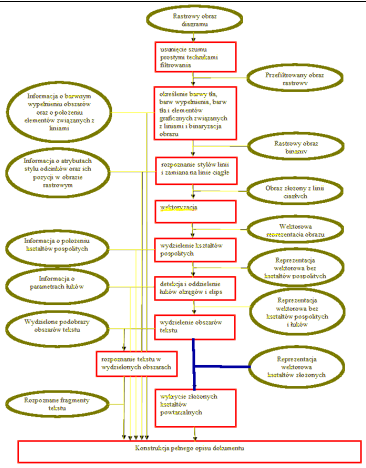


Fig.II.9. Example of compound connector

DOI: 10.15439/2015F307
ACSIS, Vol. 5

# Liveness detection in remote biometrics based on gaze direction estimation

Krzysztof Adamiak, Dominik Żurek, Krzysztof Ślot
Lodz University of Technology
ul. Stefanowskiego 18/22, 90-924 Łódź, Poland
Email: kslot@p.lodz.pl

***Abstract*—The following paper presents a simple and fast liveness detection method based on gaze direction estimation under a challenge-response user authentication scenario. To estimate a line of sight, a procedure composed of several steps, including face and eye detection, derivation of gaze direction representation and subsequent classification, has been proposed. The proposed, novel gaze orientation descriptor is easy to compute and it provides sufficiently accurate estimates for the considered task. To assess a probability of genuine biometric trait presentation, recorded gaze direction responses induced by presentation of a randomly generated on-screen object, are matched against expected patterns.**

## I. Introduction

ONE OF the main threats that exist for unattended biometric authentication systems are so called 'presentation attacks', where a system is presented with a biometric artefact. The problem becomes especially severe when easy-to-spoof biometric traits are considered, such as e.g. fingerprints, face or iris images. To enable unattended (including remote, by means of popular mobile devices) user verification, biometric systems must cope with the stated problem. For this purpose, a methodology aimed at verification of biometric trait authenticity, referred to as liveness detection, has been developed.

Existing liveness detection approaches can be broadly categorized into two main groups: methods that exploit physiological properties of tissues and organs subject to analysis and various challenge-response schemes. Liveness detection is therefore trait-specific and many diverse ways for its assessment have been proposed so far. For example, several different tests are available for iris image authenticity verification, such as application of varying intensity illumination levels to check for iris physiological responses (this approach combines challenge-response scheme and physiology), analyzing presence of saccades or analyzing spectra of reflected light. Approaches considered for liveness testing in case of face-based recognition include blink detection [1], detection of eye-movements [2] detection of presence of facial expressions [3] or lip motion detection [4]. A natural means against presentation of photos provides 3D face recognition (however, it clearly becomes vulnerable if 3D masks are used). Fingerprint validity can be assessed by analyzing perspiration processes. For some biometric traits, such as vein pattern structure, liveness detection is an intrinsic component of the adopted recognition methodology, as vein imaging (blood-flow imaging) is possible only for living tissues.

Widespread availability of mobile devices equipped with cameras and microphones brought up an interest in exploiting face images and voice in remote biometric authentication, despite their well-known limitations. One of the main problems that needs to be addressed in this scenario is clearly liveness detection. One of the most natural liveness detection schemes used in case of speaker recognition is a challenge-response procedure, where a speaker is prompted to utter some randomly generated text, and only after positive verification of the response, a biometric system proceeds with user-verification procedure. An application of a similar challenge-response scheme in video based face analysis have been proposed in [5], where a challenge requires a user to make voluntary blinks and mouth movements (opening and closing). Another interesting example of challenge-response scheme that utilizes gaze tracking and that is intended as a secure method for logging to computer systems has been proposed in [6]. A set of icons, which includes a randomly scattered subset of previously memorized ones, is displayed to a user, who is supposed to trace (with her/his eyes) a path, that defines a convex hull built upon known icons.

The presented paper proposes a simple liveness detection method that is based on verification of line-of-sight trajectory compliance (a response) with some expected pattern (a challenge). The pattern is defined by subsequent locations of a marker that gets displayed at random locations of a screen. The main element of the proposed method is a novel gaze direction estimation algorithm, which is computationally inexpensive, enabling its real-time application even on machines with limited computing power, such as mobile devices. Simplicity of the proposed gaze detection method results from its specific context: horizontal line-of-sight displacement evaluation is sufficient for execution of liveness detection procedure.

The proposed gaze detection method conforms to a general framework of the domain and includes two phases: derivation of eye-image representation that correlates well with gaze direction, followed by gaze direction evaluation. As widely available devices are considered for the method implementation, gaze tracking is performed using regular visible light cameras (the best performing gaze tracking

This work was supported by the Polish National Science Center under the research grant 2012/05/B/ST6/03647

978-83-60810-66-8/$25.00©2015, IEEE

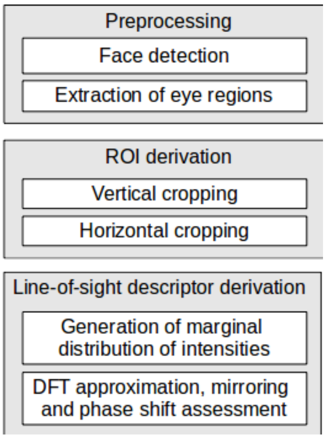

Preprocessing
Face detection
Extraction of eye regions
ROI derivation
Vertical cropping
Horizontal cropping
Line-of-sight descriptor derivation
Generation of marginal distribution of intensities
DFT approximation, mirroring and phase shift assessment

Fig. 1. Block diagram of gaze estimation procedure

methods exploit near infrared range and typically use infrared light sources [7]).

Several different eye image representations have been proposed for the purpose of gaze tracking applications. They belong to two broad categories: representations based on mutual locations of salient geometric eye features, such as inner and outer eye corners, iris/pupil centers, and appearance based representations (a broad review of relevant methods can be found e.g. in [8]). Gaze direction assessment exploiting features from the former group is basically a regression problem that can be solved using e.g. Support Vector Regression (SVR) [9] or neural networks [10]. For appearance-based descriptors, gaze direction assessment is made e.g. by computing between-region correlation coefficients [11] or using mean-shift algorithm [12]. The proposed algorithm for line-of-sight direction estimation is based both on salient feature detection and appearance-based eye modeling, and it is followed by regression based analysis, so it combines both of the presented general methodologies.

A structure of the presented paper is the following. The proposed gaze detection algorithm has been introduced in Section 2. Section 3 presents a background used for challenge-response liveness detection. Finally, experimental evaluation of the proposed method has been presented in Section 4.

## II. Gaze Estimation Algorithm

Block diagram of the proposed algorithm has been depicted in Fig. 1. A general idea of the proposed line-of-sight direction estimation is to confront information extracted from two eye images, where one of them is mirrored, so that deviations from a central fixation point (determined during a calibration phase of the procedure) get amplified.

The procedure begins with face detection, followed by eye-region detection, both performed using a well-known Viola-Jones algorithm [13]. The detected regions (containing the left and the right eye) are further refined by vertical and horizontal cropping, performed to increase processing speed and to facilitate subsequent analyses by eliminating complex yet unrelated structures, such as eyebrows. The resulting regions of interest (ROI) become a domain for gaze direction assessment, which uses a quantitative descriptor that estimates eye disks horizontal offset from their ’neutral’ position (forward gaze). The descriptor is derived from Fourier spectra of marginal distributions of vertical projections of eye image intensities and gaze direction is represented by a value of a phase shift between fundamental frequencies that approximate the considered functions. The presented approach is detailed in the following subsections.

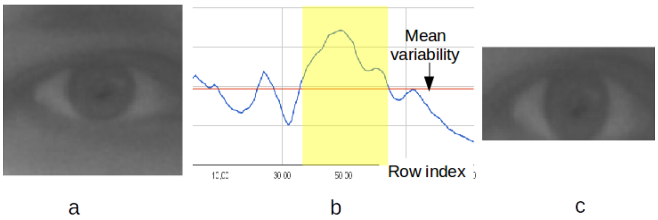


Fig. 2. Determination of vertical bounds for a rectangular image analysis window: an initial eye region produced by Viola-Jones algorithm (a), sample plot of horizontal intensity variance with the selected row interval (b) and the resulting analysis window (c)

### A. ROI Derivation

An objective of vertical cropping of initial eye regions is elimination of irrelevant upper and lower image structures, such as e.g. eyebrows. To adjust initial eye regions vertically, we propose to analyze horizontal image variability. As eye images always contains regions of extreme intensities (white cornea versus black pupil), variability of image intensities along these rows is expected to dominate over the remaining regions. A criterion for selection of vertical bounds of a rectangular window that will be used for gaze analysis is based on analysis of horizontal variance projection function. We propose to extract from an initial eye region only the widest strip, composed of rows with gray-level variability above the average level, computed for the whole region (see Fig. 2).

An objective of horizontal cropping is to produce a normalized image analysis domain, where two landmarks: inner and outer eye-corners determine a system of reference. A use of eye corners as landmarks has important advantages (they are distinctive and separated by a fixed distance from each other, which offers a basis for pose estimation).

The proposed eye-corner detection method operates on vertically-cropped image eye-regions. To compensate for illumination variations, prior to further processing, eye region images are normalized in intensity. The algorithm begins with corner detection procedure, which seeks for image points with large contents variability. Two well-known methods that differ in the adopted decision criterion were examined to do the task: Harris [14] and Shi-Tomasi [15] detectors. Both methods were able to correctly detect all salient image points, including eye corners (see Fig. 3). However, as the Shi-Tomasi method favors features that are easier to track (this is an important aspect from the standpoint of computational efficiency of the algorithm), it has been selected in further analyses.

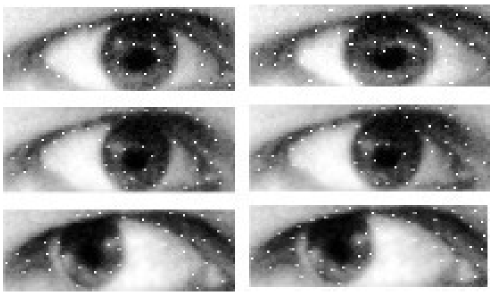


Fig. 3. Corner detection results in input images produced using Harris detector (left column) and Shi-Tomasi detector (right column).

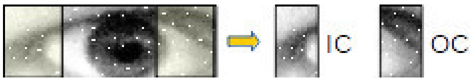


Fig. 4. Eye corner search domains (IC - inner corner search region, OC - outer corner search region).

As it can be seen from Fig. 3, corner detection procedure results in identification of a large set of salient points. To find eye corners, elements of this set are subject to subsequent analysis. Firstly, its domain gets restricted, so that inner and outer eye corners are sought only in feasible eye subregions, defined as boundary vertical bands of width set to 25% of eye image (Fig. 4).

To identify eye corners among a set of available salient points, a descriptor that summarizes image appearance within a square, 15×15 neighborhood around a salient point, has been generated. The neighborhood is divided into quadrants of size $r \times r$, $r$ =8 (with overlapping boundaries), and mean image intensity gradients are evaluated within each quadrant. A descriptor of a salient point $\mathbf{P}^{i,j}$, located in $i-$th row and $j-$th column of an image, is thus a collection of four vectors:

$$\mathbf{S}^{i,j} = \left[\mathbf{g}_{\mathrm{TL}}^{ij}, \mathbf{g}_{\mathrm{TR}}^{ij}, \mathbf{g}_{\mathrm{BL}}^{ij}, \mathbf{g}_{\mathrm{BR}}^{ij}\right] \quad (1)$$

where subscripts T,B,L,R label quadrants (top, bottom, left and right) and the mean gradient of pixel intensities $I(k,l)$ in a quadrant XY ($\mathbf{g}_{\mathrm{XY}}^{ij}$) is given by:

$$\mathbf{g}_{\mathrm{XY}}^{ij} = \frac{1}{r^2} \sum_{k,l \in \mathrm{XY}} \nabla I(k,l) \quad (2)$$

Descriptor gradient vectors were finally normalized in length, so that they sum up to unity. Salient points are matched against eye-corner models. Four separate eye-corner models (inner and outer for the left and for the right eye), of the same structure as given by (1), were generated from a set of manually labeled images (five images per each corner - see Fig. 5). Descriptors derived for training images for particular eye corners were averaged, forming the corresponding eye-corner models $\mathbf{C}^{\mathrm{IL}}, \mathbf{C}^{\mathrm{OL}}, \mathbf{C}^{\mathrm{IR}}, \mathbf{C}^{\mathrm{OR}}$, where I,O,L,R denote inner, outer, left eye and right eye respectively. As it was the case for salient point descriptors, also eye-corner model gradients were analogously normalized in length. The derived models were matched against salient points that are present in the corresponding eye-image bands (i.e. inner left eye-corner model was applied to salient points present in the 'IC' region of the left eye image etc.). A matching score was defined as a sum of dot products between components of a corner's $\alpha, \beta$ model descriptor and some considered salient point's descriptor:

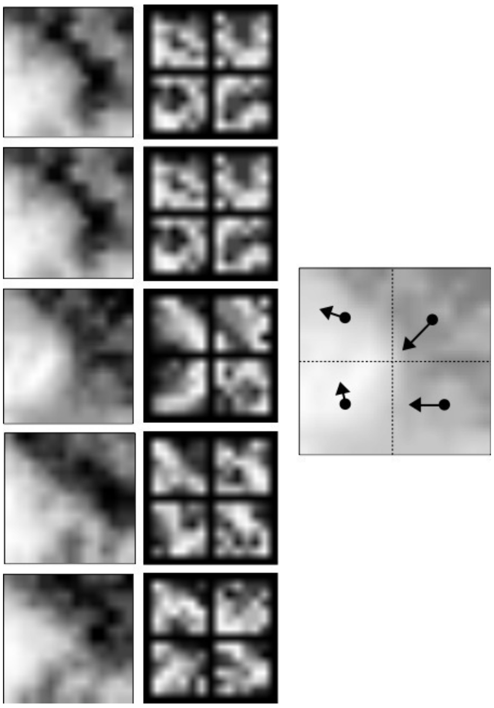


Fig. 5. Generation of eye-corner model: manually selected instances of inner left eye corner (left column), gradient magnitudes derived for quadrants (middle) and the resulting gradients (right).

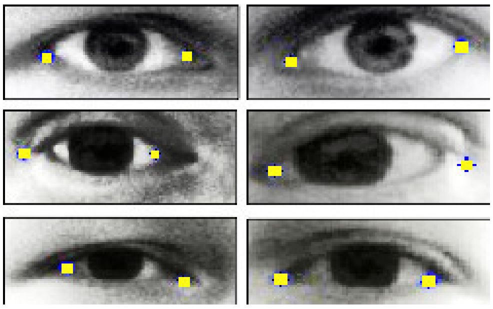


Fig. 6. Sample eye corner detection results performed for three different persons using the presented algorithm.

$$F_{\alpha,\beta}^{ij} = \sum_{k=0}^{3} \langle \mathbf{S}^{i,j}[k], \mathbf{C}^{\alpha,\beta}[k] \rangle \quad (3)$$

The score (3) gets maximized for salient point neighborhoods that match a particular eye-corner model. Sample results of eye-corner detection have been presented in Fig. 6.

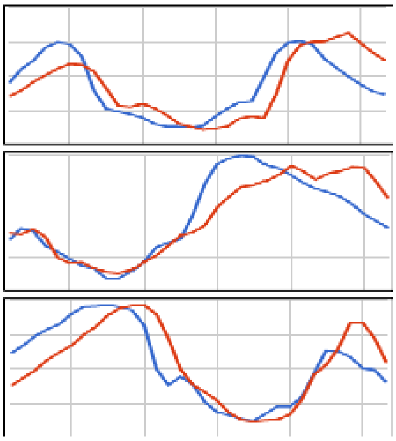


Fig. 7. Plots of image intensity distribution accumulated along a vertical axis for three different gaze directions: forward (top), left (middle) and right (bottom), for left (red) and right (blue) eyes.

### B. Line-of-sight Direction Assessment

Regions of interest, derived for both eyes, provide domains that comprise information relevant for gaze direction assessment. To provide fast and accurate estimation of gaze direction, an appropriate descriptor that can be easily computed and that is robust against possible image artifacts, needs to be derived.

To meet the formulated goals, we decided to generate gaze direction descriptor based on differences in general appearance of left and right eye strips in horizontal direction. The appearance of a strip can be summarized using marginal distribution of strip pixel intensities accumulated in vertical direction. One can observe that a general shape of the resulting function is line-of-sight direction specific (see Fig. 7). Such a general shape can be easily quantified by using leading components of any of possible signal orthogonal transformations, such as e.g. Discrete Fourier Transform (DFT). DFT has been chosen as a basis for eye appearance representation and spectra of marginal distributions of vertically accumulated intensities, derived for different images (see Fig. 8), were analyzed. We found that the first periodic component appears to be an attractive means for summarizing eye appearance, as its phase shows good correlation with a gaze direction. Examples of approximation of marginal gray-level distributions by means of a fundamental component of its DFT decomposition have been shown in Fig. 9.

To amplify sensitivity of the representation, we decided to confront the approximation produced for marginal distribution derived for one of the eyes with an approximation produced for the mirrored marginal distribution of the other eye (see Fig. 9). Thus, a final descriptor of gaze direction is a phase difference between two first harmonics, where the first one approximates the marginal intensity distribution derived for the left eye and the second one approximates a mirrored distribution derived for the right eye.

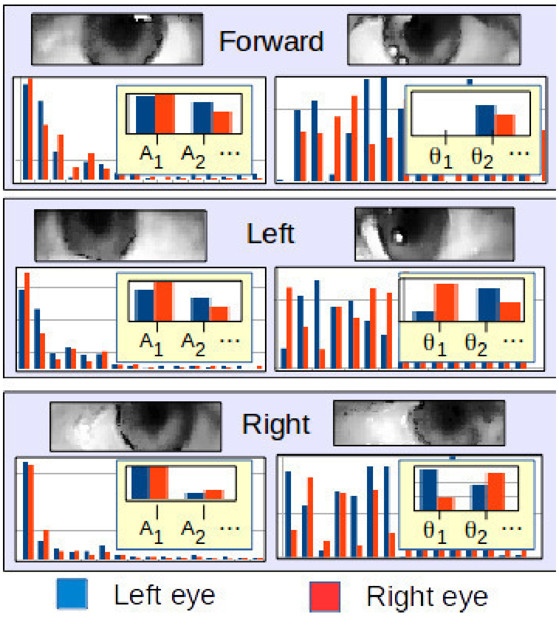


Fig. 8. Magnitude (left column) and phase (right column) spectra derived for marginal distributions of vertical image intensities for three gaze directions: forward (top), left (middle) and right (bottom) (magnitudes and phases of the two first periodic components have been enlarged for illustration clarity).

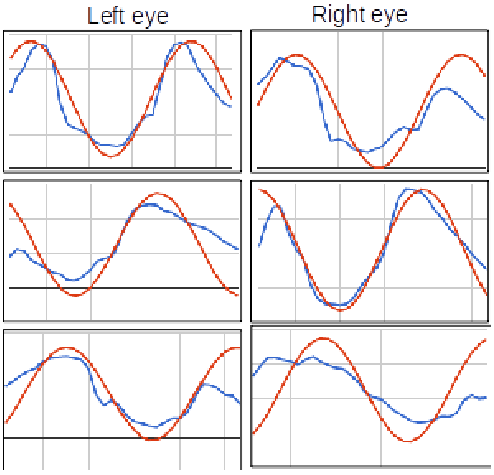


Fig. 9. Marginal distribution approximations using the first DFT harmonic component for three different line-of-sight directions: forward (top), left (middle) and right (bottom).

## III. Liveness Detection Procedure

Challenge-response scheme has been used as a framework for liveness detection, where the challenge is an on-screen presentation of a marker (a circle) in time-varying locations and the expected response is a corresponding line-of-sight direction adjustment. Objects are presented at random locations and the system is attempting to determine the induced gaze direction, thus verifying a required reaction. If probability of correct gaze detection exceeds 0.5, one can expect that successive repetitions of the procedure will eventually provide a required level of confidence that a user actually responds to a challenge. To derive a quantitative estimate of number of trials that are required for getting some predefined confidence level, we assume that the considered liveness detection scheme can be expressed in terms of a Bernoulli process. The main error in gaze direction estimation results from erroneous detection of eye corners, that can happen equally likely for any viewing angle, we assume that probabilities of correct gaze detection are the same, regardless of this angle. This justifies adoption of the Bernoulli scheme and allows to define the two outcomes required by the process in the following manner. A success occurs if a response (estimated line of sight) falls within an expected angular interval around actual position of a marker (we assume that an angular range of on-screen locations is evenly split into an even number of intervals). Otherwise, we consider that an outcome of an experiment is a failure. Given this framework, and given success and failure probabilities, we can estimate a number of challenge repetitions that is required to meet some predefined confidence level (or equivalently, liveness detection error probability). It can be shown [16] that probability of correct classification of at least $\frac{n+1}{2}$ elements of $n$-element sequence $Q$, given probability of correct entry classification $p > 0.5$ :

$$p(Q,n) = \sum_{i=1}^{(n+1)/2} \binom{n}{\frac{n-1}{2}+i} p^{\frac{n-1}{2}+i}(1-p)^{\frac{n+1}{2}-i} \qquad (4)$$

converges to unity as $n \to \infty$. Therefore, it is always possible to find such $n$ that provides some desired confidence level $T$ of an affirmative decision.

## IV. Experimental Evaluation of the Procedure

The experimental setup used for proposed algorithm evaluation was the following. An application for generating a challenge draws a single marker at time-varying, seventeen equidistant discrete locations (a marker is a circle of a fixed size). Locations of the marker change every 2 seconds. During gaze direction detection accuracy tests marker locations were periodically updated in an oscillating manner. During liveness detection tests, marker locations were randomly selected. A user was situated in front of a screen at a fixed distance, so that marker is observed within a range of angles starting from -30 degrees to +30 degrees. A simple web camera, positioned centrally atop a screen was monitoring user's responses (challenge generation and image acquisition processes were synchronized). Each recorded frame was subject to a separate analysis (no object tracking mode was used, to provide more data for evaluation of all procedure steps).

Two databases were used throughout the experiments. The first one was prepared by the Authors and comprises 415 low resolution (640 × 480) test images of three subjects with manually labeled four eye corners and with labeled line-of-sight orientations. The second source of experimental material was a CAVE database [17]. 1176 high resolution (5,184 × 3,456 pixels) images of 56 different persons, with known gaze directions, were used.

TABLE I
Eye corner detection accuracy evaluation (column labels identify a corner: OL - outer left eye corner, IL - inner left eye corner, IR - inner right eye corner, OR - outer right eye corner).

| | All | OL | IL | IR | OR |
|---|---|---|---|---|---|
| Detected | 96% | 99% | 100% | 96% | 96% |
| Identified | 79.4% | 86% | 94% | 93% | 84% |

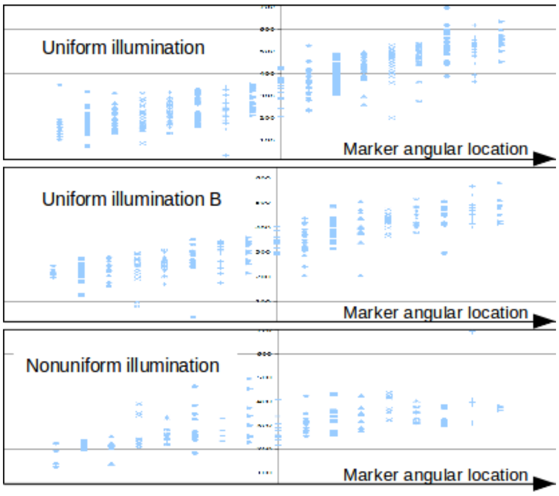


Fig. 10. Plots of estimated gaze directions with respect to actual ones for three different sets of experimental data.

The first phase of the evaluation was aimed at estimating accuracy of ROI derivation procedure. Results of eye-corner detection have been summarized in Table I. The first row shows performance of Shi-Tomasi algorithm - eye corner detection has been considered successful if any of produced salient points was sufficiently close to the considered landmark (within its 5x5 neighborhood). The second row of the table specifies performance of correct salient point identification. The first column of Table I indicates percentage of correct detection for all corners, whereas the remaining ones show scores for individual eye-corners.

The second set of experiments was concerned with evaluation of gaze detection accuracy. A set of three video sequences of a user asked to eye-track the oscillating marker were recorded and a functional relation between actual marker angular positions and positions calculated using the presented algorithm was derived. The sequences were differing in adopted illumination conditions: the first two sequences were taken under uniform illumination of different intensity, whereas in the third case a face was lit from aside. For every sequence a total of 510 frames were analyzed (an average of 30 frames per marker location). The results are summarized in Fig. 10, where plots show the computed marker locations against their actual locations. As it can be seen, there exist significant variations in gaze direction estimation, however one can identify angular intervals that can be exploited for liveness detection.

Given gaze-estimation results, it has been assumed that a marker will be displayed randomly at three different locations of a screen: two extreme positions and in the middle. An average probability of successful detection of gaze direction was evaluated to be $p \approx 0.66$. It follows from eqn. (4) that a number of required presentations of a marker necessary for obtaining a $T = 95\%$ level of confidence that a subject is actually following the marker equals 23.

The last part of experiments was concerned with evaluation of computational complexity of the proposed algorithm. The method was implemented in C++ programming language and executed on a desktop computer with i7, quad core processor, running at 2.4 GHz. The presented algorithm, excluding an initial phase of face detection and preliminary eye-region detection (both performed using Viola-Jones algorithm) took on average only 3 milliseconds to execute (the result was averaged for processing of 721 images). Although the aforementioned, initial preprocessing can be time consuming, one needs to note that it can be significantly accelerated if face tracking mode will be used for analysis of frames that follow the first one.

## V. Conclusion

The proposed algorithm proves that liveness detection can be performed using line-of-sight estimation, by using a simple camera for image acquisition. Computational complexity of the procedure is low and we believe that it can be implemented on popular mobile device platforms. There exist several elements of the procedure that need to be explored to increase gaze-direction assessment accuracy, which is important to reduce a required duration of liveness detection procedure. The main directions of further exploration will be concerned with improving eye-corner identification performance (e.g. by applying multi-resolution analysis) and with modifying the adopted eye-image representation (e.g. by including more components of DFT decomposition of eye images).

## Acknowledgment

The Authors wish to thank Dr Piotr Duch for his invaluable help in formulating the presented concepts.

## References

[1] L. Sun, G. Pan, Z. Wu, S. Lao, "Blinking-Based Live Face Detection Using Conditional Random Fields," *in Int. Conf. on Biometrics*, pp. 252–260, 2007. http://dx.doi.org/10.1007/978-3-540-74549-5-27

[2] Komogortsev, O. V.; Karpov, A.; Holland, C. D., "Attack of Mechanical Replicas: Liveness Detection With Eye Movements," *IEEE Transactions on Information Forensics and Security*, vol.10, no.4, pp. 716–725, 2015. http://dx.doi.org/10.1109/TIFS.2015.2405345

[3] J. Li, Y. Wang, T. Tan, A.K. Jain, "Live face detection based on the analysis of Fourier spectra," *in SPIE Conf. on Biometric Technology for Human Identification*, vol. 5404, pp. 296–303, 2004. http://dx.doi.org/10.1117/12.541955

[4] K. Kollreider, H. Fronthaler, M. I. Faraj, J. Bigun, "Real-time face detection and motion analysis with application in âĂIJlivenessâĂİ assessment," *IEEE Transactions on Information Forensics and Security*, vol. 2 pp. 548–558, 2007. http://dx.doi.org/10.1109/TIFS.2007.902037

[5] Singh, A. K.; Joshi, P.; Nandi, G. C., "Face recognition with liveness detection using eye and mouth movement," *in Int. Conf. on Signal Propagation and Computer Technology (ICSPCT)*, pp. 592–597, 2014. http://dx.doi.org/10.1109/ICSPCT.2014.6884911

[6] D. Weinshall, "Cognitive Authentication Schemes Safe Against Spyware," *in Proc. of IEEE Symposium on Security and Privacy*, pp. 300–306, 2006. http://dx.doi.org/10.1109/SP.2006.10

[7] Ying Qi; Zhi-Liang Wang; Zhang Chuang, "A non-contact eye-gaze tracking system for human computer interaction," *in Int. Conf. Wavelet Analysis and Pattern Recognition*, vol.1, pp. 68–72, 2007. http://dx.doi.org/10.1109/ICWAPR.2007.4420638

[8] D.W. Hansen, Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.32(3), pp. 478–500, 2010. http://dx.doi.org/10.1109/TPAMI.2009.30

[9] Zhiwei Zhu; Qiang Ji; Bennett, K. P., "Nonlinear Eye Gaze Mapping Function Estimation via Support Vector Regression," *in Proc. of Int. Conf. on Pattern Recognition*, vol.1, pp. 1132–1135, 2006. http://dx.doi.org/10.1109/ICPR.2006.864

[10] B.L. Nguyen, C. Tijus, F. Jouen, M. Molina, Y. Chahir, "Eye gaze tracking with free head movements using a single camera," *in Proc. of the 2010 Symposium on Information and Communication Technology*, pp. 108-113, 2010. http://dx.doi.org/10.1145/1852611.1852632

[11] M. Betke, J. Gips, P. Fleming, "The camera mouse: visual tracking of body features to provide computer access for people with severe disabilities", IEEE Neural Systems and Rehabilitation Engineering 10(1), pp. 1-10 (2002). http://dx.doi.org/10.1109/TNSRE.2002.1021581

[12] T. Liu, C. Pang, "Eye-gaze Tracking Research Based on Image Processing", Congress on Image and Signal Processing, IEEE, pp. 176-180 (2008). http://dx.doi.org/10.1109/CISP.2008.590

[13] P. Viola and M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," *in IEEE Conference on Computer Vision and Pattern Recognition*, vol.1, pp. 511–518 2001. http://dx.doi.org/10.1109/CVPR.2001.990517

[14] C. Harris , M. Stephens, "A combined corner and edge detector," *in Proc. of Fourth Alvey Vision Conference*, pp. 147–151, 1988. http://dx.doi.org/10.5244/C.2.23

[15] Shi, C. Tomasi, "Good Features to Track," *in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 593–600, 1994. http://dx.doi.org/10.1109/CVPR.1994.323794

[16] L. K. Hansen and P. Salamon, "Neural Networks Ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990. http://dx.doi.org/10.1109/34.58871

[17] B. A. Smith, Q. Yin, S. K. Feiner and S.K. Nayar,"Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction", *ACM Symposium on User Interface Software and Technology (UIST)*, pp. 271–280,2013.

DOI: 10.15439/2015F410
ACSIS, Vol. 5

# Fast GPU and CPU computing for Head Position Estimation

Michał Szkudlarek[1] and Maria Pietruszka[2]
Institute of Information Technology
Lodz University of Technology
ul. Wólczańska 215, 90-924 Łódź, Poland
[1] michal.szkudlarek@dokt.p.lodz.pl, [2] maria.pietruszka@p.lodz.pl

***Abstract*—The head movement based control methods in the 3D graphic applications requires the real-time face position estimation. Therefore, the tracking method at the high speed and with the minimal latency is needed. This is especially hard to achieve when the face is tracked with the use of the high resolution video image on mobile devices. In the article, we present several methods for an acceleration of the face position estimation method based on the fuzzy skin color classifier and other color-based face tracking methods. The acceleration is achieved through a highly parallel GPU computation, the precalculation of the classifier weights and through the combined computations on the GPU and the CPU. The achieved computation time is independent of the used skin color classification method, allowing for use of very complex classifiers. The presented methods provides the robust head position tracking on the high resolution video image of 1920x1080 pixels, at 300 frames per second, on the mobile device with a low computing power.**

## I. Introduction

THE head position tracking can be used as a multiplatform control method on desktop computers, laptops, game consoles and hand-held mobile devices. In the latter case it is especially important as the available touch-based control methods are not suitable for many interactive applications, due to the low precision and the fingers obscuring the screen. Moreover, when we are using the hand-held mobile device, the change of a relative head position does not require the actual movement of the head and can be changed by the rotation of the device.

The face tracking can be used, inter alia, for three-dimensional imaging technique called the *Head-Coupled Perspective* (*HCP* abbreviated) [1][2][3][9] which gives the impression of the depth of the presented 3D scene by dynamically linking the perspective of the rendered scene with the current head position. The rough head tracking in three degrees of freedom is sufficient for this imaging technique. For the depth sensation more important than the face tracking accuracy are its smoothness (i.e. the lack of the jittering and the unnoticeable difference between two consecutive estimated head positions), a high frequency of tracking and its low latency (i.e. short time between the actual head movement and the application response). These requirements are difficult to achieve in the case of the face tracking on the high resolution video image. Also it must be taken into account, that the head tracking is just one of the costly calculations that the interactive application needs to perform in a real time. On the mobile devices the additional limitation is their low computing power.

On the mobile devices it is possible to estimate the relative position of the user's face on the basis of other internal device sensors like the accelerometer or the gyroscope [6][7]. The rotation and the orientation of the device can be determined based on the readings from these sensors and therefore the head position relative to the screen can be estimated. The information from these sensors is delivered quickly (with frequency of 200 Hz on Android devices) and do not require time consuming computations [19]. Moreover, the tracking range of the sensors is not limited by the camera field of view. The drawback of the sensor based tracking is that their readings are noisy, what is especially noticeable in the *HCP* technique where a slight head movement results in the big displacement of the far background of the scene. Filtering of the sensor signals takes into account earlier sensor readings and generates the additional delay in the application response to the head movement. Yet, it may be unnoticeable with use of the Extended Kalman Filter [19]. Still, sensors like accelerometer or gyroscope are not able to notice the actual movement of the head itself, omission of which can destroy the 3D impression. Therefore, the additional use of the video based head tracking can highly improve the quality of the sensor based tracking [6][7]. Unfortunately, the face tracking methods chosen in these articles, are too slow. The tracking is several times slower than the sensors input, so its usage is limited to the occasional sensor reading correction. The use of a very rapid head tracking method may provide the actual head movement tracking, as well as the signal noise reduction without the delay generation.

The article presents the face tracking method that is able to find face coordinates in a few milliseconds and is based on the fuzzy skin color classifier first proposed in [1]. The low computational time is achieved through the parallel calculations on the GPU and the CPU and through the precalculation of all possible color weights assigned by the fuzzy classifier.

In the next section the existing rapid head tracking method are discussed. The third section presents the face position

978-83-60810-66-8/$25.00©2015, IEEE

estimation method with use of the fuzzy skin color segmentation. In the fourth section we describe precisely the methods for accelerating the color-based head position tracking. The fifth section presents the test results of the proposed solutions and the sixth section contains the conclusions and considered future work.

## II. Related Works

Fast head position finding in the camera image is possible with various head tracking and detection methods. They can be based, among others, on the background subtraction [1][2][4], template matching [2], Haar-like feature based learning [16][17][18] or the Local Binary Pattern based learning [9][15].

In [2], for the head tracking the template matching method is used, which compares the low resolution face image with the input frame fragments. To reduce the search space, the background subtraction is performed to reject still, unchanged pixels. This method of the face tracking proved to be rapid, but the used subtraction is sensitive to the illumination changes, which is frequent in the natural light. It is not applicable on the mobile devices in which camera is not still and moves with the device. Additionally, the template matching method is sensitive to the head tilting, a natural movement when controlling the application with head movement.

For the fast head detection in the camera image, the most popular method is based on the Haar-like features, proposed by Viola and Jones [18]. Method is based on the detection of simple rectangular features, used then for the AdaBoost learning algorithm. With the use of the Integral Image, wherein each pixel represents the sum of values of all pixels from the input image above and to the left of it, it is possible to quickly evaluate the features on the given position and scale. Unfortunately the calculation of the Integral Image is time consuming and requires $2*N$ operations for the image of $N$ pixels. The complexity of the feature calculations is of $O(M*N)$ for features of $M$ scales. In [16] the acceleration of this method is proposed, with use of the GPU, but still it reaches only 19 fps for an image of 1280x960 pixels. In [17] a head tracking method based on Haar-like feature detection is proposed, working at theoretical 500 fps. The search of the entire frame of 512x512 pixels can be executed at 200 fps, on a rather powerful GPU (934 GFLOPS). The search frequency of the high-resolution image of 1920x1080 pixels, on the hand-held device would not exceed 20fps.

A faster tracking method, also based on the AdaBoost learning, but using the Local Binary Pattern is proposed in [9]. The LBP method considers the surroundings of the pixels. Thanks to the calculation on GPU and CPU [15], for a picture of 1024x1024 pixels, method can process at 10fps, but on the mobile device of very low computational power. On current hand-held devices even high-resolution image could be processed at about 25 fps.

Due to the low computational complexity and the possibility of use on the mobile devices, the face tracking methods based on the skin color classification are important[1][4][5][14]. Such a classifiers extract from the RGB space (or other color space) a subspace containing the RGB values corresponding to the possible shades of the skin color. Fuzzy classifiers that determines the pixel degree of membership may provide a smoother tracking with the same resolution image [1]. Multiplatform version of the method described in [1] is presented in Section III.

Despite the linear complexity of the skin color based tracking method, analysis of high resolution images still can be very time consuming, especially on hand-held mobile devices. In this article, we proposed accelerating methods, enabling face tracking on frame of 1920x1080 pixels at about 300 fps, even on the devices with a low processing power.

## III. The Fuzzy Skin-Color Classifier

To maintain the multiplatform usability of the fuzzy head tracking method, first proposed in [1], and to allow its usage on hand-held mobile devices, the omission of the background subtraction is necessary. It is impossible to use due to the continuous motion of the hand-held device and hence its camera.

In the proposed multiplatform head tracking method, the user's head position is determined based on the degree of membership $\mu_S(p)$ of all the pixels to the fuzzy skin-color pixels multiset $S$:

$$S = \{\langle p, \mu_S(p)\rangle : p \in X_{RGB}\} \tag{1}$$

$$\mu_S(p) = \max(0, \min(1, p_R \cdot f_R + p_G \cdot f_G + p_B \cdot f_B + f_D)) \tag{2}$$

Where:

$X_{RGB} = \{<r,g,b>: r,g,b \in [0,255]\}$ – the multiset of all the image pixels.

$p=<p_R ,p_G ,p_B>$ – the components $R, G, B$ of the pixel $p$.

$f=<f_R ,f_G ,f_B, f_D>$ – the vector of the color filter.

The vector of the color filter, specified in the classifier calibration, defines a plane in the RGB space. This plane separates most of the skin color pixels from the background pixels that we are trying to discard. The filter vector specifies also the "positive" side of that plane and tells us which of the two subspaces contains skin-color pixels. The fuzziness of the multiset $S$ provides a gradient boundary between the subspaces. The further the pixel color resides from the plane on its "positive" side, the higher is its degree of membership.

The usage of the RGB space instead of other color model, in which the skin colors are easier to distinguish and isolate, is motivated by the additional computation time needed to perform the transformation from the original RGB image provided by the camera to the more preferable color model.

The usage of just two components (e. g. H and S in HSV model) does not compensate this time in the processing. As the skin tone is separable in the RGB space, the computational complexity remains the major factor in choosing the color model.

To determine the position of the face on the image and its distance from the camera, it is necessary to calculate the cardinality of the $S$ multiset, defined as the sum of the degrees of membership of all $N$ image pixels:

$$|S| = \sum_{i=1}^{N} \mu_S(p_i) \tag{3}$$

We are using all the N pixels of the image as the skin-color face pixels, because the background pixels with the degrees of membership equal 0 does not affect the results. This degrees (Eq. 2) are also the weights of the pixels used to determine the coordinates $C_X$ and $C_Y$ of the centroid $C$ of all the skin-color pixels, calculated as the ratio of the weighted sum of the image pixel coordinates to the cardinality of multiset $S$:

$$C_X = \frac{1}{|S|} \sum_{i=1}^{N} \mu_S(p_i) \cdot x_i \tag{4a}$$

$$C_Y = \frac{1}{|S|} \sum_{i=1}^{N} \mu_S(p_i) \cdot y_i \tag{4b}$$

Where:

$x_i, y_i$ – coordinates of pixel $p_i$

The found centroid coordinates are considered the coordinates of the head center in the image and are used to determine face position in the plane parallel to the screen and the camera. The ratio of the cardinality of the $S$ multiset to the cardinality of multiset of all the image pixels can be used as the $A$ measure of the face area in the image and used to calculate the head distance from the screen:

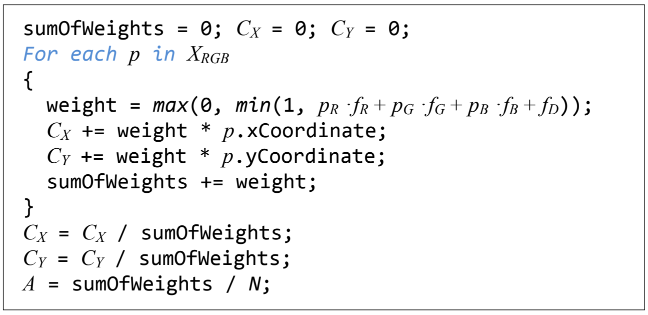
$$A = \frac{|S|}{|X_{RGB}|} \tag{5}$$

What is, according to (Eq. 3):

$$A = \frac{1}{N} \sum_{i=1}^{N} \mu_S(p_i) \tag{6}$$

The algorithm of finding the face position on the camera image is shown in the Fig. 1. The non-modified algorithm is later referred to as "*Version CPU 1"* in comparison with the accelerated variants of the algorithm.

The filter $f$ (Eq. 2), in contrast to most of the skin-color classifier methods, is not designed to match all colors that may belong to the skin of any person, of any race, skin tone, in all possible lightning condition. The primary objective of the filter $f$ is to extract the user face from the current environment visible in the camera. It requires determining the optimal values of the filter parameters for each application usage.



```
sumOfWeights = 0; $C_X$ = 0; $C_Y$ = 0;
For each $p$ in $X_{RGB}$
{
  weight = max(0, min(1, $p_R \cdot f_R + p_G \cdot f_G + p_B \cdot f_B + f_D$));
  $C_X$ += weight * $p$.xCoordinate;
  $C_Y$ += weight * $p$.yCoordinate;
  sumOfWeights += weight;
}
$C_X$ = $C_X$ / sumOfWeights;
$C_Y$ = $C_Y$ / sumOfWeights;
$A$ = sumOfWeights / $N$;
```

Fig. 1 The algorithm of finding the face position on the camera image (pseudocode).

Manual finding of the optimal values may be hard task for the user, and certainly it is uncomfortable and time-consuming. Therefore, in [8] we proposed the automatic method of finding the optimal values of skin-color filter. The method of automatic parameters calculation is based on the analysis of the image with arbitrarily marked area of the face. The user moves the head to place it in the oval-shaped mask visible in the preview of the camera. Once approved, several consecutive frames of the video image are capture for the analysis. Pixels of the obtained images are then use as factors in the objective functions *G(f)* which maximum is searched. To accelerate the computation of the optimum parameters the clustering of the input data transformed to the RGB space is performed with fast grid-based clustering method proposed in [8] where the entire process is described in detail.

The automatic parameters calculation allows for extracting most of the face pixels even from a difficult background with colors close to the skin tone or in the poor lightning. The automatically calculated parameters provide the method stability, ensuring that the $|S|$ value in the Equations (4a) and (4b) does not tend to zero. Still, in some particularly difficult conditions the method does not filter out all of the background pixels, assigning a small part of these pixels with high weights. In practice, the tracking is then still effective. Although the unfiltered pixels "attract" the found centroid to their center of gravity, reducing the amplitude of the estimated head movement determined by the tracking, still the direction of the motion is preserved and its speed is proportional to the actual speed of the head movement. As a result, the perspective of the virtual scene can still be coupled with the head movements, resulting in an immersive sense of depth in the Head-Coupled Perspective technique or effective control in other applications. Therefore, although the method is error-prone in cases when additional skin-colored body parts (e.g. neck or chest) or other faces are visible to the camera, the method still provides sufficient results for proposed applications. Especially on the hand-held mobile devices, the user head is always the biggest skin-colored object visible to the camera and cannot be dominated by other objects. When the head leaves the field of view, due to the device rotation, it may be also recorded by other internal sensors which can then substitute the head tracking.

To minimalize the resulting errors of classification in the particularly difficult conditions, we can utilize more computationally complex skin-color models, like described in [14] Gaussian model or used in [10] elliptical Gaussian chrominance probability density function. We can also transform pixel colors to another color space, e.g. *HSV,* normalized *RGB* or *CIE-XYZ* (with all components divided by the sum of all components), or proposed in [10] the *STV* space. The possible skin colors in these spaces are easier to separate [10] and the impact of the lightning on the classification is reduced. Unfortunately, both methods of improvement increase the computational cost of calculating the pixels weights, almost proportionally increasing the computation time, which actually need to be reduced. Therefore, besides the direct computation time reduction, it is desirable to decouple the head tracking cost from the complexity of the used skin color classifier.

## IV. Accelerating the Head Position Tracking

The computational cost of the presented above head tracking method is of the order *O(N)*, where *N* is the number of image pixels. The method requires a few operations per pixel, and each is processed once and individually. Although the linear computation cost seems to be low, for high resolution video image of 1920 x 1080 pixels the processing of the full frame in the real time is hard to achieve, even without the additional CPU load.

The accelerating methods that decrease the problem size, reduce at the same the tracking quality. The downscaling of the input image [15][16] requires additional computation, decreases the number of input data and reduces the angular resolution of the tracking. Searching for the face only in the neighborhood of the previously found head position [9][16][17] makes the tracking sensitive to the fast motion of the head, providing just a slight acceleration, due to the large area occupied by the face on the image of the narrow-angle camera.

The described below accelerating methods reduce the computational cost of the head tracking without decreasing the number of the analyzed pixels, to maintain high quality of the tracking. When the lower quality of the tracking is allowed, these methods can be successfully combined with the problem size reduction for the further tracking acceleration.

### *A. The Head Tracking Acceleration on the GPU*

The acceleration of the head tracking can be achieved by transferring the calculations to the graphics processor unit, which allow for the parallel analysis of many pixels. Besides the possible computation time reduction, this approach may additionally decrease the energy usage of the mobile devices during the frame processing [9]. The use of the Nvidia® CUDA® framework allow for using the full capabilities of graphics processors compatible with this architecture. Such GPUs are widely installed on the laptop computers and they are recently available on the hand-held mobile devices. The tests of the methods described below are performed on the laptop computer equipped with the Intel® Core™ i5-2450M mobile processor and the Nvidia® GeForce® GT 630M mobile GPU.

As the classification of the pixel colors in our method is performed for each pixel individually, the acceleration is seemingly easy to achieve by performing the computations on the graphics card. The parallelization of the calculation of the pixels degrees of membership to the skin-color multiset *S* is trivial on GPU. Each thread must calculate the weight of one pixel, according to Equation 2. A part of the kernel (i.e. a function in CUDA executed in parallel by multiple threads on GPU) responsible for the pixels weights calculation is as follows:

```
uint index=threadIdx.x + blockIdx.x*blockDim.x;
Pixel p=frame[index];
float weight=f[0]*p.r + f[1]*p.g + f[2]*p.b+ f[3];
if(weight>0.0f){ if(weight>1.0f)weight=1.0f; }
else weight=0.0f;
[…]
```

A constant delay is generated by the transfer of the image to the GPU memory. For an image of 1920x1080 pixels, the transfer takes about 2 milliseconds.

For the further analysis, the calculation of the head position is split into two parts:

1) Calculating the degrees of membership from Eq. 2.
2) Finding the centroid coordinates (Eq. 4) and the sum of all the pixel weights need in Eq. 6.

On the CPU, the second part of the calculation represents only a small fraction of all the computations. Conversely, on the GPU the computation of the second part can be several times longer than the first part. The time of 2) on the graphics processor depends largely on the used algorithm. Below, different methods of accelerating computation on GPU are compared.

#### *1) Version GPU 1 – Atomic adding to one global value*

In the naive approach to calculate the coordinates of the centroid, all the threads can add its calculated values directly to the output sums, using the *atomic* add function:

```
[…]
atomicAdd(x,weight*(index%width));
atomicAdd(y,weight*(index/width));
atomicAdd(w,weight);
```

*Atomic* functions in CUDA allows multiple threads to modify common data, with the guarantee of receiving the correct result, which means that every thread perform the operation exactly once, without data loss from the simultaneous access to the output value.

Unfortunately, in this approach the summing is performed sequentially, as only one thread at a time increases the output sum. Moreover this variant requires a very frequent access to the very slow global memory. It is very inefficient approach, which do not use the full power of the GPU and prolongs the calculation compared to the CPU computing.

### *2) Version GPU 2 – Atomic adding to the shared memory*

The acceleration can be obtained when first the thread values are atomically added to local sums of the blocks in the faster shared memory, and then only this local sums are atomically added to the global output values, by one thread per block:

```
__shared__ float sX,sY,sW;
sX=0.0f;
sY=0.0f;
sW=0.0f;
[…]
atomicAdd(&sX,weight*(index%width));
atomicAdd(&sY,weight*(index/width));
atomicAdd(&sW,weight);

__syncthreads();

if( 0 == threadIdx.x )
{
  atomicAdd(x,sX);
  atomicAdd(y,sY);
  atomicAdd(w,sW);
}
```

In this approach a part of the internal aggregations is executed in parallel between several thread blocks. Therefore this variant is more than three times faster than the *Version GPU 1.*

### *3) Version GPU 3 – Parallel aggregation with divide and conquer*

Although in *Version GPU 2* part of the summation is performed in parallel between blocks, still only one thread per block can add its value to the local sum at the same time. To execute more parallel addition inside a thread block it is possible to use the divide and conquer method, decreasing the complexity order of the summation for each *N*-thread block from *O(N)* to *O(logN)*. In this case, half of the block threads must sum up pairs of values calculated by the consecutive threads. In the next steps, the sums from previous step are summed in pairs until there is only one final sum of all the block values (Fig. 2).

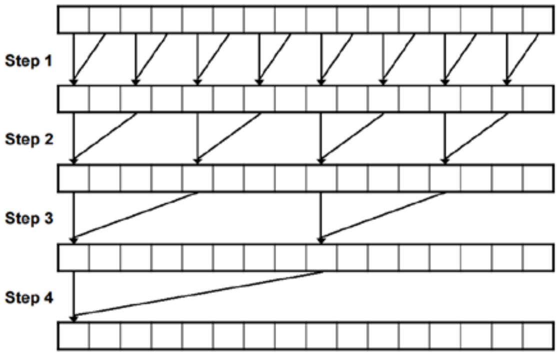


Fig. 2 The parallel aggregation in block

The kernel realizing directly this approach is as follows:

```
__shared__ float tempX[THREADS_PER_BLOCK];
__shared__ float tempY[THREADS_PER_BLOCK];
__shared__ float tempW[THREADS_PER_BLOCK];
[…]
uint s,temp;
for(s=1;s<THREADS_PER_BLOCK;s<<=1)
{
  if(0==threadIdx.x%(s<<1))
  {
    __syncthreads();
    temp=threadIdx.x+s;
    tempX[threadIdx.x]+=tempX[temp];
    tempY[threadIdx.x]+=tempY[temp];
    tempW[threadIdx.x]+=tempW[temp];
  }
}
if( 0 == threadIdx.x )
{
  atomicAdd(x,tempX[0]);
  atomicAdd(y,tempY[0]);
  atomicAdd(w,tempW[0]);
}
```

This approach decrease the computational time by 40% compared with the *Version GPU 2*. The acceleration is less than expected due to the highly divergent branching of the code in the condition `(0==threadIdx.x%(s<<1))`, what is discussed by Harris [11] on an analogous example.

### *4) Version GPU 4 – Without divergent branching*

Further acceleration can then be obtain by alternatively engaged threads, as shown in Fig. 3.

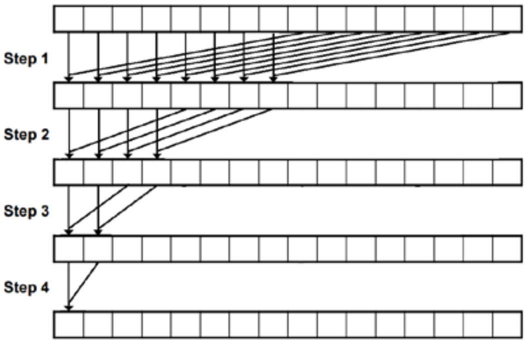


Fig. 3 The parallel aggregation without the divergent branching and bank conflicts

This approach omits the problem of divergent branching, without additional shared memory bank conflicts. Analogous solution for similar problem was used in [13].

The internal aggregation in this version looks as follows:

```
[…]
uint s,temp;
for(s = THREADS_PER_BLOCK>>1; s>0; s>>=1)
{
     if(threadIdx.x<s)
     {
       __syncthreads();
       temp=threadIdx.x+s;
       tempX[threadIdx.x]+=tempX[temp];
       tempY[threadIdx.x]+=tempY[temp];
       tempW[threadIdx.x]+=tempW[temp];
     }
}
if( 0 == threadIdx.x )
{
  atomicAdd(x,tempX[0]);
  atomicAdd(y,tempY[0]);
  atomicAdd(w,tempW[0]);
}
```

This variant is 2.5 times faster than *Version GPU 3*.

### *5) Version GPU 5 – Multiple aggregation per thread*

In the *Version GPU 4* during the iterative aggregation an average of ¾ of the block threads is idle, from the half in the first iteration, to all but one in the final aggregation. At the same time, due to the limit of thread number per block, we receive a large number of blocks. The possible solution, proposed by Harris for similar problem [11], is to use threads to iteratively add up a greater number of values. In our case it requires also the calculation of more pixel weights (i. e. degrees of membership to skin color multiset) per thread. Such a solution can significantly better harness the GPU computational power. Before we get to the inter-thread aggregation, all the threads are engaged in a long non-synchronized work.

After this modification, kernel is as follows:

```
__shared__ float tempX[THREADS_PER_BLOCK];
__shared__ float tempY[THREADS_PER_BLOCK];
__shared__ float tempW[THREADS_PER_BLOCK];
uint ti = threadIdx.x;
uint index = ti+blockIdx.x * THREADS_PER_BLOCK;
uint gridSize = THREADS_PER_BLOCK * gridDim.x;
Pixel p;
float weight;
tempX[ti]=0;  tempY[ti]=0;  tempW[ti]=0;
while(index<N)
{
  p=frame[index];
  weight=f[0]*(p.r)+ f[1]*(p.g)+ f[2]*(p.b)+ f[3];
  if (weight>0){if(weight>1.0f) weight=1.0f;}
   else weight=0.0f;
  tempX[ti]+=weight*(index%width);
  tempY[ti]+=weight*(index/width);
  tempW[ti]+=weight;
  index+=gridSize;
}
uint s,temp;
for(s= THREADS_PER_BLOCK>>1;s>0;s>>=1)
{
  if(ti<s)
  {
    __syncthreads();
    temp=ti+s;
    tempX[ti]+=tempX[temp];
    tempY[ti]+=tempY[temp];
    tempW[ti]+=tempW[temp];
  }
}
if( 0 == ti )
{
  atomicAdd(x,tempX[0]);
  atomicAdd(y,tempY[0]);
  atomicAdd(w,tempW[0]);
}
```

Due to the reduced number of blocks, this variant is 2.5 times faster than *Version GPU 4*.

### *6) Version GPU 6 – Double memory access*

Further acceleration by about 10% can be obtained by adding two values per iteration and grouping operations. It is the result of the GPU's memory access, where reading two consecutive 4-bytes words has a similar cost to reading just one word [12]. The number of blocks is halved:

```
uint index = ti+blockIdx.x*THREADS_PER_BLOCK * 2;
uint gridSize = gridDim.x*THREADS_PER_BLOCK * 2;
[…]
while(index<N)
{
 index2=index + THREADS_PER_BLOCK;
 p=frame[index];
 p2=frame[index2];
 weight=f[0]*(p.r)+ f[1]*(p.g)+ f[2]*(p.b)+ f[3];
 weight2=f[0]*(p2.r)+f[1]*(p2.g)+f[2]*(p2.b)+f[3];
 if (weight>0){ if(weight>1.0f) weight=1.0f; }
  else   weight=0.0f;
 if(weight2>0){if(weight2>1.0f) weight2=1.0f;}
  else weight2=0.0f;
 tempX[ti]+=weight*(index%width)
           +weight2*(index2%width);
 tempY[ti]+=weight*(index/width)
           +weight2*(index2/width);
 tempW[ti]+=weight + weight2;
 index+=gridSize;
}
```

### 7) Version GPU 7 – Aggregation in local memory

In versions *GPU 5* and *GPU 6* designed on the basis of [11], highly ineffective is the iterative aggregation of all the values calculated by the thread to the shared memory. During this aggregation other threads do not need any access to this temporary sum. Our proposed solution is to aggregate the thread values in its local memory, as it is faster than the shared memory. Only the final thread sum should be copied to the shared memory for access of other threads.

```
[…]
float tX,tY,tW;
tX=0;  tY=0;  tW=0;
while(index<N)
{
 […]
 tX+=weight*(index%width)+weight2*(index2%width);
 tY+=weight*(index/width)+weight2*(index2/width);
 tW+=weight + weight2;
index+=gridSize;
}
tempX[ti]=tX;
tempY[ti]=tY;
tempW[ti]=tW;
[…]
```

This approach decrease the computational time by 10% compared with the *Version GPU 6*.

### 8) Version GPU 8 – Multiple pixels per thread with atomic adding to shared memory

The divide and conquer parallelization in the *Version GPU 4* leads to a four times faster computing, compared to *Version GPU 2,* where the tread values are atomically added to the block sum. But since each thread aggregates weights of hundreds of pixels, the synchronization of threads before each iteration of inter-thread summing is very time consuming. It appears that the atomic summation to one value is more effective in that case. Even though the acceleration of such approach is only about 3%, the additional profit is the reduction of the shared memory occupancy and a slight kernel code simplification:

```
__shared__ float tempX,tempY,tempW;
if(0==threadIdx.x)
{
  tempX=0; tempY=0; tempW=0;
}
uint ti=threadIdx.x;
uint index=ti+blockIdx.x* THREADS_PER_BLOCK *2;
uint gridSize = THREADS_PER_BLOCK *2*gridDim.x;
Pixel p;
float weight,weight2;
unsigned int index2;
float tX,tY,tW;
tX=0;  tY=0;  tW=0;
while(index<N)
{
 index2=index + THREADS_PER_BLOCK;
 p=frame[index];
 p2=frame[index2];
 weight=f[0]*(p.r)+ f[1]*(p.g)+ f[2]*(p.b)+ f[3];
 weight2=f[0]*(p2.r)+f[1]*(p2.g)+f[2]*(p2.b)+f[3];
 if(weight>0) { if(weight>1.0f)  weight=1.0f;}
  else weight=0.0f;
 if(weight2>0){if(weight2>1.0f) weight2=1.0f;}
 else  weight2=0.0f;
 tX+=weight*(index % width)
     +weight2*(index2 % width);
 tY+=weight*(index / width)
     + weight2*(index2 / width);
 tW+=weight + weight2;
 index+=gridSize;
}
atomicAdd(&tempX,tX);
atomicAdd(&tempY,tY);
atomicAdd(&tempW,tW);
__syncthreads();
if( 0 == ti )
{
  atomicAdd(x,tempX[0]);
  atomicAdd(y,tempY[0]);
  atomicAdd(w,tempW[0]);
}
```

The final version of the kernel is almost 45 times faster than the *Version GPU 1.*

## B. The Head Tracking Acceleration on the CPU

Although the computation of head position on the GPU can be very fast (with frame computation below 5 ms), it can be insufficient in the 3D graphic applications like video games, where the 3D rendering loads the GPU to its limits. To maintain the performance of the application, we may need to limit the GPU computations.

Unfortunately, the head position estimation on the CPU is time consuming, as its computational cost is of the order *O(N).* Even with an effective implementation of the algorithm (Fig. 1), on the testing platform the 30 fps is not achieved (see Table 1). Therefore it is desirable to accelerate also the CPU computing.

### 1) Version CPU 2 – The weights precalculation

The acceleration on the CPU can be obtain with the precalculation of all the possible values of $\mu_S(p)$ (Eq. 2) and referring to them instead of calculating this value for every input pixel individually. The number of all the RGB colors with 8-bits channels is limited and amounts to $256^3$. It is 8 times more values than in a frame of 1920x1080 pixels, but the precalculation can be performed only once before the start of the application, as the colors weights change only when the filter *f* values change.

The precalculated weights assigned by the classifier to all the RGB values, can be stored in an array *W*[], in which the pixel *p* weight is located at the given position:

$$W\left[p_R \cdot 256^2 + p_G \cdot 256 + p_B\right] = \mu_S(p) \qquad (7)$$

With the indexing relevant to the input pixel format (i. e. its byte order), the integer value written on the four bytes of the pixel is also the position in the array *W,* at which the pixel weight is stored.

With the use of the precalculated weights, the CPU computations are faster only by 35%. Even though this solution is interesting for other reason. As was mentioned in the previous section, it is desirable to decouple the computing time from the color classification method, for the possible classification improvement without the reduce of the tracking speed. With this solution, the usage of more complex color model or the transformation to the other color space, increases only the once performed precalculation time and the tracking time remain unchanged, so that aim is fully achieved.

Applying this concept for the GPU is not recommended, as it results in prolonged, twice as long calculation time. This is caused by the "random", irregular access to the array *W,* as the adjacent pixels can have different colors, distant in the RGB space. As a result the global memory access time is increased, slowing the entire computing. Moreover, on the GPU computation cost is already almost independent of the used color model. A slight gain from the precalculation may be achieved only with use of a very complex color classifiers.

### *2) Version CPU 3 – The precalculation and multicore processing*

The parallelization of the calculations is not limited to the GPU computing. Most modern CPUs have at least two independent processing units (called *cores*). Therefore, the equal distribution of the calculations to more threads may results in almost direct proportional time reduction.

By dividing the problem, i.e. the input frame, between multiple CPU cores, we can compute pixel weights from each part of the image, and the centroid of these pixels. The centroid of all the partial centroids is also the centroid of the whole image.

### *3) Version GPU 8 + CPU 3– The joint GPU and CPU calculations.*

As in the above example, the problem can be divided into two parts, one of which is implemented on the CPU, and the second on the GPU. Such division may be dictated by the GPU limitation, or by the need for utilizing all the available computing power for the further tracking acceleration.

In the second case, in order to achieve the greatest acceleration, the division of the problem between the CPU and the GPU should not be symmetrical as in multicore processing. For the highest resulting speed, the two processors should complete their computations at the rather same time. If average times of the full frame computation on CPU and GPU are respectively $T_C$ and $T_G$, and they analyze sub-images $I_C$ and $I_G$ of the image I (I= $I_C$ +$I_G$), than the sizes of the sub-images should be:

$$I_C = \frac{T_G}{T_C + T_G} \cdot I \qquad (8a)$$

$$I_G = \frac{T_C}{T_C + T_G} \cdot I \qquad (8b)$$

Hence, in our case, at the testing laptop, the highest speed is achieved when 70% of the image pixels is processed on the GPU (with *Version GPU 8*) and 30% on the CPU (with *Version CPU 3*). The comparison of the computation times of all the methods is presented in the next section.

### *4) Other possible acceleration methods*

The possible further acceleration of the computation may be achieved by the replacement of the floating point operations with the fixed point calculations and especially integer operations. In order to do this, we must upscale filter *f* values by *u* to the integer values. The maximum weight assigned by the classifier (1 in the Equation 2) must also be upscaled by the same *u* value. This way the *S* multiset would no longer meets the definition of the fuzzy set, but the results of head position estimation would not change if we only divide the *A* (Eq. 5 and 6) by the *u*. The problem of this solution is that we may must restrict the maximum resolution of the image or the resolution of the classifier (i.e. number of different weights it may assign) to ensure that we do not exceed the maximum of the 32 bits variables. Yet, it may be acceptable in some applications and the resulting acceleration may be significant.

Another way to accelerate CPU computation may be the usage of Streaming SIMD Extensions which allow for parallel processing of up to four values. However the usage of SSE may be restricted and may depends on the target platform. Altough this solution was not applied it is worth consideration.

## V. RESULTS

The fast head position tracking methods proposed in the Section IV were compared with the original algorithm of *Version CPU 1* shown in Fig. 1. The tests was performed on the laptop computer equipped with the mobile processor Intel® Core™ i5-2450M (2x2,5GHz) with processing performance of 34,5 GFLOPS and the mobile GPU Nvidia® GeForce® GT 630M with 307,2 GFLOPS processing performance.

The comparison of the average computation times for frames of 1920x1080 pixels is shown in the Fig. 4 and in the

Table 1. Although the times of CPU and GPU should not be compared due to the different architectures, the scale of the possible acceleration of the head tracking can be seen in the Fig. 4. The computations on the GPU (*Version GPU 8*) are over seven times faster, including the time of transfer to the GPU memory, and it allows for tracking at 200 frames per second. In an architecture where the camera image is saved directly in the GPU memory, without need for additional copying, over 300 fps can be achieved.

The acceleration methods basing on the CPU also give good results and over three times faster computing compared to the original algorithm. Using CPU with more processing cores, the further acceleration is possible.

When the processing powers of GPU and CPU are not restricted, the method combining the calculations on both units may result in the greatest tracking speed. On the testing platform the analysis of the full frame was performed in about 3,6 milliseconds (Fig. 4 and Table 1).

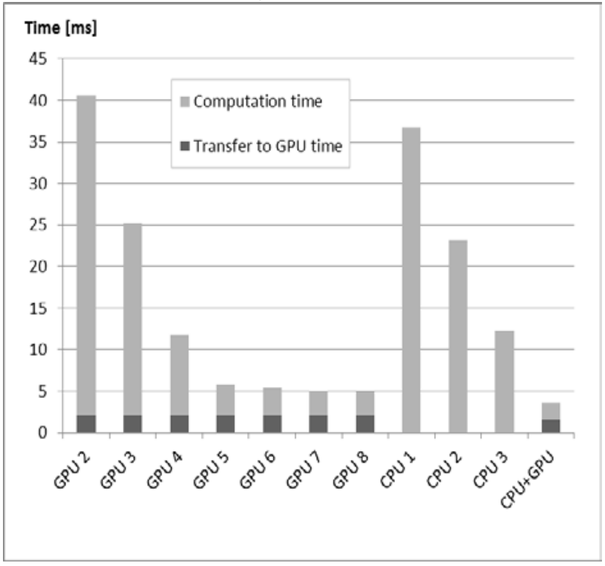


Fig. 4 The processing times of one frame of 1920x1080 pixels (average of 100 trials).

## VI. Conclusions and Future work

Although the device used in the tests was not a hand-held device, its CPU and GPU has respectively 50% and 15% lesser processing performance than the Nvidia® Shield™ Tablet with Tegra K1 mobile processor equipped with CPU ARM Cortex-A15 R3 (4x2,3GHz), with the processing performance of 70,4 GFLOPS and CUDA-enabled GPU Kepler with 364,8 GFLOPS processing performance. Also, the popular smartphone processors Qualcomm® Snapdragon™ 810 has a similar processing performance. Hence, the testing laptop computer represents well the computing power of the today hand-held devices.

The achieved head tracking times leave a large margin of error for the real-time tracking with over 60 fps, even in the case of a highly loaded GPU and CPU or when performed on the devices with a lot less computing power. With the achieved tracking times it becomes possible to use the found head position for effective Head-Coupled Perspective implementation in combination with the interactive 3D applications on hand-held mobile devices.

The future work includes testing the proposed methods on the actual hand-held devices and designing the model for determining the head position relative to the device with use of the internal sensors (accelerometer and gyroscope) combined with the proposed fast head tracking method.

## References

[1] M. Szkudlarek, M. Pietruszka, "Head-Coupled Perspective in Computer Game", In: Journal of Applied Science, vol. 21, no. 2, pp. 165-179, 2013.

[2] J. Rekimoto, "A vision-based head tracker for fish tank virtual reality-VR without head gear," Virtual Reality Annual International Symposium, 1995. Proceedings, IEEE, pp.94-100, 1995.

[3] J. Francone, "Using the User's Point of View for Interaction on Mobile Devices", 23rd French Speaking Conference on Human-Computer Interaction, pp. 41-48, New York, 2011.

[4] W. Gaver, G. Smets, K. Overbeeke, "A Virtual Window on media space". In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95), New York, pp.257-264, 1995.

[5] A. Bulbul, "A Face Tracking Algorithm for User Interaction in Mobile Device", CyberWorlds, 2009. CW '09. International Conference on, IEEE, pp. 385 – 390, 2009.

[6] J. Hwang, J. Jung, and G. J. Kim. "Hand-held virtual reality: A feasibility study", in ACM Virtual Reality Software and Technology, pp. 356–363, 2006.

[7] N. Joshi, A. Kar, and M. Cohen. "Looking at you: fused gyro and face tracking for viewing large imagery on mobile devices". In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12), pp. 2211-2220, 2012.

[8] M. Szkudlarek, M. Pietruszka, "Fast Grid-Based Clustering Method for Automatic Calculation of Optimal Parameters of Skin Color Classifier for Head Tracking". In Proceedings of 2015 IEEE 2nd International Conference on Cybernetics (CYBCONF), 2015.

[9] M. B. Lopez, J. Hannuksela, O. Silven, F. Lixin, "Head-tracking virtual 3-D display for mobile devices," Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on , pp.27-34, 2012.

[10] J.-C. Terrillon, M. David, "Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments," Automatic Face and Gesture Recognition, Proceedings.

Table I.
The processing times of one frame of 1920x1080 pixels (average of 100 trials).

| | GPU 1 | GPU 2 | GPU 3 | GPU 4 | GPU 5 | GPU 6 | GPU 7 | GPU 8 | CPU 1 | CPU 2 | CPU 3 | CPU + GPU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Transfer time [ms] | 2,05 | 2,05 | 2,05 | 2,05 | 2,05 | 2,05 | 2,05 | 2,05 | - | - | - | 1,54 |
| Calculation [ms] | 131,34 | 38,59 | 23,19 | 9,68 | 3,81 | 3,36 | 3,06 | 2,94 | 36,74 | 23,12 | 12,34 | 2,12 |
| Total time [ms] | 133,39 | 40,64 | 25,24 | 11,73 | 5,86 | 5,41 | 5,11 | 4,99 | 36,74 | 23,12 | 12,34 | 3,66 |
| CPU 1 / Total time | 0,28 | 0,90 | 1,46 | 3,13 | 6,27 | 6,79 | 7,19 | 7,36 | 1,00 | 1,59 | 2,98 | 10,04 |
| Frames per Second | 7,50 | 24,61 | 39,62 | 85,25 | 170,65 | 184,84 | 195,69 | 200,4 | 27,22 | 43,25 | 81,04 | 273,22 |

Third IEEE International Conference on,pp.112-117, 1998.

[11] M. Harris, “Optimizing Parallel Reduction in CUDA”, NVIDIA Developer Technology, 2007.

[12] J. Luitjens, S. Rennich, “CUDA Warps and Occupancy”, GPU Computing Webinar, 2011.

[13] D. Xie, L.Dang, R. Tong, “Video Based Head Detection and Tracking Surveillance System”, 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012), IEEE, pp. 2832-2836, 2013.

[14] Y.-W. Wu, X.-Y. Ai, "Face Detection in Color Images Using AdaBoost Algorithm Based on Skin Color Information," Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on , pp.339-342, 2008.

[15] M. B. López, H. Nykänen, J. Hannuksela, O. Silvén, M. Vehviläinen, “Accelerating image recognition on mobile devices using GPGPU”, IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, pp. 78720R-78720R, 2011.

[16] B. Sharma, R. Thota, N. Vydyanathan, A. Kale, "Towards a robust, real-time face processing system using CUDA-enabled GPUs," High Performance Computing (HiPC), 2009 International Conference on, pp.368-377, 2009.

[17] I. Ishii, H. Ichida, T. Takaki, "GPU-based face tracking at 500 fps", Image Processing (ICIP), 2011 18th IEEE International Conference on, pp.557-560, 2011.

[18] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features," Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol.1, pp.I-511,I-518, 2001.

[19] J. Gośliński, M. Nowicki, P. Skrzypczyński, "Performance Comparison of EKF-Based Algorithms for Orientation Estimation on Android Platform," Sensors Journal, IEEE, vol. 15, no. 7, pp. 3781 - 3792, 2015.

DOI: 10.15439/2015F85
ACSIS, Vol. 5

# Visual Detection of Objects by Mobile Agents using CBVIR Techniques of Low Complexity

Andrzej Śluzek

Khalifa University, Abu Dhabi Campus
P.O. Box 127788, Abu Dhabi, UAE
Email: andrzej.sluzek@kustar.ac.ae

***Abstract*—Visual search for objects of interest in complex environment is an important (and still challenging) problem in mobile robotics. In particular, the usage of *content-based visual information retrieval* (CBVIR) methods, which are a natural choice for such tasks, is often constrained by the real-time requirements, and the mobility of searching agents is sometimes not sufficiently exploited in the search model. In this paper, a CBVIR-based scheme is proposed, which takes into account motion of the searching agents to achieve a low-cost and high-speed detection of objects of interest in cluttered scenes, with good overall performances. We combine standard CBVIR tools, i.e. MSER detector and SIFT descriptor (quantized into sufficiently large vocabularies) assuming additionally that objects become *objects of interest* only when approached closely enough by the mobile agent, i.e. when seen at an adequately large scale. Thus, an object of interest is considered detected only if a sufficient number of keypoints from the current video-frame are matched (including the corresponding matches of scales) to the keypoints from the database images of the object. Preliminary experiments on a limited-size dataset confirm performances of the scheme, although in the classical task of video-frame retrieval the scheme cannot compete with more sophisticated CBVIR algorithms. The scheme can prospectively become more flexible if combined with a range-finding device so that the approximate distances to the scene components within the currently inspected part of the image can be used to proportionally modify the scale correspondences.**

## I. Introduction and Background Works

Visual search for unpredictably located objects of interest (where such object are represented by their exemplary images) remains one of important tasks for mobile robotics. Understandably, CBVIR is a natural source of algorithms for such a task. Actually, one of the fundamental CBVIR concepts of *visual words* was originally proposed for videos [1]. Algorithms for sub-image retrieval (where the objective is to identify images containing fragments near-duplicate to the query image) are particularly important. Those algorithms are usually based on detecting keypoints and, subsequently, comparing their visual words. Generally, however, three major differences exist between real-time visual search by mobile agents and classical CBVIR tasks:

1) In CBVIR, infrequently arriving query images (submitted by the users) are matched against large/huge datasets, e.g. [2], [3], while in a camera-based visual search very small datasets (template images of objects of interest) are matched against large numbers of continuously arriving queries (video-frames acquired by a single camera or by simultaneously working multiple cameras attached to a mobile agent).
2) Because of (1), the computational costs of image preprocessing (i.e. keypoint detection and description, quantization into visual words, etc.) are in visual search as critical as the complexity of the actual image matching and retrieval. In classical CBVIR, the costs of image preprocessing are considered negligible.
3) In visual search, the objective is to detect all instances of interesting objects, where each instance is represented by a sequence of frames in the input video stream. However, not all such frames have to be perfectly identified. Thus, in the video search (unlike in standard CBVIR tasks) *recall* of individual frame retrieval can be compromised, but *precision* should be as high as possible.

As an illustrative example, Fig. 1 shows an object of interest and an exemplary video-frame returned by the search algorithm.

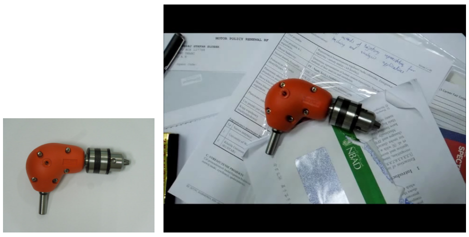

(a) (b)

Fig. 1. The object of interest (a) and an exemplary frame containing it (b).

In this paper, we propose and preliminarily evaluate a scheme which exploits the above characteristics of visual search (and mobility of the searching agent) to achieve at low costs a high performance object detection in cluttered scenes. As the basic components of this scheme, we use standard CBVIR tools, i.e. MSER detector [4] and SIFT descriptor [5] (in its RootSIFT variant [6]). Descriptors are quantized into a large vocabulary of one million words (to assure satisfactory *precision*). To avoid high costs of descriptor quantization into words, a simple quantization method based on the statistical properties of descriptors is used. Details of the image preprocessing phase are described in Section II.

978-83-60810-66-8/$25.00©2015, IEEE

Image matching is performed using the most straightforward criterion, i.e. the number of keypoint correspondences (e.g. [7], [5]) where a match is defined by identical visual words. However, we reject matches for which the ellipses of MSER keypoints are not in the correspondingly similar scales. Because the search is conducted by mobile agents, this requirement indicates that the agent finds an object at a specific distance (defined by the scale of the object images in the database). More details and explanations are given in Section III.

In Section IV, preliminary experimental results of the proposed approach are overviewed. In particular, performances of the method are compared to alternative solutions.

## II. Principles of Image Pre-processing

For a feature-based real-time visual search, efficient detection and description of keypoints is a critical factor. In particular, the number of keypoints should not be excessively large (and controllable in some sense). Thus, we use MSER detector which is affine invariant, has good performances (as reported in [8]), low complexity (which can be further reduced by using special techniques proposed for MSER detection in video sequences, e.g. [9]) and a few tuning parameters to control the numbers of detectable keypoints. Actually, recent implementations of MSER detectors in hardware, e.g. [10], [11], achieve a throughput approaching hundreds of frames per second, which indicates that MSER keypoint extraction is not a critical factor in a real-time image pre-processing, even if several cameras are simultaneously used.

Hardware implementations of SIFT descriptors (and detectors too) have been reported as well (most recently in [12]). Notably, development of a system-on-chip SIFT descriptor for affine-invariant keypoints is currently under way in our organization as well. Therefore, the feasibility of real-time SIFT description of detected keypoints can be considered documented. Even the Matlab implementation of MSER detector combined with a SIFT description module in C++ provide a throughput of 2-3 frames/sec (including all disk read/write overheads) which can be accepted as near real-time performances for slowly moving agents.

The final step of image pre-processing for CBVIR is *quantization* of descriptors into *visual words*. The size of vocabularies can be very diversified, typically ranging from a few thousand to a few million. Although small vocabularies provide better *recall* of keypoint matching, *precision* is generally unacceptably low. *Precision* obviously improves (at the expense of *recall*) with the growing size of vocabulary, but if the vocabulary becomes too large, the quantization intervals could be smaller that natural fluctuations of descriptors, and it might be difficult to find matches even in pairs of almost identical images.

Published results (e.g. [13], [3]) indicate that the recommended sizes of visual vocabularies are in the range of millions of words, especially if *precision* is more important than *recall* (which is the case in visual search by mobile agents, see Point (3) in Section I). Thus, RootSIFT descriptors are quantized into a vocabulary of 1M words (although tests have been conducted using smaller sizes as well - see Section IV). For such a large vocabulary, the standard quantization of descriptors into words by the (approximate) nearest neighbour approach could be a bottleneck in real-time processing of video frames. Instead, the descriptor space has been partitioned off-line into hypercubes of similar probabilities (the probability density was estimated using over 500 million keypoints from diversified images). Then, the descriptor quantization requires only a small number of additions and comparisons, and the processing time is negligibly small.

## III. Image Matching and Object Detection

In visual search, the objective is to identify video fragments (sequences of frames) containing the object(s) of interest, regardless the background visual contents. From CBVIR perspective, this is a problem of *partial near-duplicate* detection, for which a fully satisfactory solution has not been found yet. Nevertheless, most of the *state-of-the-art* methods seem to follow the same two-step approach. First, similarities between individual keypoints are established (using descriptors or visual words). Then, the geometric consistencies between groups of preliminarily matched keypoints are verified to detect clusters of similarly transformed keypoints (which are considered the near-duplicate fragments). Either more advanced algorithms, like the Hough transform, RANSAC, etc. are used (e.g. [14], [15], [16], [17]) to provide more credible results at higher computational costs, or simplified approaches (e.g. [18], [2], [3]) more suitable for large-scale applications are alternatively employed to verify the consistencies.

In the proposed scheme, we detect partial near-duplicates (presumably representing the objects of interest) in a way that merges the first step with a very simple variant of the geometric verification (where only the scale consistency of matching keypoints is verified). Altogether, the level of similarity between a query image (i.e. a video frame) and a database image of an object is defined by the number of keypoint correspondences, where two MSER keypoints $K_1$ (a query keypoint) and $K_2$ (a database keypoint) match if:

Definition 1.

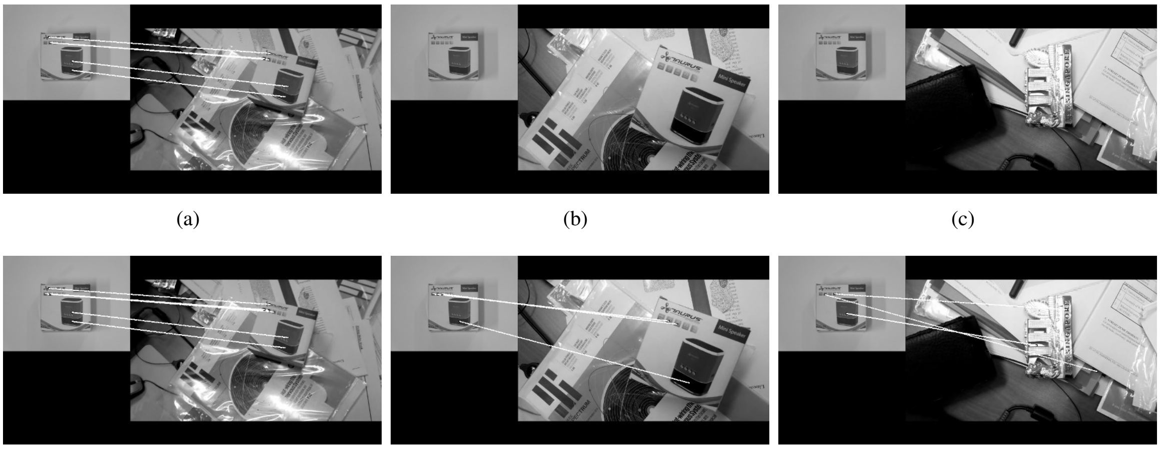
1) The keypoints are described by the same visual word, i.e. $word(K_1) = word(K_2)$.
2) The keypoints have similar scales. Assuming that $M$ and $m$ indicate, correspondingly, the length of major and minor axes of the keypoint ellipses, the conditions for the scale consistency are:

$$0.8M(K_2) \leq M(K_1) \leq 1.2M(K_2), \tag{1}$$

$$0.8m(K_2) \leq m(K_1) \leq 1.2m(K_2). \tag{2}$$

The second requirement of the above definition can be justified as follows:

> *The visual scale of an object in a captured video obviously corresponds to the distance between this object and the camera. When a mobile agent explores its environment, it is expected to recognize*



(a) (b) (c)

(d) (e) (f)

Fig. 2. Exemplary matches obtained by using the scale verification (a, b, c), and without such a verification (d, e, f). A vocabulary of 1M words is applied. In (b) the object is not detected because its scale is too large (but it is detected in (e) where the scale is not verified). In (c), scale verification prevents detection of a non-existing object (falsely detected in (f)).

*objects of interest when they are approached at a sufficiently close distance, i.e. at least at a predefined threshold distance. Thus, the database should contain images of objects of interest in the* ***reference scales*** *approximately corresponding to such threshold distances. The images in larger scales are not needed because the objects should be detected earlier (at the threshold distance) while smaller scales represent objects too distant to be interesting for the agent. Therefore images in larger or smaller scales are not included in the database.*

The axes length tolerance in Eqs 1 and 2 is rather wide (and taken independently for major and minor axes) so that not only small scale deviations but also minor viewpoint changes (up to approx. $30^o$) are generally accepted by the matching algorithm.

A similar philosophy (although with much less efficient tools for keypoint detection and matching) was behind the results presented in [19].

Eventually, two images are considered partial near-duplicates (i.e. a part of the query frame matches the object of interest) if at least 4 pairs of keypoint correspondences are found according to Definition 1. This is the minimum number of matches needed for the verification of affine transformation consistency between images (three pairs to build the transformation, and the fourth one to verify it). Although currently only the scale consistency is applied, such a geometric verification might be used in the future for more advanced tasks (e.g. in determining the number of the same objects of interest in a single video frame).

This matching method is sufficiently fast for visual search tasks considered in this paper. If the images are pre-processed (i.e. MSER keypoints are extracted and assigned visual words, which are very fast operations as outlined in Section II) even the Matlab implementation provides a throughput of approx. 50-60 video frames of VGA resolution per second (i.e. the search can be conducted using 2-3 simultaneously working cameras).

Examples in Fig. 2 highlight the principles and specific characteristics of object detection using the proposed image matching technique.

## IV. Experimental verification

The proposed scheme has been preliminarily verified on a number of short (i.e. $30 - 60$ seconds) videos captured in heavily cluttered indoor environments. A small collection of objects of interest has be arbitrarily proposed (see Fig. 3).

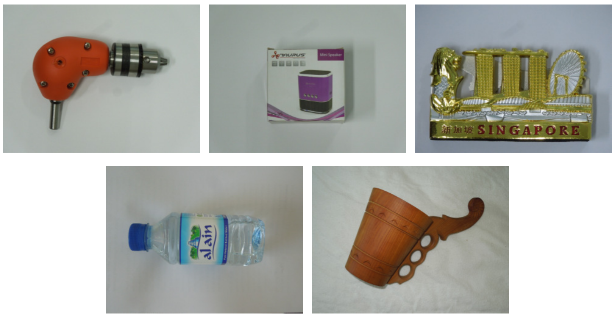


Fig. 3. Examples of objects of interest.

The objective is to identify all instances (but not necessarily all frames of the video) of the objects which are seen for some time at the reference or larger scale (i.e. more distant appearances of the objects are not counted). Fig. 4 shows a few frames from an exemplary beginning (when the object becomes sufficiently large) and from an exemplary end (when the object becomes too small and/or disappears from the field of view) of such instances. In all cases, the ground truth data, i.e. the initial and terminal frames of the instances, are established manually.

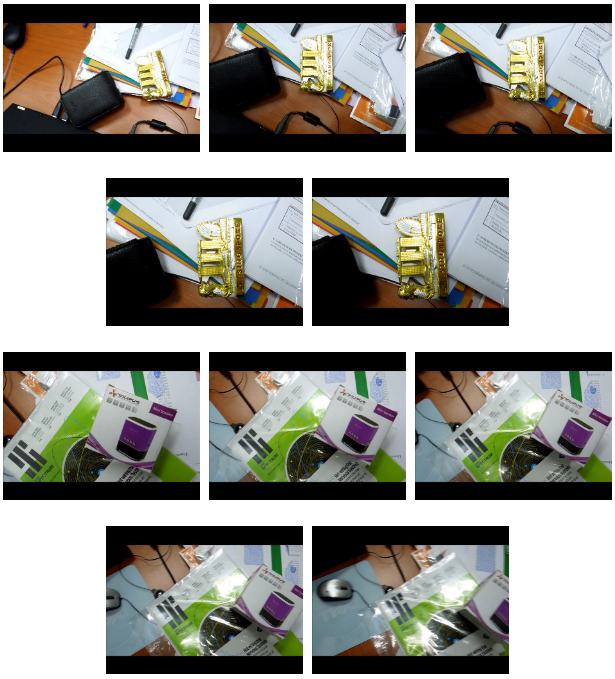


Fig. 4. Examples of frames from a typical initial part (two top rows) and a typical terminal part (two bottom rows) of a ground-truth instance of an object of interest.

The scheme's performances are evaluated by comparing *ground-truth instances* and so-called *active sequences* extracted by the scheme.

Definition 2.

> An *active sequence* is initiated whenever the algorithm identifies a frame matching (according to the specification in Section III) a database image of an object (e.g., Fig. 2A). Then, the active sequence continues until there are at least five consecutive frames which do no match the same object database images.

The value 5 has been established empirically; it corresponds to approx. 0.2sec during which the object may be temporarily invisible (due to sudden flashes of light, temporary camera defocusing, etc.). However, when an active sequence is terminated, it does not mean the object is not visible anymore. Actually, the following cases are possible:

- The object is too close to the camera so that its scale is too large for a match with database images.
- The object becomes too distant (its scale is too small for a match) which means it is no more an object of interest.
- The object actually disappears from the field of view.

Regardless the reason for which an *active sequence* is terminated, the following requirements define a fully reliable object detection scheme:

(a) Each *active sequence* is fully enclosed within a *ground-truth instance*, i.e. non-existing objects are never detected.

(b) Each *ground-truth instance* overlaps at least one *active sequence*, i.e. each genuine instance of an object is detected at least by a single active sequence.

Using the above specifications, we straightforwardly define *precision* (*PA*) of active sequence extraction and *recall* (*RI*) of ground-truth instance detection in a visual search process by a mobile agent as follows:

$$PA = \frac{AS_{(a)}}{AS}, \tag{3}$$

where $AS_{(a)}$ is the number of active sequences satisfying the above Requirement (a), and $AS$ is the total number of extracted active sequences.

$$RI = \frac{GTI_{(b)}}{GTI}, \tag{4}$$

where $GTI_{(b)}$ is the number of ground-truth instances satisfying the above Requirement (b), and $GTI$ is the total number of ground-truth instances.

It was mentioned in Section I that in object detection by visual search *recall* of the individual frame retrieval can be compromised, but *precision* should be as high as possible. However, both *RI recall* and *PA precision* values should be at the highest possible levels to reliably detect object instances.

Performances (based on Eqs 3 and 4) of the scheme for the test dataset of videos and objects are summarized in the top row of Table I. To illustrate advantages of the proposed scheme over the other choices, we include in Table I the results for three alternative scenarios. First, the same vocabulary of 1M words is used but without the scale verification (the second row of Table I). Secondly, a much smaller vocabulary of 64k words is used (with the scale verification) instead of the proposed 1M vocabulary (the third row of Table I). The last scenario included in Table I will be discussed later.

Although each of the three schemes detects all instances of objects visible within the test dataset of videos, there are significant differences in the numbers of extracted active sequences, and (consequently) in the reliability of detection. When scale verification is ignored, or the size of vocabulary is significantly reduced, the number of active sequences grows disproportionally ($5-6$ times in our experiments) and *PA precision* falls dramatically. The explanations are similar for both cases. On one hand, credibility of individual keypoint correspondences is limited (even for larger vocabularies) if no means of geometric verification are used. On the other hand, if the vocabulary is small, the number of keypoint correspondences can be so large that even the scale verification is unable to delete all false positives. As a result, large numbers of false active sequences are extracted from the incoming stream of frames. Even though most of those incorrect active sequences are short ($1-2$ frames) they should not be ignored because there are some cases when the ground-truth instances are represented only by such short active sequences.

Fig. 5 shows examples of incorrect matches (some parts of Fig. 2 are also illustrative) including a rather unusual (since *PA precision* is equal to 98.8%) case of a false positive for matching with 1M words and scale verification.

It should be emphasized that high performances of the proposed low-complexity scheme are achievable for object

TABLE I
PERFORMANCES OF OBJECT DETECTION USING THE PROPOSED SCHEME AND THREE ALTERNATIVE SCHEMES.

| **Scheme** | Ground-truth instances | Active sequences | *RI recall* (Eq. 4) | *PA precision* (Eq. 3) |
|---|---|---|---|---|
| **1M words with scale verification** | 58 | 329 | 100.0% | **98.8%** |
| **1M words without scale verification** | 58 | 1886 | 100.0% | 20.6% |
| **64k words with scale verification** | 58 | 1618 | 100.0% | 15.6% |
| **the method from [16]** | 58 | 117 | 60.3% | 100.0% |

TABLE II
PERFORMANCES OF INDIVIDUAL FRAME RETRIEVAL. IF THE SCALE VERIFICATION IS USED, ONLY THE FRAMES CONTAINING THE OBJECT IN APPROX. THE REFERENCE SCALE ARE CONSIDERED THE GROUND TRUTH.

| **Method** | Ground-truth frames | Retrieved frames | *Recall* | *Precision* |
|---|---|---|---|---|
| **1M words with scale verification** | 5429 | 1342 | 22.1% | 89.4% |
| **1M words without scale verification** | 15479 | 19741 | 60.1% | 47.1% |
| **64k words with scale verification** | 5429 | 18078 | 73.3% | 22.0% |
| **the method from [16]** | 15479 | 7514 | 47.9% | 98.6% |

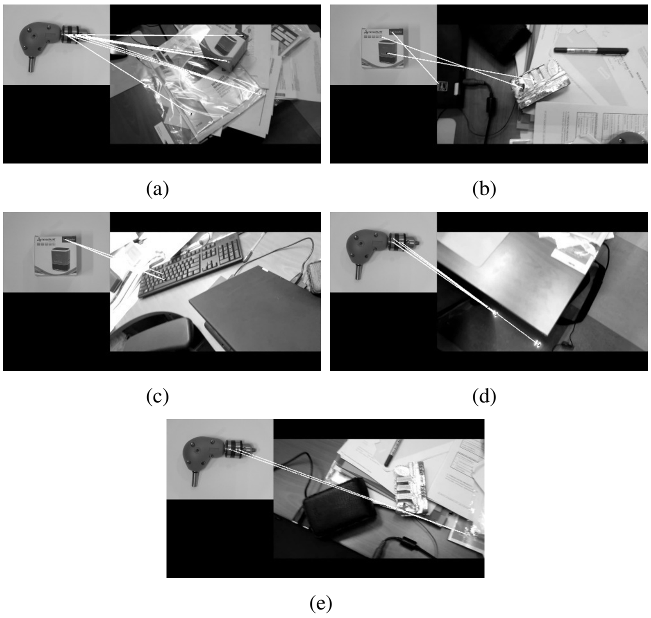

(a) (b) (c) (d) (e)

Fig. 5. Examples of false positive matches. In (a,b) 1M words without scale verification are used, while in (c,d) a vocabulary of 64k words is applied with scale verification. A very unusual case of a false positive for 1M words with scale verification is shown in (e).

detection task only. For a classical CBVIR problem of relevant frame retrieval, i.e. detection of ALL frames partially near-duplicate to the database images of objects of interest, the results are much poorer as shown in Table II presenting performances of various schemes in such a classical *relevant frame retrieval* task (using standard CBVIR definitions of *precision* and *recall* as the scores). Thus, as the last scenario, we included a high-performance (and high-complexity) image matching method proposed in [16] (the original executables of this method have been used). The last row of Table II clearly illustrates superiority of this advanced CBVIR method. The approach proposed in this paper satisfy only the requirement of high *precision*. Therefore, it is not surprising that methods similar to the proposed approach are rather seldom considered for typical CBVIR tasks.

Nevertheless, the algorithm of [16] performs poorer in the problem of object detection. As shown in the last row of Table I, its *RI recall* of instances detection is at unsatisfactory level of only 60%, which is much less than all other schemes presented in this table (which score 100%). Admittedly, it achieves 100% of *PA precision* but (as mentioned earlier) both parameters should be as high as possible in a reliable object detection scheme.

## V. CONCLUDING REMARKS AND FUTURE RECOMMENDATIONS

The paper proposes a CBVIR-based scheme for visual detection of predefined objects of interest in cluttered environments. The scheme seems to be an attractive option for low-cost mobile agents equipped with vision devices. The proposed scheme provides (as preliminarily confirmed by our limited-scale experiments) sufficiently high performances using only low-complexity CBVIR mechanisms (borrowed from a classical CBVIR problem of partial near-duplicate retrieval). Additionally, we assume that the encountered objects become *objects of interest* when seen from a sufficiently short distance. This threshold distance defines the scale at which the database template images of objects should be recorded. Then, a frame containing an object would be recognized as an interesting one if two conditions are met. First, it is similar to a database image of some object of interest (i.e. numerous keypoint correspondences defined by identical visual words from a sufficiently large vocabulary exist) and, secondly, a significant number of those matching keypoint pairs are in approximately the same scale (i.e. the object is approached at approximately the threshold distance).

The second requirement can be considered a limitating factor, especially if the threshold distance (or rather the scale corresponding to this distance) cannot be specified or it fluctuates, e.g. because of the camera zoom. However, in typical modern applications (in robotics in particular) the mobile agents are usually equipped with some kind of range-sensing devices which can provide the agent with the depth data, e.g. [20], [21]. Then, an estimate of the distance to the observed part of the environment can be used to correspondingly modify the reference scale to be applied in the scheme (so that an adaptive reference scale is used). The only modification needed is a minor change in Eqs 1 and 2, which should be rewritten as

$$0.8M\,(K_2) \leq S_F \cdot M(K_1) \leq 1.2M(K_2), \tag{5}$$

$$0.8m\,(K_2) \leq S_F \cdot m(K_1) \leq 1.2m(K_2), \tag{6}$$

where $S_F$ is the scale adaptation factor which can be estimated using the depth data from a range sensor and/or the current camera focus data.

## References

[1] [1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Conf. ICCV 2003,* vol. 2, Nice, 2003. doi: 10.1109/ICCV.2003.1238663 pp. 1470–1477. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2003.1238663

[2] H. Jegou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision,* vol. 87, no. 3, pp. 316–336, 2010. doi: 10.1007/s11263-009-0285-2. [Online]. Available: http://dx.doi.org/10.1007/s11263-009-0285-2

[3] H. Stew´ enius, S. Gunderson, and J. Pilet, "Size matters: Exhaustive geometric verification for image retrieval," in *Proc. ECCV 2012,* vol. II, Florence, 2012. doi: 10.1007/978-3-642-33709-3 48 pp. 674–687. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33709-3 48

[4] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *Image and Vision Computing,* vol. 22, pp. 761–767, 2004. doi: 10.1016/j.imavis.2004.02.006. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2004.02.006

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[6] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. CVPR 2012,* 2012. doi: 10.1109/CVPR.2012.6248018 pp. 2911–2918. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2012.6248018

[7] Y. Ke, R. Sukthankar, and L. Huston, "An efficient parts-based nearduplicate and sub-image retrieval system," in *Proc. ACM Multimedia Conf.,* 2004. doi: 10.1145/1027527.1027729 pp. 869–876. [Online]. Available: http://dx.doi.org/10.1145/1027527.1027729

[8] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *International Journal of Computer Vision,* vol. 65, pp. 43–72, 2005. doi: 10.1007/s11263-005-3848-x. [Online]. Available: http://dx.doi.org/10.1007/s11263-005-3848-x

[9] M. Donoser and H. Bischof, "Efficient maximally stable extremal region (mser) tracking," in *Proc. IEEE Conf. CVPR 2006,* 2006. doi: 10.1109/CVPR.2006.107 pp. 553–560. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.107

[10] F. Kristensen and W. MacLean, "Real-time extraction of maximally stable extremal regions on an fpga," in Proc. IEEE Symp. ISCAS 2007, 2007. doi: 10.1109/ISCAS.2007.378247 pp. 165–168. [Online]. Available: http://dx.doi.org/10.1109/ISCAS.2007.378247

[11] E. Salahat, H. Saleh, A. Sluzek, M. Al-Qutayri, B. Mohammed, and M. Ismail, "Architecture and method for real-time parallel detection and extraction of maximally stable extremal regions (msers)," *U.S. Patent Application* No. 14/482,629, 2014.

[12] J. Jiang, X. Li, and G. Zhang, "Sift hardware implementation for real-time image feature extraction," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 24, no. 7, pp. 1209–1220, 2014. doi: 10.1109/TCSVT.2014.2302535. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2014.2302535

[13] D. Nist´ er and H. Stew´ enius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. CVPR 2006,* vol. 2, 2006. doi: 10.1109/CVPR.2006.264 pp. 2161–2168. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.264

[14] O. Chum and J. Matas, "Matching with prosac - progressive sample consensus," in *Proc. IEEE Conf. CVPR 2005,* San Diego(CA), 2005. doi: 10.1109/CVPR.2005.221 pp. 220–226. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.221

[15] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Computer Vision,* vol. 2, 1999. doi: 10.1109/ICCV.1999.790410 pp. 1150–1157. [Online]. Available: http://dx.doi.org/10.1109/ICCV.1999.790410

[16] M. Paradowski and A. Śluzek, "Local keypoints and global affine geometry: Triangles and ellipses for image fragment matching," in *Innovations in Intelligent Image Analysis,* H. Kwasnicka and L. Jain, Eds. Springer-Verlag, 2011, vol. SCI339, pp. 195–224. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-17934-1 9

[17] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Conf. CVPR 2009,* Miami Beach, 2009. doi: 10.1109/CVPR.2009.5206566 pp. 25–32. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2009.5206566

[18] O. Chum and J. Matas, "Large-scale discovery of spatially related images," *IEEE PAMI,* vol. 32, no. 2, pp. 371–377, 2010. doi: 10.1109/TPAMI.2009.166. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2009.166

[19] M. Islam and A. Śluzek, "Relative scale method to locate an object in cluttered environment," *Image and Vision Computing,* vol. 26, no. 2, pp. 259–274, 2008. doi: 10.1016/j.imavis.2007.06.001. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2007.06.001

[20] K. Shubina and J. Tsotsos, "Visual search for an object in a 3d environment using a mobile robot," *CVIU,* vol. 114, pp. 535–547, 2010. doi: 10.1016/j.cviu.2009.06.010. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2009.06.010

[21] K. Sjo, D. Lopez, C. Paul, P. Jensfelt, and D. Kragic, "Object search and localization for an indoor mobile robot," *J. Computing and Inf.Technology,* vol. CIT 217, no. 1, pp. 67–80, 2009. doi: 10.2498/cit. [Online]. Available: http://http://hrcak.srce.hr/index.php?show=toc&idbroj=3662

# 5th International Workshop on Artificial Intelligence in Medical Applications

THE workshop on Artificial Intelligence in Medical Applications – AIMA'2015 - provides an interdisciplinary forum for researchers and developers to present and discuss latest advances in research work as well as prototyped or fielded systems of applications of Artificial Intelligence in the wide and heterogenious field of medicine, health care and surgery. The workshop covers the whole range of theoretical and practical aspects, technologies and systems based on Artificial Intelligence in the medical domain and aims to bring together specialists for exchanging ideas and promote fruitful discussions.

## Topics

The topics of interest include, but are not limited to:

- Artificial Intelligence Techniques in Health Sciences
- Knowledge Management of Medical Data
- Data Mining and Knowledge Discovery in Medicine
- Health Care Information Systems
- Clinical Information Systems
- Agent Oriented Techniques in Medicine
- Medical Image Processing and Techniques
- Medical Expert Systems
- Diagnoses and Therapy Support Systems
- Biomedical Applications
- Applications of AI in Health Care and Surgery Systems
- Machine Learning-based Medical Systems
- Medical Data- and Knowledge Bases
- Neural Networks in Medicine
- Ontology and Medical Information
- Social Aspects of AI in Medicine
- Medical Signal and Image Processing and Techniques
- Ambient Intelligence and Pervasive Computing in Medicine and Health Care

## Event Chairs

**Lasek, Piotr,** University of Rzeszow, Poland

**Paja, Wiesław,** University of Rzeszów, Poland

**Pancerz, Krzysztof,** University of Management and Administration in Zamość, Poland

## Program Committee

**Basarici, Samsun M.,** (Medical) Image Processing, Yasar University, Turkey

**Deserno, Thomas M.,** Uniklinik RWTH Aachen University, Germany

**Drahansky, Martin,** Brno University of Technology, Czech Republic

**Hashimoto, Hiroshi,** Advanced Institute of Industrial Technology, Japan

**Hassanien, Aboul Ella,** Cairo University, Egypt

**Iantovics, Barna,** Petru Maior University, Romania

**Kountchev, Roumen,** Technical Univerity of Sofia, Bulgaria

**Krawczyk, Bartosz,** Wroclaw University of Technology, Poland

**Kumar, Sajeesh,** University of Tennessee, Health Science Center, United States

**Majernik, Jaroslav,** Pavol Jozef Safarik University in Kosice, Slovakia

**Min, Fan,** Zhangzhou Normal University, China

**Olszewska, Joanna Isabelle,** University of Gloucestershire, United Kingdom

**Sawada, Hideyuki,** Kagawa University, Japan

**Shieh, Jiann-Shing,** Dept. of Mechanical Engineering, Yuan Ze University, Taiwan

**Sirakoulis, Georgios,** Department of Electrical & Computer Engineering, Democritus University of Thrace, Greece

**Strzelecki, Michal,** Lodz University of Technology, Poland

**Wtorek, Jerzy,** Gdańsk University of Technology, Poland

**Wysocki, Marian,** Rzeszow University of Technology, Poland

**Yanushkevich, Svetlana,** University of Calgary, Canada

**Zaitseva, Elena,** University of Zilina, Slovakia

DOI: 10.15439/2015F194
ACSIS, Vol. 5

# Comparison of SVM and k-NN classifiers in the estimation of the state of the arteriovenous fistula problem

Marcin Grochowina
University of Rzeszów
al. Rejtana 16, 35-310 Rzeszów, Poland
Email: gromar@ur.edu.pl

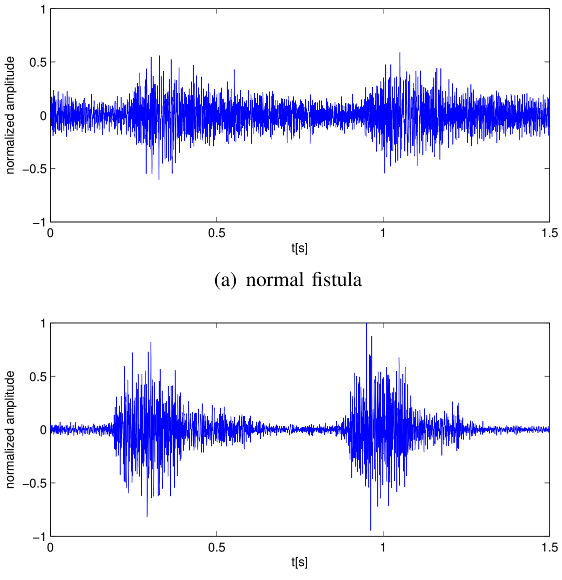
Lucyna Leniowska
University of Rzeszów
al. Rejtana 16, 35-310 Rzeszów, Poland
Email: lleniow@ur.edu.pl

***Abstract*—The paper presents a concise report on the comparison of the classifiers k-NN and SVM in the case of a fuzzy classification of the arterio-venous fistula based on audio recordings. What has been used in the studies are the acoustic signals taken from both healthy patients as well as those diagnosed with the narrowing of a fistula in a mild and major degree of stenosis. In the publication there have been selected two features, each presenting one- time and frequency domain, which enable a quite clear depiction of the classification result. The aim of the study is to develop a solution enabling the detection of fistula's pathologies at an early stage.**

## I. Introduction

THE MAINTENANCE of a properly functioning arterio-venous fistula is extremely important for those who undergo a haemodialysis. Its good condition providing an access to arterial blood with a great degree of flow enables an efficient process of extracorporeal blood filtration. The condition of a fistula is examined by a stethoscope auscultation each time before connecting the patient to a dialysis apparatus. In particular examples diagnostics may be extended to USG examination with the Doppler's facility. Unfortunately, the most abnormalities in functioning of a fistula are diagnosed when they are so advanced that they might be detected just with a use of a stethoscope. The most frequent problem is a deformation of a fistula's vessel which result in a decreasing of its diameter. The narrowing causes a decrease of an arterial blood supply and as a result it could lengthen the process of haemodialysis or even it could make it impossible. That is why it is crucial to determine the methods of an early detection of any pathology within a fistula.

According to the studies conducted, the character of a sound emitted by the blood flowing inside a fistula's vessel vary depending on the state of a fistula. The blood flowing through a fistula in a normal state is continous and it undergoes a very slight changes in a heart rhythm period. As a result a noise of a flowing blood can be heard well all the time. In a stenosed fistula a blood flow is hindered, what effects in rhythm changes of the intensity of a sound, compatible with a heart rhythm. The amplitude of changes is then even bigger and the stenosis of a fistula is more advanced. The differences in the picture of an acoustic signal of a properly working fistula as well as a pathologically changed fistula are presented in the fig.1.



(a) normal fistula

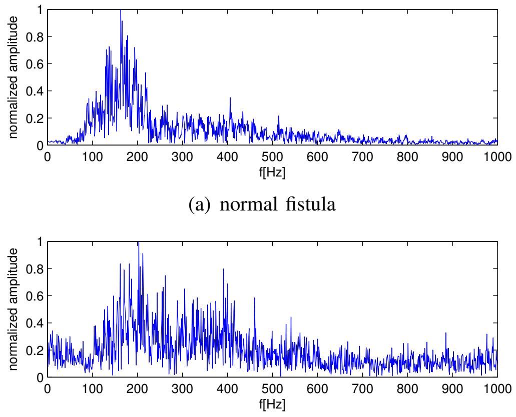
(b) stenosed fistula

Fig. 1. An image of the time-domain audio signal emitted by the arterio-venous fistula

Not all anomalies which may occur within the fistula's area result in a decreasing of the flow. When it comes to the hyperkinetic fistula the blood flow increases excessively, what may result in cardiological complications or even strokes. However, each change of geometry of a fistula causes the change in a character of the flow of the transported blood. Usually, the flow becomes turbulent and causes changes in a frequency spectrum of a signal, what has been show in the fig.2.

Presented examples of trajectories in both the time and the frequency domain represent extreme cases, moreover, the fistula diagnosed as pathological qualifies for a reconstruction through a surgical treatment. Another problem is detecting an anomaly in an early stage, what enables a preventive

978-83-60810-66-8/$25.00©2015, IEEE



(a) normal fistula

(b) stenosed fistula

Fig. 2. The frequency spectrum of the acoustic signal emitted by the arterio-venous fistula

intervention enabling a suppression of unwanted changes. In order to do that what should be determined ,apart from the collections of data from extreme cases, are concentrations of data referring to average cases with the use of which the classifier should be trained enabling their proper assessing.

## II. Materials and Methods

The research material has been collected from 9 patients, all male, age from 35 to 52. Two patients were diagnosed with an extreme malfunction (occlusion) of a fistula, another two had stenosis in various degrees of development, in the case of the last five, the fistula was normal. The examination concerned only end-to-side radio-cephalic fistulas located in a wrist. The acoustic signal was taken from the distance of 5 cm from the anastomosis point. The point of a signal intake is pictured in the fig.3

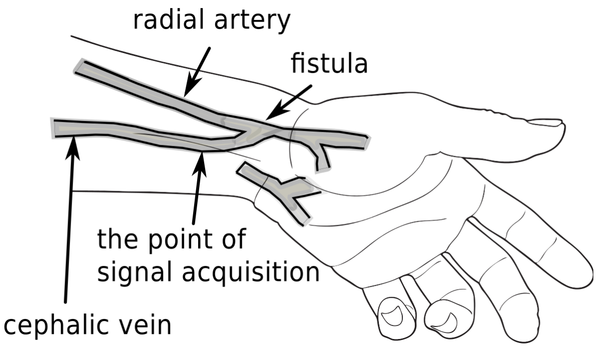


Fig. 3. The point of a signal intake

The material has been taken with the use of a specially created measuring area (fig.4) which includes:

a) A header made from the ABS material which is shaped in a way that it enables the most possible anatomic position of its end during the signal intake. Inside the main body of the header electret microphone Ringford CZ034 has been installed with responsivity -42dB (0dB=1V/Pa, 1KHz, 8mV/Pa) and S/N ratio better than 60dB.
b) A PC computer with Xubuntu Linux 14.10 operating system, equipped with a sound card Sound Blaster PCI 64 constructed according to the Ensoniq AS1371 chip. Dedicated application serving as an registrant of a sound written in C++ with the use of the Qt library, communicating also with the equipment thanks to the ALSA sound server. The sample rate was determined on 8kHz level, with a resolution 16 bits per sample.
c) Headphones used for the organoleptic monitoring of the quality of the recorded material.

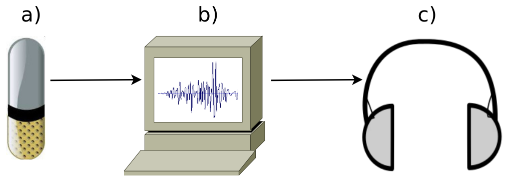


Fig. 4. data acquisition system diagram

From the taken material there have been isolated 100 samples from each group. The material has been randomly divided into a training sets including 60 samples of each kind and the testing set having 40 samples of each kind.

A common problem in classifiers building tasks is a great difference in the amount of positive and negative training patterns. In order to receive a set pattern with a similar number for each of the classes there have been registered more samples for each person with a stenosed or occluded fistula.

Registered material has been divided into parts corresponding to a single heart rhythm period, next, there has been distinguished properties being components of vector's features.

Altogether, there have been determined over 20 properties in a time domain and a frequency domain, and for the visualization needs of this elaboration, two have been chosen, each from one domain-time and frequency.

### A. A property in a time domain.

When it comes to time domain, there has been chosen one property based on a difference between the maximal and the minimal value of the envelope of a signal (fig.5).

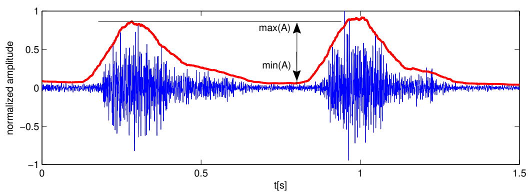


Fig. 5. A property in a time domain

In order to get an envelope signal, it has been straighten and filtrated with the use of a low-pass filter. A moving average filter with the length of 1000 elements has been used.

This property has been marked as $f_1$ and calculated from the equation (1).

$$f_1 = \frac{A_{min}}{A_{max}} \tag{1}$$

The initial analysis has shown that in the case of a fully efficient fistula, the flowing blood generates a sound of a slightly reducing intensity during the whole heart rhythm period. The intensity of a sound emitted by a stenosed fistula definitely characterizes with a greater fluctuation, as a result of a periodical decrease of a flowing blood speed. In connection to that, for an efficient fistula the value of its properties $f_1$ will be higher than in the case of a narrowed fistula.

### B. A property in a frequency domain

In order to get the properties in a frequency domain, a signal had been firstly divided into fragments corresponding to the heart rhythm and then transformed to the frequency domain with using FFT transformation. In order to eliminate a spectral leakage the Hamming time window has been used.. From the spectral module possessed, there have been distinguished two divisions: $FR_1 = [125HZ, 175Hz]$ and $FR_1 = [375Hz; 425Hz]$ (fig.6).

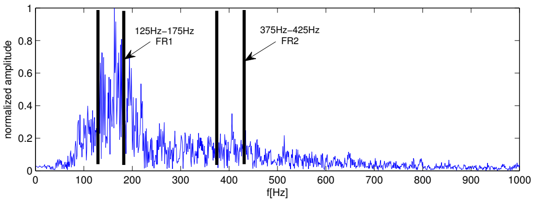


Fig. 6. A property in a frequency domain

The value of a spectral amplitude within these divisions have been summarized and calculated its property $f_2$ according to the equation (2).

$$f_2 = \frac{FR1}{FR2} = \frac{\sum_{f=125}^{175} fft}{\sum_{f=375}^{425} fft} \tag{2}$$

An analysis of signals coming from normal as well as stenosed fistulas, has shown that in the case of those stenosed ones, the value of a $FR2$ factor is definitely higher. Therefore, the value of a $f_2$ property will be smaller when the condition of a fistula will be better.

### C. The use of an SVM classifier in a fuzzy classification task

The classifier used in the elaboration comes from the libSVM library in 3.2 version by Chih-Chung Chang and Chih-Jen Lin. The library is written in the C++ language owns interfaces which enable using with various programming languages, including the "m" language in Matlab and Octave environment, which were used in order to work out the results of this elaboration.

The used classifier in the classifying tasks gives back the value of -1 or 1 indicating the adjunction of an examined feature vector in a particular category. In the regression tasks what is given back is the real value, which may go beyond the $[-1, 1]$ interval.

$$y_{SVM} = \begin{cases} -1 & \text{for } SVM_{out} < -1 \\ SVM_{out} & \text{for } SVM_{out} \in [-1; 1] \\ 1 & \text{for } SVM_{out} > 1 \end{cases} \tag{3}$$

Taking the specifics of the examined problem, it became necessary to limit the set of values given back through the interval classifier $[-1, 1]$. It has been accomplished by an application of a function described as (3) pattern, where $SVM_{out}$ is a regression value calculated through the classifier.

### D. The use of the classifier k-NN in a fuzzy classification task

Normally, the algorithm k-NN is used in classification tasks, what means it gives back a discreet value determining an adjunction to the one of the classes determined beforehand. On the account of the requirements determined by the assumptions of this elaboration, it has become necessary to create an own implementation of the k-NN algorithm in a way that it gives back a real value from the continuum between two extreme cases.

In order to do that, after having found the k-nearest neighbors and calculating their Euclidean distance from the examined point, the transformation described with a equation (4) has been applied, where $d(n)$ is a distance of the umpteenth pattern from the examined point.

$$s(n) = \frac{e^{-d(n)}}{\sum_{i=1}^{k} e^{-d(i)}} \tag{4}$$

In effect, the distance measurement has been changed into the proximity measure $s(n)$, simultaneously receiving

$$\sum_{n=1}^{k} s(n) = 1$$

to calculate definitively an expected degree of belonging of the examined point to the model classes.

$$y_{k-NN} = \sum_{n=1}^{k} s(n)X(n) \tag{5}$$

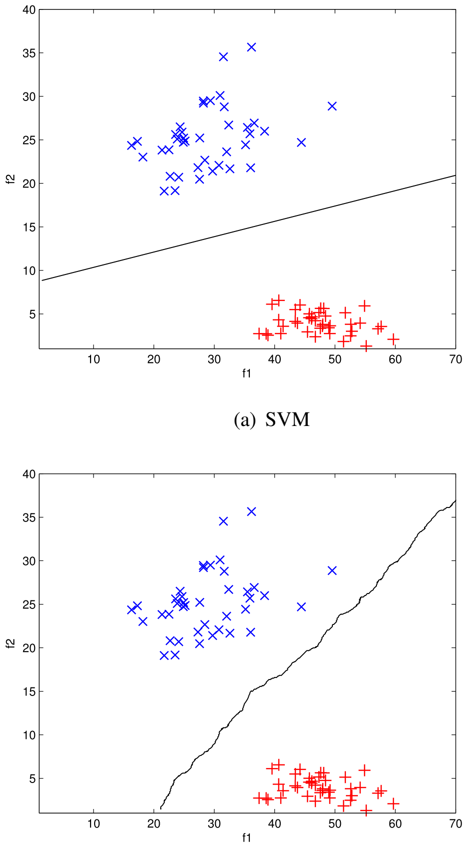
Assuming that all the model vectors are ascribed to the classes from the range $[-1, 1]$, the result of operation of such constructed classifier will be within the same range.

## III. Results and discussion

The SVM and k-NN classifiers have been compared in three cases of using. For each case the accuracy has been determined and the analysis with the use of a confusion matrix has been conducted. In the first example the classifiers have been used in a model classification task, however in two other cases in a fuzzy classification task. For the k-NN classifier, the k=19 parameter has been established and used in all three cases. In the first two cases for the SVM classifier the parameter c=1 has been established and there hasn't been any transformations of the feature's space used. In the third case, because of the lack of a linear separation of the data, for the SVM classifier the parameter C=100 has been established and a kernel transformation with the use of the Gauss's function has been used.

### A. Classification

In this task, the classifiers have been trained with a training set including only the extreme data, namely, ascribed to one of the two sets labeled "-1" or "1". Since the sets of data of both classes are substantially distant from each other and moreover, linearly separated, none of the classifiers had any problems with their proper division, what is presented in the fig.7.



(a) SVM

(b) k-NN

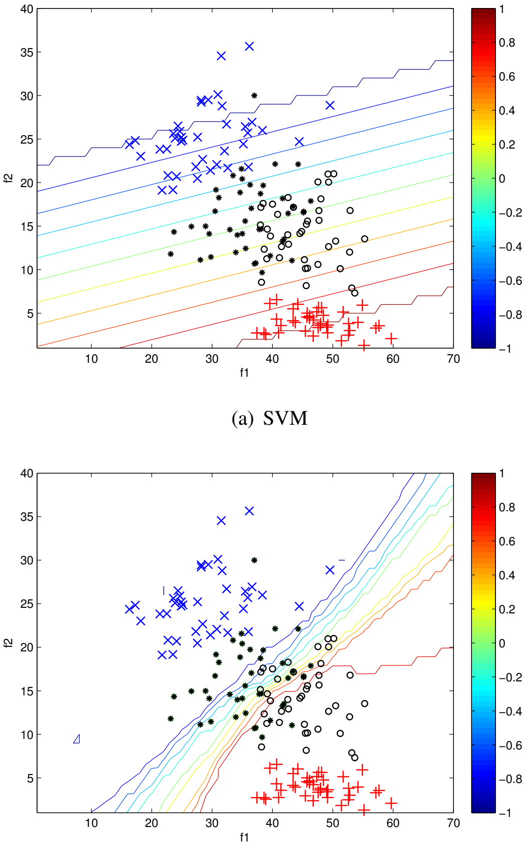
Fig. 7. Classification

In both cases, all the examined points have been accurately recognized in both classes, what gave 100% of accuracy. The classifiers differs only in the course of hyperplane of the division of the classes, what happened to be irrelevant in this particular case.

### B. The fuzzy classification - version 1

In cases of a real assessment of an arterio-venous fistula's condition classifiers cannot determine the adjunction of the examined point to one from the two extreme classes. The task is problematic since it requires determining the position of an examined point in a space between the extreme cases. Therefore, as a result of its operation, a classifier cannot give back the value of -1 or 1. It is necessary to adjust a classifier in a way it could give back the value from the $[-1, 1]$ range. These values indicate a degree of adjunction of an examined point to the extreme classes.

In the first version of a fuzzy classification a teaching set containing only extreme cases has been applied. In effect, the functions mapping the $[f1, f2]$ features' space in the $[-1, 1]$ range has been obtained as presented graphically in the fig.8.



(a) SVM

(b) k-NN

Fig. 8. Fuzzy classification - version 1

In testing, the set containing average cases have been used, marked on a basis of an expert's assessment with the labels -0.33 and 0.33, what gave, in total, four classes of training patterns labeled with: $\{-1, -0.33, 0.33, 1\}$.

In order to make the analysis the values given back by the classifiers have been digitized to four ranges corresponding to four entering classes, and next, compared with the real adjunction of the testing patterns. The assignation of the ranges of the starting classifiers to the appropriate classes is shown in the table.I

TABLE I
ASSIGNATION OF THE RANGES OF THE STARTING CLASSIFIERS TO THE APPROPRIATE CLASSES

| range | class |
|---|---|
| $[-1, -0.66)$ | -1 |
| $[-0.66, 0)$ | -0.33 |
| $[0, 0.66)$ | 0.33 |
| $[0.66, 1])$ | 1 |

The result of the operation of the classifiers has been compared in the table II and table III.

TABLE II
CONFUSION MATRIX - K-NN 0/1

| accuracy:0.59 | | Actual value | | | | |
|---|---|---|---|---|---|---|
| | | -1 | -0.33 | 0.33 | 1 | class precision |
| Predicted | -1 | **40** | 19 | 1 | 0 | 0.67 |
| | -0.33 | 0 | **7** | 3 | 0 | 0.70 |
| | 0.33 | 0 | 7 | **8** | 0 | 0.53 |
| | 1 | 0 | 9 | 28 | **40** | 0.52 |
| | class recall | 1.00 | 0.17 | 0.20 | 1.00 | |

TABLE III
CONFUSION MATRIX - SVM 0/1

| accuracy:0.78 | | Actual value | | | | |
|---|---|---|---|---|---|---|
| | | -1 | -0.33 | 0.33 | 1 | class precision |
| Predicted | -1 | **35** | 1 | 0 | 0 | 0.97 |
| | -0.33 | 5 | **24** | 11 | 0 | 0.60 |
| | 0.33 | 0 | 15 | **26** | 0 | 0.63 |
| | 1 | 0 | 9 | 3 | **40** | 0.93 |
| | class recall | 0.88 | 0.60 | 0.65 | 1.00 | |

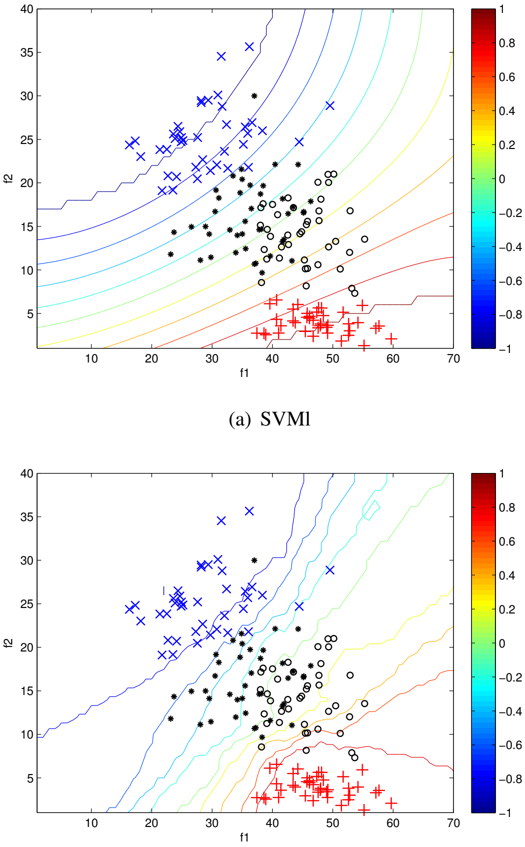
The analysis of the confusion matrix shows definitely better parameters of the SVM classifier. Although in the case of the extreme classes labeled with -1 and 1 recall for the classifier k-NN equal 100% and it is better than for SVM, for the average classes, the values between ten and twenty percent disqualify the k-NN classifier completely. However, the level of accuracy of 80% obtained for the SVM classifier is also not very satisfying, all the more, for the classes labeled negative it has a tendency to make a better diagnosis.

### *C. The fuzzy classification - version 2*

In order to improve the classifiers' parameters, the next step was to insert the vectors from the sets labeled -0.33 and 0.33 into the training sets. In effect, the functions mapping the features' space $[f1, f2]$ into range $[-1, 1]$ has undergone a significant modification what has been pictured in the fig.9.

The modification improved the quality of operating of both of the classifiers. The comparison of the results of the classifiers' testing is presented in the table IV and table V.

For the SVM classifiers the improvement is slight. The accuracy increased from 78% to 81%, similarly, the class precision and the class recall parameters have undergone in some cases a slight correction down.



(a) SVMl

(b) k-NN

Fig. 9. Fuzzy classification - version 2

TABLE IV
CONFUSION MATRIX - KNN FUZZY

| accuracy:0.85 | | Actual value | | | | |
|---|---|---|---|---|---|---|
| | | -1 | -0.33 | 0.33 | 1 | class precision |
| Predicted | -1 | **37** | 2 | 0 | 0 | 0.95 |
| | -0.33 | 3 | **29** | 6 | 0 | 0.76 |
| | 0.33 | 0 | 9 | **30** | 0 | 0.77 |
| | 1 | 0 | 0 | 4 | **40** | 0.91 |
| | class recall | 0.93 | 0.72 | 0.72 | 1.00 | |

The results of the operation of the k-NN classifier have undergone a significant improvement. The accuracy increased from 59% to 85% and it exceeded the value obtained by the SVM classifier. Only class recall for the "-1" class has slightly reduced.

In both cases what can be observed is unfortunately the problem of an accurate classification of the average classes.

## IV. CONCLUSION

As a result of the conducted tests, what has been tentatively stated is that both the SVM and the k-NN classifiers can be applied in a task of an estimation of a fistula's condition and enable obtaining a relatively high co-factor of accuracy. What provides proper operation of the examined classifiers is putting into a set of pattern data not only extreme patterns determining the limits of a case's space but also, when it is possible, the whole continuum. On the basis of obtained results it cannot

TABLE V
CONFUSION MATRIX - SVM FUZZY

| accuracy:0.81 | | Actual value | | | | |
|---|---|---|---|---|---|---|
| | | -1 | -0.33 | 0.33 | 1 | class precision |
| Predicted | -1 | **37** | 1 | 0 | 0 | 0.97 |
| | -0.33 | 3 | **27** | 12 | 0 | 0.64 |
| | 0.33 | 0 | 12 | **24** | 0 | 0.67 |
| | 1 | 0 | 0 | 4 | **40** | 0.91 |
| | class recall | 0.93 | 0.68 | 0.60 | 1.00 | |

be stated which of the applied classifiers will be the best for the task. The differences in the accuracy of the classification are slight and in most cases they are within the limits of the statistical error.

In order to obtain more reliable results in continuing of the examinations, the realization of a few conclusions seems to be essential:

- it is necessary to collect a significantly greater number of an overall data since a small amount of data used in testing does not guarantee a solution of the problem,
- the signal samples have to be taken from the greater number of patients,
- it is recommended to obtain the possible biggest amount of data from the fistula's in an average state between the fully efficient to the occluded ones,
- it should be considered to widen or to change the set of the features' signal included in the analysis of the features' vector.

The application of the above-mentioned conclusions gives a great chance for the fruitful continuation of the examinations and in effect obtaining a solution enabling identification of pathological states within a fistula in the earliest state possible.

## REFERENCES

[1] Marcin Grochowina, Lucyna Leniowska, Piotr Dulkiewicz, "Application of Artificial Neural Networks for the Diagnosis of the Condition of the Arterio-venous Fistula on the Basis of Acoustic Signals," *Brain Informatics and Health, Lecture Notes in Computer Science Volume 8609,* Springer, 2014, pp 400-411.

[2] Marcin Grochowina, Lucyna Leniowska, "Analiza parametrów akustycznych prototypu głowicy do akwizycji sygnału z przetoki tętniczo-żylnej," *Mechanika w Medycynie,* Uniwersytet Rzeszowski, 2014, pp 63-72.

[3] Mikkel Grama , Jens Tranholm Olesena , Hans Christian Riisa , Maiuri Selvaratnama and Michalina Urbaniaka, "Stenosis detection algorithm for screening of arteriovenous fistulae," *15th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC 2011),* Springer, 2011, pp. 241–244.

[4] Fan, Rong-En and Chen, Pai-Hsuen and Lin, Chih-Jen, "Working set selection using second order information for training support vector machines," *The Journal of Machine Learning Research vol.6,* JMLR. org, 2005, pp. 1989–1918.

[5] Vesquez, PO and Marco, MM and Mandersson, Bengt, "Arteriovenous fistula stenosis detection using wavelets and support vector machines," *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE,* IEEE, 2009, pp. 1298–1301.

[6] Castillo, Oscar and Melin, Patricia and Ramírez, Eduardo and Soria, José, "Hybrid intelligent system for cardiac arrhythmia classification with Fuzzy K-Nearest Neighbors and neural networks combined with a fuzzy system," *Expert Systems with Applications vol.39,* Elsevier, 2012, pp. 2947–2955.

[7] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM – A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[8] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin "A Practical Guide to Support Vector Classification," Department of Computer Science National Taiwan University, Taipei 106, Taiwan.

[9] Tadeusz Morzy, "Eksploracja danych - metody i algorytmy," PWN, 2013.

[10] Pawel Delimata, Zbigniew Suraj , "Reducts Evaluation Methods Using Lazy Algorithms," *Rough Sets and Knowledge Technology, 4th International Conference, RSKT 2009, Gold Coast, Australia, July 14-16, 2009. Proceedings,* Springer, 2009, pp. 120–127.

DOI: 10.15439/2015F318
ACSIS, Vol. 5

# Detection of Breast Abnormalities of Thermograms based on a New Segmentation Method

Mona A. S. Ali[4,*], Gehad Ismail Sayed[5,*], Tarek Gaber[1,2,*],
Aboul Ella Hassanien[5,*], Vaclav Snasel[3], Lincoln F. Silva[6]
[1]*Faculty of Computers and Informatics, Suez Canal University, Egypt*
[2]*IT4Innovations, VSB-TU of Ostrava, Czech Republic*
[4]*Faculty of Computers and Information,Minia University, Egypt*
[5]*Faculty of Computers and Information, Cairo University, Egypt*
[3]*FEECS, Department of Computer Science and IT4Innovations, VSB-TU of Ostrava, Czech Republic*
[6]*Department of Computer Science, Fluminense Federal University, Brazil*
[*]*Scientific Research Group in Egypt, (SRGE), http://www.egyptscience.net*

*Abstract*—**Breast cancer is one from various diseases that has got great attention in the last decades. This due to the number of women who died because of this disease. Segmentation is always an important step in developing a CAD system. This paper proposed an automatic segmentation method for the Region of Interest (ROI) from breast thermograms. This method is based on the data acquisition protocol parameter (the distance from the patient to the camera) and the image statistics of DMR-IR database. To evaluated the results of this method, an approach for the detection of breast abnormalities of thermograms was also proposed. Statistical and texture features from the segmented ROI were extracted and the SVM with its kernel function was used to detect the normal and abnormal breasts based on these features. The experimental results, using the benchmark database, DMR-IR, shown that the classification accuracy reached (100%). Also, using the measurements of the recall and the precision, the classification results reached 100%. This means that the proposed segmentation method is a promising technique for extracting the ROI of breast thermograms.**

## I. Introduction

BREAST cancer is the most common one among women, and it most likely cause of women death in the world wide. In the United States, one death of 4 is due to cancer [1]. There are many breast imaging techniques used for identifying early stage of breast cancer. One of them is mammography. It is most commonly used for screening of breast cancer. However the false negative rate can reach up to 30% in addition it expose patient to ionizing radiation effect [2]. Thermography uses a special heat-sensing camera to measure and map the heat on the surface of the breast [3], [4].

As long as cancerous tumors have an increased blood supply and cell growth they tend to be warmer than the surrounding normal tissue and it is the idea behind using Thermal images instead mammogram. Thermography play an important rule as a visualization technique that can be used in many different fields of physics and science. Unlike mammography, there is no compression of the breast so women may find it more comfortable, thermal imaging does not expose the woman to any radiation, as occurs with mammography (a type of X-ray), fast, low cost and sensitive method [5].

The extraction of the Region Of Interest, ROI segmentation, from the breast image is very important for detecting the cancer from the breast images. ROI segmentation aims to separate the regions of the breast from the other parts of the body. It can range from a completely manual to a fully automatic process [6]. Once the breast region is separated from the rest of thermal images, then some features are extracted. Then some kindS of artificial classifications algorithms are applied to classify the organs in analysis as normal or abnormal [7].

For the automatic segmentation approaches [7], [8], the level set technique [9] used to extract the blood vessels in a thermal image. The level set function is evolved using the gradient magnitude and direction of an edge map provided by few initial points selected in region of interest. In [7], automatic segmentation approach is proposed using active contour and level set method without re-initialization. It is used to extract the breast regions from breast thermograms. Before applying the level set, a statistical based noise removal and contrast-limited-adaptive histogram equalization are used to improve signal to noise ratio and the contrast of thermal images. Verification and validation of the segmented results are carried out using 60 images against the ground truths. The segmented areas are observed to be in good correlation with the ground truth areas as the correlation coefficient was 98%.

For breast abnormalities, several attempts [10], [11], [12], [13], [14] have been made to demonstrate the ability of different features in detecting breast cancer abnormalities. Texture features have been used to detect abnormal thermograms using support vector machine (SVM) [11] and artificial neural networks [12]. Statistical features ranging from first order statistical to higher order statistical features have been extracted for automatic classification of abnormal breast conditions using different classifiers like SVM [13], linear discriminant classifier, minimum distance classifier and Parzen window [14].

This paper proposes an automatic breast segmentation approach based on the distance between body and camera is 1 meter (dynamic protocol). In order to evaluate the accuracy of the proposed segmentation approach, different selective

978-83-60810-66-8/$25.00©2015, IEEE

features were extracted from the segmented regions and then the SVM classifier was used to detect the breast abnormalities

This paper is organized as follows. Section II reviews the existed segmentation and the extracted features methods. Section III gives overview of the thermal imaging protocols of breast cancer while Section IV presents the proposed approach to segment region of interest. In Section V the proposed segmentation is used to detect the abnormalities of thermogarms images and IN Section VII the experimental results are discussed. Finally, we conclude the paper and gives some future directions in Section VIII.

## II. Literature Review

To make our proposed approach, in this paper comparable with its related work, we have limited this related work to the efforts done using the DMR-IR database [15]. These efforts can be classified into two classes: automatic segmentation of breast regions [7]] [8] and classification based on the asymmetry analysis to normal and abnormal cases [16], [17], [18].

For the automatic segmentation [7], [8], the level set technique [9] has been used to extract the blood vessels in a thermal image. The Level set function was evolved using the gradient magnitude and direction of an edge map provided by few initial points selected in region of interest. In [7], an automatic segmentation approach, using active contour and level set method without re-initialization, was proposed to extract the breast regions from breast thermograms. Before applying the level set, a statistical based noise removal technique and contrast limited adaptive histogram equalization were used to improve signal to noise ratio and contrast of thermal images. Verification and validation of the segmented results were carried out using 60 images against the ground truths. The segmented areas were observed to be in good correlation with the ground truth areas as the correlation coefficient was 98

Another automatic segmentation approach has been proposed in [8] to segment the frontal breast tissues from breast thermograms. This approach made use of the Modified Phase Based Distance Regularized Level Set (MPBDRLS) method. The method was further modified by adopting an improved diffusion rate model. The segmented region of interests was evaluated using 72 gray scale images of size 320 X 240 pixels and against the ground truth images. The overlap measures showed that the average similarity between four sets ground truths and segmented region of interests was 97%.

The asymmetric-based classification is based on the asymmetric abnormalities which can be identified by comparing the features extracted from the breast regions (right and left). Several statistical and fractal features are found to be useful features in identification of pathological conditions of breast tissues [18]. Using DMR-IR database, in [16], an approach was proposed to classify the normal and abnormal (carcinoma, nodule and fibro adenoma) breast thermograms Gabor wavelet transform. First, the segmentation of the breast tissues was performed using ground truth masks and the raw images. Gabor features [19] were then extracted for the detection of the abnormalities. The results showed that from total of 20 images, used of the approach evaluation, there were 9 images with carcinomas, 6 with nodules, and 5 with fibro adenomas.

## III. INFRARED IMAGING PROTOCOLS

Infra-red imaging protocols in general have the following parts: Recommendations to the patient, conditions of the examination room; preparation, cooling and positioning of the patient; capturing positions and parameters. For recommendation to the patient, patient asked to avoid: Alcohol, caffeine, physical exercises, and nicotine for at least two hours before the examination process. And for condition of examination room, Room temperature must be maintained between 20°C and 22 °C, no airflow directed to the patient no windows, no opening.

For preparation of the patient, she is asked to remove earrings, necklaces or other accessory. The patient's body temperature is checked by clinical thermometer and her hair stuck with a cap. The patient is positioned in front of the camera with the hands on the head. And for capturing positioning and parameters, the standard distance between the camera and the patient is 1 $m$. But it depend on the size of the patient, a distance of 0.8m or 1.2m is adopted for a better frame of the region of interest (breast and armpit) in the image. The distance, room temperature and the relative humidity of the air are recorded and inserted as parameters in the IR camera settings [20].

In all reproduced protocols, the initial procedures are equal to each volunteer. These procedures are i) to check the central temperature by a clinical thermometer; ii) to check whether the volunteer followed the recommendations of the protocol which are at least two hours before: iii) to ask the patient to remove beads, earrings and other accessories which can be viewed in the thermal images and hold the hair with a shower cap; and iv) to ask the patient to remove clothing from the waist up.

In all protocols below the volunteer stays with his hands on his head during the whole process.

1) First Static Protocol: The volunteer rests for 10 minutes to stabilize the skin temperature of the breasts and armpit. After that, a frontal image is captured.
2) Second Static Protocol: Similar procedures of the previous protocol, but in this, the volunteer rests per 15 minutes.
3) First Dynamic Protocol: An electric fan is turned on and directed to the volunteer's breasts and armpits per 2 minutes. After this, six images are captured, one each minute.
4) Second Dynamic Protocol: Alcohol is applied at the region of the volunteer's breasts and armpits, then an electric fan is turned on and guided to this region per 30 seconds. After this, six images are captured, one each minute.

## IV. The New Segmentation Method

The proposed segmentation method depends on the predefined parameter (i.e. th distance between camera and patient) of the data acquisition protocol used to collect thermal breast image in [20]. In our case, the breast's region occupies nearly half of the image height while the other areas include women shoulders and stomach. Figure 1 confirm these facts.

The main idea of the proposed segmentation method is based on the following facts [20]:

1) The distance between body and camera is 1 meter (dynamic protocol).
2) As can be seen from Figure 1, the image includes only the upper part of the patient that contains a part of the stomach and arms with nick.
3) The breast engaged a specific position in human body which is nearly at the center of the image, see Figure 1.

As a result of these facts, we followed the algorithm as see in Algorithm (1).

**Algorithm 1** Segmentation Method

```
Read original grayscale thermal image, $I$
Read M = I's height
Read N = I's width
Read the corrdinates $Y_1$ and $Y_2$ where $Y_1 = 1/4 * M$ and $Y_2 = M - 0.2 * M$
Extract the ROI where $ROI = imcrop(I, [X_1, X_2, Y_1, Y_2])$, where $X_1 = 0$ and $X_2 = N$
Convert the ROI to binary image by using threshold with value equal to 0.4 (trial and error) to differentiate body from background
Remove columns from the image width having $value = 0$
```

In case, as a women body is not always at center of image then after getting initial ROI extracted, threshold will be used to binaries the image. Then full column with zero value will be removed in order to focus on breast region. Figures 3 and 2 show the extracted ROI from original grayscale thermal image in case of normal and abnormal.

To explain how this new segmentation method works, we give the following example. Suppose, there is an image with 480 height and 640 width as seen in Figure 4. Based on Algorithm (1), the coordinate of the ROI are $X$ where $X$ ranges form 1 to 640 and $Y$ where $Y$ from 120 to 384.

## V. Breast Abnormalities Detection Approach

To evaluate the new segmentation method, we proposed a breast abnormalities detection approach. As shown in Figure 5, this approach consists of four phases: ROI segmentation, ROI enhancement, feature extraction and classification.

*a) ROI Segmentation:* : At this phase, our method, introduced in Section (IV) will be applied to extract the ROI (i.e. the breast region) from original grayscale thermal image.

*b) ROI Enhancement:* Image enhancement is important step in image processing. It used to bring out detail that is obscured, or simply to highlight certain features of interest in an image. Then histogram equalization is used to enhance contrast of the image and increase classification accuracy.

*c) Features Extraction:* Two types of features (first order statistical and texture) are extracted from the enhanced ROI of breast thermograms. The first order statistical features include *mean, standard deviation, median, mode, skewness and kurtosis* whereas the texture ones consists 15 features which were extracted from gray-level co-occurrence matrix (GLCM) [21] with the distance parameter $d = 1$. These features are *Energy, Contrast, Correlation, Sum of square variance, Homogeneity, Dissimilarity, Inverse difference moment (IDM), Inverse difference normalized (INN), Information measure of correlation1 (IMC1), Information measure of correlation2 (IMC2), Difference entropy, Difference variance, Sum entropy, Sum variance and Sum average*. So, the total extracted features is 21 features.

Table I and Table II show the features extracted from the ROI of both normal an abnormal cases. To increase the performance of the proposed approach, before using these features in the classification phase, they were analyzed using student's $t$ hypothesis testing the mean of two independent samples. The features with $P > 0.05$ were selected for the classification process. Based on this test, as seen Table I, the Median, Mode and Standard deviation are statistically insignificant where $P > 0.05$ whereas the Mean, Skewness and Kurtosis are found to be statistically significant where$P < 0.05$. Also, form Table II Homogeneity and Information Measure of Correlation1 (IMC1) are found insignificant features and the Energy, Contrast, Correlation, Sum of square variance, Dissimilarity, Inverse difference moment (IDM), Inverse difference normalized (INN), Information measure of correlation2 (IMC2), Difference entropy, Difference variance, Sum entropy, Sum variance and Sum average were found highly significant features where $P < 0.05$. Therefore, only 16 out of 21 features were used in the classification phase.

TABLE I: First order statistical features analysis for normal group (NG) and abnormal (AbG) group

| Feature | Average of NG | Average of AbG | p |
|---|---|---|---|
| Mean | 105.622 | 123.905 | 0.023 |
| Median | 110.485 | 126.742 | 0.061 |
| Mode | 59.885 | 68.571 | 0.581 |
| Standard Deviation | 14.123 | 16.169 | 0.089 |
| Skewness | 0.8105 | 1.8128 | 0 |
| Kurtosis | 15.245 | 31.415 | 0 |

*d) Classification:* The Support vector machine (SVM) were used to evaluate the feature extracted from the ROI. The SVM is a supervised learning method that transforms input data to high-dimensional feature space using various kernel functions (Linear, polynomial, RBF, and quadratic) such that the transformed data becomes more separable that that of the original input data [22][23][24]. The SVM was first trained with the selected 16 features (i.e. features with $p < 0.05$) to

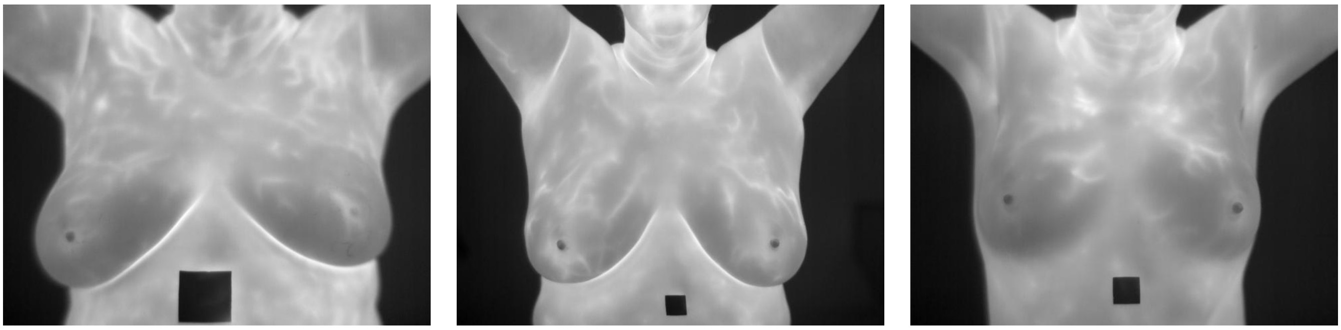


Fig. 1: Original Patient Image

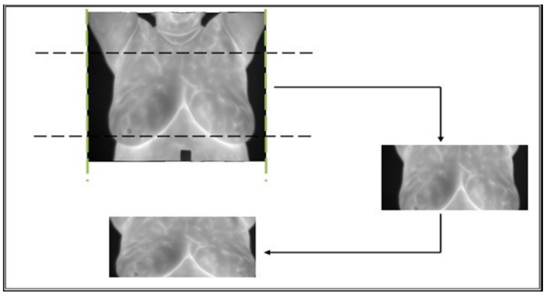


Fig. 2: Abnormal Segmentation Case

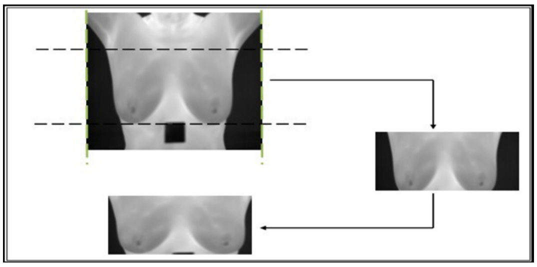


Fig. 3: Normal Segmentation Case

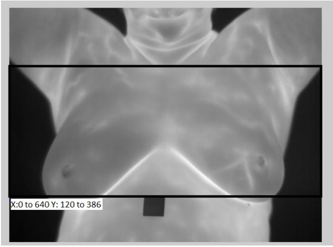


Fig. 4: Segmented ROI

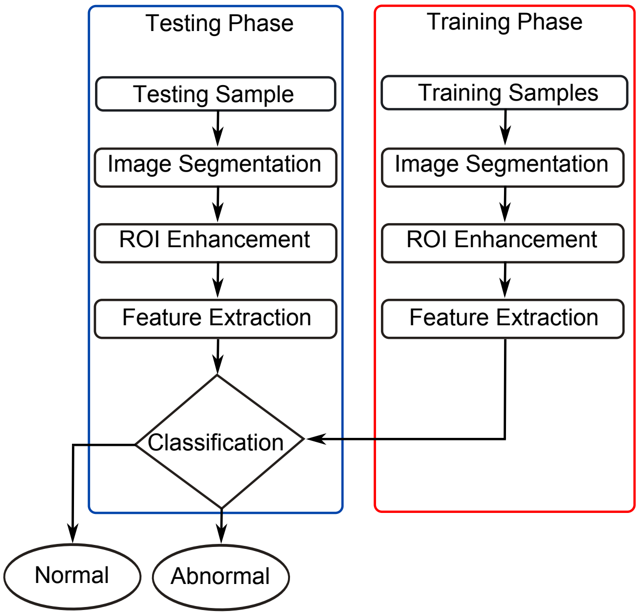


Fig. 5: Proposed Approach Model

build a classification model which was then used in the testing phase to classify breast image to normal or abnormal.

*e) Support Vector Machine (SVM):* SVM is one of the classifiers, which deals linearly or non-linearly to classify unknown objects based on increase the margin between classes. It deals successfully with high dimensional datasets ([19], [25]).

Given a training dataset or a feature matrix, $x_i$ and $y_i$, where $i = 1, 2, 3, \ldots, N$, $N$ represents the number of training samples, $x_i$ is a feature vector, and $y_i$ represent the class labels of the training sets. In case of binary classification $y_i \in \{-1, +1\}$ is the target label, $y = +1$, for samples belong to class $C_1$ and $y = -1$ denotes to samples belong to class $C_2$. The default SVM deals linearly with the classification problem, but using different kernels it may solve the problem of non-linear classification [19]. SVM algorithm tries to find an optimal hyperplane with the maximal margin to separate two different classes, which requires solving the optimization problem in Equation (1).

$$\begin{gathered} maximize \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j . K(x_i, x_j) \\ \text{subject to: } \sum_{i=1}^{n} \alpha_i y_i, 0 \leq \alpha_i \leq C \end{gathered} \tag{1}$$

where, $\alpha_i$ represents the weight assigned to the training sample $x_i$ (if $\alpha_i > 0$, then $x_i$ is called a support vector); $C$ is a

TABLE II: Second order statistical features analysis for normal group (NG) and abnormal (AbG) group

| Feature | Average of NG | Average of AbG | p |
|---|---|---|---|
| Energy | 0.0025 | 0.0027 | 0.044 |
| Contrast | 0.0369 | 0.047 | 0.013 |
| Correlation | 0.7951 | 0.9288 | 0.028 |
| Sum of Square variance | 18496 | 21703 | 0.024 |
| Homogeneity | 0.2828 | 0.3108 | 0.236 |
| Dissimilarity | 9.2618 | 11.448 | 0.004 |
| IDM | 0.8264 | 0.9652 | 0.024 |
| INN | 0.7967 | 0.9322 | 0.0255 |
| IMC2 | 0.8169 | 0.9681 | 0.0244 |
| IMC1 | -0.373 | -0.4241 | 0.0971 |
| Difference Entropy | 1.969 | 2.3652 | 0.0078 |
| Difference Variance | 369.02 | 470.77 | 0.0013 |
| Sum Entropy | 4.15 | 4.8719 | 0.0215 |
| Sum Variance | 7147.6 | 8377.8 | 0.0232 |
| Sum Average | 214.7 | 251.8 | 0.0226 |

regulation parameter; and $K$ is a kernel function, which is used to map or transform the features into higher dimensional space to discriminate between different samples

## VI. Experimental Results

### A. Breast cancer Dataset

A benchmark database [20] used to evaluate our proposed approach. This public database are constructed by collecting the IR images from UFF University's Hospital and publicly published under the approval of the ethical committee where every patient should sign consent. 63 IR single breast images (640 × 480 pixels) from this database were used in this paper (29 healthy and 34 malignant).

### B. Experimental Results

This section presents two type of the results: segmentation results obtained by our new segmentation method and the classification results based on this new method.

*a) Segmentation Results:* Due to the fact that there are different size of the breast of each woman and this is the case of the DMR-IR database [20], four scenarios ( small size breast, medium size breast, large size breast, and the asymmetric breast size) were designed to evaluate our new segmentation method against these different cases. The segmentation results of all these scenarios using our new method are shown in Figure 6 which shows that automatic segmentation of the ROI of thermograms under the various image cases.

*1) Classification Results:* In order to test the efficiency of segmentation results obtained by our new method, two main scenarios were designed. The first scenario was planned to understand the effect of using different number of images for the training and testing on the accuracy of the classification when using statistical and texture features. Different combinations between the number of training and testing images are illustrated in Table III.

TABLE III: Various scenarios for evaluating the proposed approach at different training and testing thermograms

| Scenarios | Training Data | | Testing Data | |
|---|---|---|---|---|
| | *Normal* | *Abnormal* | *Normal* | *Abnormal* |
| **1st Scenario** | 19 | 24 | 10 | 10 |
| **2nd Scenario** | 14 | 19 | 15 | 15 |
| **3rd Scenario** | 8 | 15 | 21 | 19 |
| **4th Scenario** | 4 | 9 | 25 | 25 |

All the scenarios are evaluated by the SVM functions, i.e. Linear, polynomial, RBF, and quadratic, and the summary of their results are given Table (IV), and Table (V).

The second scenario was designed to investigate whether the integration between the features of texture and statistical is better for detecting the abnormalities in thermograms. Under the same sub-scenarios illustrated in Table (III), the results, evaluated by the accuracy, are shown in Figure (7. The accuracy is calumniated according to Equation (2),

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{2}$$

where $T_P$ represents true positive, $T_N$ represents true negative, $F_P$ represents false positive and $F_N$ represents false negative.

## VII. Discussion

As shown in Figure 6, it can be seen that the new segmentation method can successfully extract the ROI (the breasts area) of thermograms. We have some cases where a small part at the bottom of the breast is not included in the extract ROI. However, this part would not affect on the diagnosis of breast abnormality. As reported in [26] that the pectoral region is the most important part for brest cancer diagnosis and it is reported that near 50% of the breast cancer is located in this region which is successfully extracted by our new segmentation method.

The results of this segmentation method was further evaluated by classifying the segmented ROI to normal and abnormal and the results of different scenario are shown in Tables (IV and V) and Figure (7). From these results, it can be seen that the combination between the statistical and texture features of the ROI gave 100 % which is much better than using each of them individually.

$$Recall = \frac{T_P}{T_P + F_N} \tag{3}$$

$$Precision = \frac{T_P}{T_P + F_P} \tag{4}$$

The results of the feature combination were further evaluated in terms of the recall and precision described in equation (3) and (4) and the results of these recall and precision are summarized in Figure (8) and Figure (9), respectively. From these two tables and Table (7), it can be noticed that (a) the SVM-RBF gave the best results (100%) (b) the more training images were used, the high accuracy was obtained as in the first scenario where the accuracy, the recall and the precision reached 100%. These excellent classification results based on

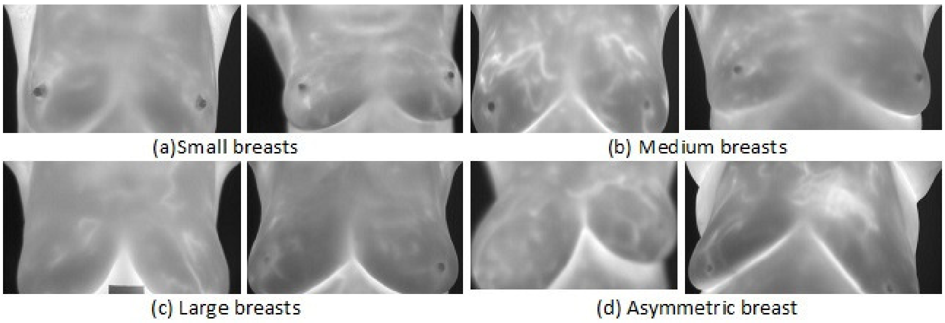

(a)Small breasts
(b) Medium breasts
(c) Large breasts
(d) Asymmetric breast

Fig. 6: Different cases Segmentation Results

TABLE IV: Accuracy of the Results of the First Order Statistical Features

| | *First Scenario* | *Second Scenario* | *Third Scenario* | *Fourth Scenario* |
|---|---|---|---|---|
| **Quadratic** | 85 | 83.33 | 70 | 60 |
| **Polynomial** | 80 | 66.67 | 70 | 56 |
| **RBF** | 55 | 60 | 57.14 | 58 |
| **Linear** | 85 | 76.67 | 72.5 | 58 |

TABLE V: Accuracy of the Results of Texture Features

| | *First Scenario* | *Second Scenario* | *Third Scenario* | *Fourth Scenario* |
|---|---|---|---|---|
| **Quadratic** | 80 | 60 | 62.5 | 34 |
| **Polynomial** | 80 | 56.67 | 47.5 | 50 |
| **RBF** | 70 | 80 | 65 | 52 |
| **Linear** | 65 | 56.67 | 62.5 | 50 |

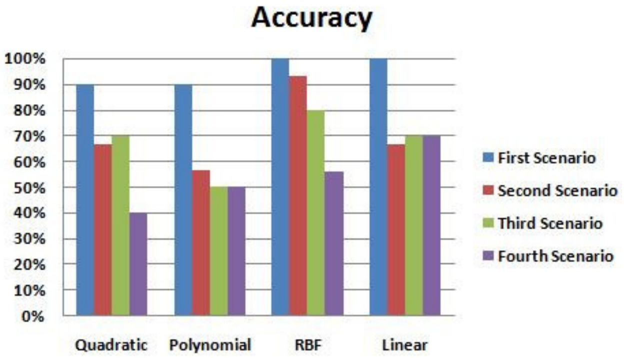


Fig. 7: Accuracy performance measurements

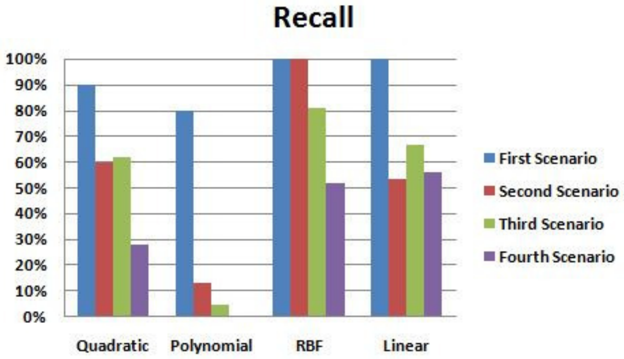


Fig. 8: Recall performance measurements

our new segmentation method means that this method could be used for the automatic segmentation of the thermal breast cancer images.

## VIII. Conclusion and future research directions

In this paper, an automatic segmentation methods for themograms have been proposed. This method is based on the data acquisition protocol parameter (the distance from the patient to the camera) and the image statistics. The proposed segmentation results prove its reliability in extracting the ROI for different cases. This method was evaluated using the segmented ROI in a proposed approach for the detection of the abnormalities of breast thermograms. This approach made use of statistical and texture features from the segmented ROI and the SVM with its kernel function was used to detect the normal and abnormal breasts. Based the experimental results, it was found that the SVM-RBF gave the best results (100%). Also, using the measurements of the recall and the precision, the classification results reached to 100%. These excellent classification results, based on our new segmentation method, concludes that our segmentation method could be used for the automatic segmentation of the thermal breast cancer images. In the future, we plan to increase the dataset used in order to test the reliability of the proposed segmentation method and the classification approach.

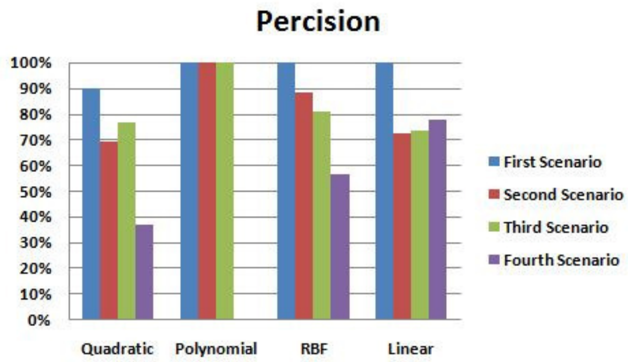


Fig. 9: Percision performance measurements

## IX. Acknowledgment

This paper has been elaborated in the framework of the project "New creative teams in priorities of scientific research", reg. no. CZ.1.07/2.3.00/30.0055, supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic and supported by the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070), funded by the European Regional Development Fund and the national budget of the Czech Republic via the Research and Development for Innovations Operational Programme and by Project SP2015/146 "Parallel processing of Big data 2" of the Student Grant System, VSB- Technical University of Ostrava

## References

[1] R. Siegel, J. Ma, Z. Zou, and A. Jemal, “Cancer statistics, 2014,” *CA: a cancer journal for clinicians*, vol. 64, no. 1, pp. 9–29, 2014.

[2] X. Yao, “A comparison of mammography, ultrasonography, and far-infrared thermography with pathological results in screening and early diagnosis of breast cancer,” *Asian Biomed*, vol. 8, no. 1, 2014.

[3] T. B. Borchartt, A. Conci, R. C. Lima, R. Resmini, and A. Sanchez, “Breast thermography from an image processing viewpoint: A survey,” *Signal Processing*, vol. 93, no. 10, pp. 2785–2803, 2013.

[4] L. Silva, G. Sequeiros, M. L. Santos, C. Fontes, D. C. Muchaluat-Saade, and A. Conci, “Thermal signal analysis for breast cancer risk verification,” in *MEDINFO’15 - 15th World Congress on International Health and Biomedical Informatics*, 2015.

[5] N. Arora, D. Martins, D. Ruggerio, E. Tousimis, A. J. Swistel, M. P. Osborne, and R. M. Simmons, “Effectiveness of a noninvasive digital infrared thermal imaging system in the detection of breast cancer,” *The American Journal of Surgery*, vol. 196, no. 4, pp. 523–526, 2008.

[6] D. Machado, G. Giraldi, A. Novotny, R. Marques, and A. Conci, “Topological derivative applied to automatic segmentation of frontal breast thermograms,” 2013.

[7] S. Suganthi and S. Ramakrishnan, “Semi automatic segmentation of breast thermograms using variational level set method,” in *The 15th International Conference on Biomedical Engineering*. Springer, 2014, pp. 231–234.

[8] S. S. Srinivasan and R. Swaminathan, “Segmentation of breast tissues in infrared images using modified phase based level sets,” in *Biomedical Informatics and Technology*. Springer, 2014, pp. 161–174.

[9] Q. Zhou, Z. Li, and J. K. Aggarwal, “Boundary extraction in thermal images by edge map,” in *Proceedings of the 2004 ACM symposium on Applied computing*. ACM, 2004, pp. 254–258.

[10] L. F. Silva, G. O. S. Olivera, S. Galvao, J. B. Silva, A. A. S. M. D. Santos, D. C. Muchaluat-Saade, and A. Conci, “Análise de séries temporais de sinais térmicos da mama para detecção de anomalias (analysis of time series of breast thermal signs for anomaly detection),” in *WIM - XIV Workshop de InformÃątica MÃľdica*. Anais CSBC, 2014, pp. 1818–1827.

[11] U. R. Acharya, E. Y.-K. Ng, J.-H. Tan, and S. V. Sree, “Thermography based breast cancer detection using texture features and support vector machine,” *Journal of medical systems*, vol. 36, no. 3, pp. 1503–1510, 2012.

[12] T. Jakubowska, B. Wiecek, M. Wysocki, C. Drews-Peszynski, and M. Strzelecki, “Classification of breast thermal images using artificial neural networks,” *Journal of Medical Informatics & Technologies*, vol. 7, pp. 41–50, 2004.

[13] S. V. Francis, M. Sasikala, and S. Saranya, “Detection of breast abnormality from thermograms using curvelet transform based feature extraction,” *Journal of medical systems*, vol. 38, no. 4, pp. 1–9, 2014.

[14] M. C. Araújo, R. C. Lima, and R. M. De Souza, “Interval symbolic feature extraction for thermography breast cancer detection,” *Expert Systems with Applications*, vol. 41, no. 15, pp. 6728–6737, 2014.

[15] U. R. Acharya, E. Y.-K. Ng, S. V. Sree, C. K. Chua, and S. Chattopadhyay, “Higher order spectra analysis of breast thermograms for the automated identification of breast cancer,” *Expert Systems*, vol. 31, no. 1, pp. 37–47, 2014.

[16] S. Suganthi and S. Ramakrishnan, “Analysis of breast thermograms using gabor wavelet anisotropy index,” *Journal of medical systems*, vol. 38, no. 9, pp. 1–7, 2014.

[17] S. Prabha, K. Anandh, C. Sujatha, and S. Ramakrishnan, “Total variation based edge enhancement for level set segmentation and asymmetry analysis in breast thermograms,” in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 6438–6441.

[18] E. Rodrigues, A. Conci, T. Borchartt, A. Paiva, A. C. Silva, and T. MacHenry, “Comparing results of thermographic images based diagnosis for breast diseases,” in *Systems, Signals and Image Processing (IWSSIP), 2014 International Conference on*. IEEE, 2014, pp. 39–42.

[19] A. Tharwat, T. Gaber, and A. E. Hassanien, “Cattle identification based on muzzle images using gabor features and svm classifier,” in *Advanced Machine Learning Technologies and Applications*. Springer, 2014, pp. 236–247.

[20] L. Silva, D. Saade, G. Sequeiros, A. Silva, A. Paiva, R. Bravo, and A. Conci, “A new database for breast research with infrared image,” *Journal of Medical Imaging and Health Informatics*, vol. 4, no. 1, pp. 92–100, 2014.

[21] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

[22] H. Xu, C. Caramanis, and S. Mannor, “Robustness and regularization of support vector machines,” *The Journal of Machine Learning Research*, vol. 10, pp. 1485–1510, 2009.

[23] A. Tharwat, T. Gaber, A. E. Hassanien, H. A. Hassanien, and M. F. Tolba, “Cattle identification using muzzle print images based on texture features approach,” vol. 303, pp. 217–227, 2014.

[24] A. Tharwat, T. Gaber, A. E. Hassanien, M. Shahin, and B. Refaat, “Sift-based arabic sign language recognition system,” vol. 334, pp. 359–370, 2015.

[25] N. A. Semary, A. Tharwat, E. Elhariri, and A. E. Hassanien, “Fruit-based tomato grading system using features fusion and support vector machine,” in *Intelligent Systems’ 2014*. Springer, 2015, pp. 401–410.

[26] M. Etehadtavakol, E. Ng, V. Chandran, and H. Rabbani, “Separable and non-separable discrete wavelet transform based texture features and image classification of breast thermograms,” *Infrared Physics & Technology*, vol. 61, pp. 274–286, 2013.

DOI: 10.15439/2015F147
ACSIS, Vol. 5

# Brain Image Classification Based on Automated Morphometry and Penalised Linear Discriminant Analysis with Resampling

Eva Janousova, Daniel Schwarz
Masaryk University, Institute of
Biostatistics and Analyses,
Kamenice 3, 625 00 Brno,
Czech Republic
Email: {janousova,
schwarz}@iba.muni.cz

Giovanni Montana
Imperial College London,
SW7 2AZ, London,
United Kingdom
Email: g.montana@imperial.ac.uk

Tomas Kasparek
Masaryk University, Department
of Psychiatry, Jihlavska 20, Brno,
Czech Republic
Email: tkasparek@fnbrno.cz

***Abstract*—This paper presents a new data-driven classification pipeline for discriminating two groups of individuals based on the medical images of their brain. The algorithm combines deformation-based morphometry and penalised linear discriminant analysis with resampling. The method is based on sparse representation of the original brain images using deformation logarithms reflecting the differences in the brain in comparison to the normal template anatomy. The sparse data enables efficient data reduction and classification via the penalised linear discriminant analysis with resampling. The classification accuracy obtained in an experiment with magnetic resonance brain images of first episode schizophrenia patients and healthy controls is comparable to the related state-of-the-art studies.**

## I. Introduction

The last two decades have witnessed an explosive growth in the ability to “understand the human brain” – a key to progress in neuroscience, to promote and protect brain health, and to develop treatments for restoring, regenerating, and repairing diseased brain functions. The motivations for that are clear: as populations inevitably grow older, mental disorders will increase dramatically – implying economic and social implications. The identification, characterization and validation of biomarkers for the major mental disorders would facilitate accurate prediction of disease risk, course, and therapeutic responses and ultimately lead to knowledge-based treatment and preventive strategies.

Computational neuroanatomy is a growing field of powerful applications of imaging modalities and computational techniques in neuroscience. It promises an automated methodology to characterize neuroanatomical configuration of structural magnetic resonance imaging (MRI) brain scans. One of the crucial techniques in this methodology is image registration. Its task is to find a spatial transformation which maps each point of an image onto its corresponding point of another image. Atlas-based registration is a special technique of computational neuroanatomy – not seen widely in other fields of biomedical imaging. It performs the task of spatial normalization of images according to a common reference anatomy termed as a brain atlas. Together with techniques adopted from inferential statistics and hypothesis testing, it allows to uncover brain regions with significant morphological differences between normal and clinical populations. Such techniques have been already used also in modern psychiatry research to seek for biomarkers and neurobiology of various mental diseases [1]–[4].

The real challenge for psychiatry is, however, to move from group analysis between patients and healthy volunteers to computer aided diagnostics on the level of an individual patient. Although pioneering works employing machine learning techniques have recently borne fruit in case of neurological diseases, this is extremely difficult in mental diseases. For instance, in schizophrenia – a disease with a complex neurobiology – the brain-imaging measurements in patients show considerable overlap with the normal range [5].

Algorithms, which have been proposed in the diagnostics of neurodegenerative disorders, relied on brain image data classification between patients and healthy controls. The most commonly used classification methods have been the linear discriminant analysis (LDA) [6]–[10], support vector machines (SVM) [11]–[13] or the k-nearest neighbour algorithm [14]–[16]. Due to the large amount of features obtained from 3-D medical images, the classification is often preceded by data reduction performed by principal component analysis [17], independent component analysis [18], selection of regions of interests (ROI) [6]–[8], and other data reduction methods. So far, only few studies have presented complex pipelines for data reduction and classification, such as the COMPARE method [19], which combines deformation-based morphometry with machine learning methods (watershed segmentation algorithm and support vector machine-recursive feature elimination

This work was supported by the grant IGA MH CZ NR No. NT 13359-4/2012.

technique). In [19] the COMPARE algorithm, classification of schizophrenia patients with very high classification accuracy (91.8% for female subjects and 90.8% for male subjects) was applied. Thus, the complex pipeline seems to enable classification with a higher efficiency than other commonly used methods that have reported classification accuracy between 70% and 90% [5]–[18].

The aim of this paper is to present a new data-driven complex classification pipeline consisting of deformation-based morphometry (DBM) and penalised linear discriminant analysis (pLDA) with resampling. The DBM-pLDA algorithm enables efficient data reduction and subsequent classification as it is based on sparse representation of the original image data. The algorithm starts with an application of the DBM on original MRI data to create 3-D deformations, which are then log-transformed and used in the pLDA with resampling to identify the brain regions with different local volumes in patients and controls. The last step comprises of classifying the brain images into groups of patients and healthy controls based on the features representing automatically detected brain regions.

The pLDA has been successfully employed in our previous imaging-genetics study [20] in which image phenotypes, using pre-selected pLDA, were used for searching the genotypes most associated with Alzheimer's disease. To our knowledge, this is the first study that uses pLDA for selecting brain imaging features for the purposes of distinguishing diseased individuals from healthy controls in schizophrenia research.

The rest of the paper is organized as follows. In Section II, we describe all the necessary steps in the classification pipeline. Section III shows the application of the proposed DBM-pLDA algorithm on T1-weighted MRI data of first-episode schizophrenia patients and healthy controls. Section IV discusses the results and concludes the paper.

## II. Methods

The proposed data-driven algorithm for image classification is based on combining the deformation-based morphometry and penalised linear discriminant analysis with resampling. In the DBM, high-resolution nonlinear registration [21] of 3-D brain images with a digital brain atlas is performed. The resulting 3-D deformations represented by the displacement fields or their Jacobian determinants clearly show how the brain anatomy of a subject differs from the normal template anatomy in terms of local volume contractions and expansions. After logarithmic transformation, the 3-D deformations tend to be sparse, i.e. numerous voxels have zero or close to zero values. The sparse representation of the image data together with pLDA with resampling leads to effective selection of the most discriminating regions, as explained below. The selected brain voxels are then used as features in image classification. The DBM and pLDA are described in more details in the following two subchapters.

### *A. Deformation-Based Morphometry*

Here, a brief summary is given on our algorithm for high-resolution deformation-based morphometry with the underlying registration method, based on a spatial deformation model that allows for large deformations while preserving the topology of the images. Details can be found in [21].

The registration method operates directly on image intensity values with no data reduction by segmentation or classification. The 3-D displacement field which maximizes global mutual information between a reference image and a floating image is searched in an iterative process that involves computation of the local forces as a gradient of point similarity measures and their regularization using the spatial deformation model. The regularization involves two Gaussian spatial filters forming the combined elastic-incremental model [22]. The first spatial filter regularizes displacement improvements that are proportional to the applied forces. These displacements are integrated into the final deformation, which is done iteratively by summation. The second part of the model represents the property of elastic materials in which displacements wane upon retracting the forces. This is ensured by a second Gaussian smoother. The resulting deformations preserve the topology of the images, i.e. only one-to-one mappings, termed diffeomorphic, are obtained. This requirement is satisfied by controlling the standard deviations of the Gaussian filters that affect the behaviour of the spatial deformation model. Standard deviations are incremented each time the minimum Jacobian determinant drops below a predefined threshold. The deformation should capture subtle anatomical variations among the studied images; therefore, the standard deviations of the Gaussians are decremented as well whenever the minimum Jacobian determinant starts rising during the registration process.

### *B. Penalised Linear Discriminant Analysis with Resampling*

Prior to the penalised linear discriminant analysis with resampling, the logarithms of the Jacobian determinants computed from 3-D deformation fields are transformed into 1-D vectors and arranged in $(n \times p)$ matrix $\mathbf{X}$, where $n$ is the number of individuals in a data set and $p$ is the number of voxels in each deformation. All columns of the matrix $\mathbf{X}$ are mean-centred and have unit variance. It is assumed that all $n$ individuals have been labelled as one of the two classes, which are denoted by $D$ (diseased individuals) and $H$ (healthy controls). The number of individuals in each class is $n_D$ and $n_H$, respectively, and $n = n_D + n_H$.

The common LDA aims at finding a direction vector $\mathbf{v}$ that best discriminates two classes within a data sample via maximizing the between-class variance and simultaneous minimizing of the within-class variance. The between-class scatter matrix, denoted as $\mathbf{S}_B$, is calculated as:

$$\mathbf{S}_B = (\mathbf{m}_H - \mathbf{m}_D)^T(\mathbf{m}_H - \mathbf{m}_D), \tag{1}$$

where $\mathbf{m}_H = \frac{1}{n_H}\sum_{i=1}^{n_H}\mathbf{x}_{i.}$ is the mean vector of class *H*, $\mathbf{m}_D = \frac{1}{n_D}\sum_{i=1}^{n_D}\mathbf{x}_{i.}$ is the mean vector of class *D* and $\mathbf{x}_{i.}$, $i = 1, ..., n$, are rows of the matrix $\mathbf{X}$.

The within-class scatter matrix, denoted as $\mathbf{S}_W$, is defined as:

$$\begin{aligned}\mathbf{S}_W &= \sum_{i=1}^{n_H}(\mathbf{x}_{i.} - \mathbf{m}_H)^T(\mathbf{x}_{i.} - \mathbf{m}_H) \\ &+ \sum_{i=1}^{n_D}(\mathbf{x}_{i.} - \mathbf{m}_D)^T(\mathbf{x}_{i.} - \mathbf{m}_D).\end{aligned} \tag{2}$$

The direction vector **v** is then a solution of the following optimization problem:

$$\max_{\mathbf{v}}\{\mathbf{v}^T\mathbf{S}_B\mathbf{v}\} \text{ subject to } \mathbf{v}^T\mathbf{S}_W\mathbf{v} = 1. \tag{3}$$

In the penalised LDA [23], a penalty is imposed on the $l_1$ norm of the direction vector **v** which leads to setting of coefficients $\mathbf{v}_j$, $j = 1, ..., p$, of the least discriminative features to zero. If the input data into the pLDA are already sparse, the amount of selected features (i.e. features with non-zero coefficients) is smaller and the classification results tend to be more stable than while using original brain images in the data analysis. In pLDA, the optimisation problem changes to:

$$\max_{\mathbf{v}}\left\{\mathbf{v}^T\mathbf{S}_B\mathbf{v} - \lambda\sum_{j=1}^{p}s_j|v_j|\right\} \text{ subject to } \mathbf{v}^T\mathbf{S}_W^*\mathbf{v} = 1, \tag{4}$$

where $\mathbf{S}_W^*$ is the diagonal estimate of $\mathbf{S}_W$, $\operatorname{diag}(\mathbf{S}_W) = (s_1^2, ..., s_p^2)$ and $\lambda$ is the regularization parameter that controls the number of selected features. Specifically, when $\lambda$ is exactly zero, no penalty is imposed and all *p* features contribute in the direction vector **v**. As $\lambda$ increases from zero, less features contribute in **v**. At its maximum value $\lambda_{\max}$, all coefficients of **v** are set to zero. A common approach for tuning $\lambda$ involves cross-validating the prediction error for a grid of values of $\lambda$ and selecting the value of $\lambda$ that leads to the smallest cross-validated error. However, this approach may be prone to sampling errors. Therefore, we opted for a resampling method proposed in [24] for sparse predictive modelling. This procedure aims to calculate selection probabilities for each feature by repeatedly fitting the pLDA model on random subsets of the data set, while keeping track of the features associated to non-zero coefficients of **v**. The final set of the most discriminative features consists of voxels with selection probability higher than 0.5.

It should be noted that by using the data resampling, a set of features over the range $[\lambda_{\min}, \lambda_{\max}]$ with stable classification results is selected instead of tuning the regularization parameter $\lambda$ [24]. The final set of selected features is then used for classification of individuals into the class *D* or *H* using LDA.

## III. Experiment and Results

### *A. Imaging Data Used in the Experiment*

The proposed classification pipeline was tested in an experiment with magnetic resonance brain imaging data of 52 male patients with first-episode schizophrenia and 52 sex- and age-matched healthy control subjects. The median age of the patients and controls was 22.9 years (range 17-40 years) and 23.0 (range 18-38 years), respectively. Thirty-nine patients and the same number of controls took part in our previous study [25]. All subjects signed an informed consent before entering the study.

All 52 patients were recruited from males hospitalised in the all-males unit of the Department of Psychiatry, Masaryk University in Brno for first-episode schizophrenia. The diagnosis was established during a clinical interview guided by the International Statistical Classification of Disease and Related Health Problems (ICD-10) research criteria and was focused on information about the family and personal history, the somatic conditions, substance abuse, pharmacological history and the current treatment, previous psychiatric conditions and, finally, on the current clinical manifestation, the presenting symptoms, the duration, and the functional impact. Next, the patients were physically examined, including blood (haematology and biochemistry) and urine analysis (biochemistry and toxicology). If abnormal findings were present, their origin was traced by additional examination. A fully trained senior psychiatrist (board certified in psychiatry) reviewed all information, established the diagnosis and suggested the case for inclusion in the study. MRI examination was performed during the first episode, that is, all patients were treated with antipsychotics for 3-14 weeks only at the time of MRI.

The 52 healthy control subjects without substance dependence, family or personal history of axis I psychiatric conditions, neurological or somatic conditions affecting the structure or function of the brain, and contraindications for MRI examination, were recruited from the community, the local staff and medical students.

All subjects were scanned with 1.5T Siemens Symphony machine. Whole head T1-weighted images were obtained using 3-D acquisition with IR/GR sequence, TR 1700 ms, TE 3.93 ms, TI 1100 ms, flip angle 15°, 160 slices, voxel size 1.17 × 0.48 × 0.48 mm, FOV 246 × 246 mm, and matrix size 512 × 512 voxels.

The 3-D T1-weighted images were checked for abnormalities and then pre-processed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). Specifically, the images

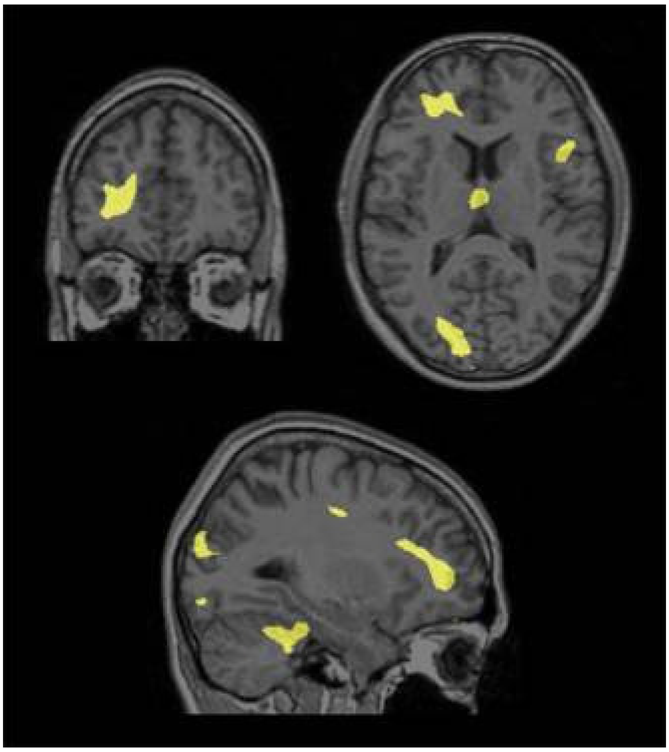


Fig. 1 Coronal, transversal and sagittal slices showing the automatically detected highly discriminative voxels (in yellow).

were corrected for bias-filed inhomogeneity and spatially normalised, i.e. transformed into stereotactic space.

### B. Classification efficiency

The DBM-pLDA algorithm was applied on the pre-processed images. The classification efficiency of the DBM-pLDA algorithm was evaluated with the leave-one-out cross-validation technique to avoid biased results. Consecutively, each of the $n$ subjects was selected as a testing subject and the remaining $n$-1 subjects were used for training the classifier. The testing subject was classified into the patient or healthy control class. Then, the resulting class was compared to the true classification label. The classification performances for all subjects were combined in order to create the overall classification performance measures, namely accuracy, sensitivity, specificity and precision, defined as:

$$sensitivity = \frac{TP}{TP+FN}, \tag{1}$$

$$specificity = \frac{TN}{TN+FP}, \tag{2}$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{3}$$

$$precision = \frac{TP}{TP+FP}, \tag{4}$$

TABLE I
NUMBER OF SELECTED VOXELS AND CROSS-VALIDATED CLASSIFICATION PERFORMANCE MEASURES IN PERCENTAGE FOR VARIOUS $\lambda_{\min}$ VALUES.

| | $\lambda_{\min} = 0.3$ | $\lambda_{\min} = 0.4$ | $\lambda_{\min} = 0.5$ | $\lambda_{\min} = 0.6$ |
|---|---|---|---|---|
| # voxels | 315,123 | 107,967 | 30,461 | 5,113 |
| Accuracy | 82.7 | 84.6 | 85.6 | 83.7 |
| Sensitivity | 84.6 | 86.5 | 84.6 | 78.8 |
| Specificity | 80.8 | 82.7 | 86.5 | 88.5 |
| Precision | 81.5 | 83.3 | 86.3 | 87.2 |

where $TP$, $TN$ represent numbers of true positive and true negative results respectively, and $FP$, $FN$ represent numbers of false positive and false negative results respectively.

### C. Experiment Results

The performance measures obtained in the classification of 104 schizophrenia patients and healthy controls are summarized in Table I. Our results demonstrate that the classification accuracy is stable for meaningful ranges of $\lambda$, i.e. ranges leading to an adequate number of selected voxels. In this experiment, $\lambda_{\max}$ was fixed to 0.9 and so, the table shows the results for the ranges $[\lambda_{\min}, 0.9]$. The best cross-validated classification accuracy was 85.6% (sensitivity 84.6%, specificity 86.5%) while selecting about 30,000 most discriminative voxels. The selected features composed of the most discriminating voxels are shown in Fig. 1. They form connected regions in the left prefrontal cortex, the right anterior insula, the medial parts of the thalamus, and the cerebellar cortex.

In order to compare the proposed DBM-pLDA algorithm to other classification methods, the leave-one-out cross-validation procedure was carried also with other classifiers, which have been used on neuroimaging data frequently: (i) LDA and (ii) SVM with linear kernel. Features for these classifiers were selected with the use of mass univariate analysis (Student's t-test, $p < 0.01$), so that only the significant local volume changes in patients when compared to healthy controls were input to the classifiers. The resulting classification performance measures are summarized in Table II. The results show an improved classification efficiency of the proposed DBM-pLDA algorithm, when compared to both DBM-LDA and DBM-SVM algorithms.

## IV. DISCUSSION AND CONCLUSIONS

A classification pipeline for discriminating two groups of individuals based on brain images has been presented. The fully automated data-driven algorithm consists of deformation-based morphometry and penalised linear discriminant analysis with resampling. Firstly, sparse representation of the original brain images using logarithms of deformations, that are the results of high-dimensional nonlinear registration of the brain images with a digital brain atlas, is acquired. Secondly, the sparse data is reduced using

TABLE II
CLASSIFICATION EFFICIENCY OF THE PROPOSED DBM-PLDA ALGORITHM COMPARED TO DBM-LDA AND DBM-SVM ALGORITHMS.

| | Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|---|
| DBM-pLDA | 85.6 | 84.6 | 86.5 | 86.3 |
| DBM-LDA | 66.3 | 59.6 | 73.1 | 68.9 |
| DBM-SVM | 70.2 | 69.2 | 75.0 | 73.5 |

pLDA with resampling and then classified into the group of patients or healthy controls.

To our knowledge, this is for the first time when pLDA is used for selecting highly discriminative brain imaging features in schizophrenia. In our previous study, pLDA was successfully applied for selecting image phenotypes that were used in searching for genotypes most associated with Alzheimer's disease [20]. Here, pLDA is used in combination with DBM in the pipeline for the classification of MRI data of patients with first episode schizophrenia (FES) and healthy controls. Only FES patients were used in this study as it is known that longer duration of the illness leads to higher magnitude of morphological changes in the brain of schizophrenia patients [26]. Thus, the classification results can be overestimated if the data set contains chronic schizophrenia patients as well. To avoid biased results, only males were used in the analysis, as it was shown that there are differences in structural brain abnormalities in males and females [27].

The proposed algorithm uses the whole volume of the brain and is fully automated. Thus, it overcomes limitations of traditional ROI-based classification analyses. The ROI-based methods need prior knowledge about the regions that might be affected by the disease. As schizophrenia is a large-scale disorder of neurocognitive networks rather than confined to specific regions [1]–[4], whole brain analysis is more appropriate. Moreover, the fully automated method is less time consuming and error-prone than manual tracing of the ROIs.

High classification accuracy (85.6%) has been achieved while applying the proposed DBM-pLDA algorithm on T1-weighted MRI data of first-episode schizophrenia patients and healthy controls. The efficiency of the algorithm is comparable or superior to the other state-of-the-art studies dealing with the classification of schizophrenia patients [5]-[18]. However, the classification performance is smaller than in [19], in which the COMPARE algorithm enabled classification of schizophrenia and healthy females with accuracy equal to 91.8% and the classification of diseased and healthy males with accuracy of 90.8%. Nevertheless, Fan et al. [19] used a mixed data set of patients with first-episode schizophrenia and chronic schizophrenia. The fact that the morphological changes in chronic schizophrenia patients progress during the course of the disease, while the first episode schizophrenia is characterized only by subtle morphological abnormalities, could overestimate the results reported in [19].

A by-product of the classification pipeline is a selection of the most discriminating brain morphological features between patients and controls. The automatically detected discriminating brain regions were located in the left prefrontal cortex, the right anterior insula, the medial parts of the thalamus, and the cerebellar cortex. These results are consistent with those published in previous studies [1]–[4]. Moreover, it is known that these brain areas are involved in higher cognitive, integrative and regulatory functions that are impaired in schizophrenia [1]–[4].

Even though the results are promising, further experiments are necessary to investigate whether the DBM-pLDA algorithm can assist in the early diagnosis of schizophrenia. A limitation of this study might be the slightly limited sample size, as Nieuwenhuis et al. [12] recommend to use more than 130 subjects. However, our data set containing 104 subjects is larger than the data sets used in most of the schizophrenia studies [5]–[8], [11], [14]–[15], [17]–[19]. The next step in our research will be a replication using a completely independent set of schizophrenia subjects. We would also like to test the performance of the algorithm on other patient groups.

In conclusion, the associations between the automatically detected discriminating morphology features and their significance in the neurobiology of schizophrenia, as well as the high accuracy of classification of patients and healthy controls, demonstrate the face validity of our approach that combines DBM and pLDA with resampling.

## REFERENCES

[1] I. C. Wright, S. Rabe-Hesketh, P. W. R. Woodruff, A. S. David, R. M. Murray, and E. T. Bullmore, "Meta-analysis of regional brain volumes in schizophrenia," *Am. J. Psychiatry*, vol. 157, no. 1, pp. 16–25, Jan. 2000.

[2] M. E. Shenton, C. C. Dickey, M. Frumin, and R. W. McCarley, "A review of MRI findings in schizophrenia," *Schizophr. Res.*, vol. 49, no. 1–2, pp. 1–52, Apr. 2001.

[3] M. A. Niznikiewicz, M. Kubicki, and M. E. Shenton, "Recent structural and functional imaging findings in schizophrenia," *Curr. Opin. Psychiatry*, vol. 16, no. 2, pp. 123–147, Mar. 2003.

[4] R. Honea, T. J. Crow, D. Passingham, and C. E. Mackay, "Regional deficits in brain volume in schizophrenia: A meta-analysis of voxel-based morphometry studies," *Am. J. Psychiatry*, vol. 162, no. 12, pp. 2233–2245, Dec. 2005.

[5] D. Sun, T. G. M. van Erp, P. M. Thompson, C. E. Bearden, M. Daley, L. Kushan, M. E. Hardt, K. H. Nuechterlein, A. W. Toga, and T. D. Cannon, "Elucidating a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis: Classification Analysis Using Probabilistic Brain Atlas and Machine Learning Algorithms," *Biol. Psychiatry*, vol. 66, no. 11, pp. 1055–1060, Dec. 2009.

[6] C. M. Leonard, J. M. Kuldau, J. I. Breier, P. A. Zuffante, E. R. Gautier, D. C. Heron, E. M. Lavery, J. Packing, S. A. Williams, and C. A. DeBose, "Cumulative effect of anatomical risk factors for schizophrenia: An MRI study," *Biol. Psychiatry*, vol. 46, no. 3, pp. 374–382, Aug. 1999.

[7] K. Nakamura, Y. Kawasaki, M. Suzuki, H. Hagino, K. Kurokawa, T. Takahashi, L. Niu, M. Matsui, H. Seto, and M. Kurachi, "Multiple structural brain measures obtained by three-dimensional magnetic resonance imaging to distinguish between schizophrenia patients and normal subjects," *Schizophr. Bull.*, vol. 30, no. 2, pp. 393–404, 2004.

[8] Y. Takayanagi, Y. Kawasaki, K. Nakamura, T. Takahashi, L. Orikabe, E. Toyoda, Y. Mozue, Y. Sato, M. Itokawa, H. Yamasue, K. Kasai, M. Kurachi, Y. Okazaki, M. Matsushita, and M. Suzuki, "Differentiation

of first-episode schizophrenia patients from healthy controls using ROI-based multiple structural brain variables," *Prog. Neuropsychopharmacol. Biol. Psychiatry*, vol. 34, no. 1, pp. 10–17, Feb. 2010.

[9] M. Ota, N. Sato, M. Ishikawa, H. Hori, D. Sasayama, K. Hattori, T. Teraishi, S. Obu, Y. Nakata, K. Nemoto, Y. Moriguchi, R. Hashimoto, and H. Kunugi, "Discrimination of female schizophrenia patients from healthy women using multiple structural brain measures obtained with voxel-based morphometry," *Psychiatry Clin. Neurosci.*, vol. 66, no. 7, pp. 611–617, Dec. 2012.

[10] E. Janousova, D. Schwarz, and T. Kasparek, "Combining various types of classifiers and features extracted from magnetic resonance imaging data in schizophrenia recognition," *Psychiatry Res. Neuroimaging*, to be published.

[11] K. M. Pohl and M. R. Sabuncu, "A Unified Framework for MR Based Disease Classification," in *Information Processing in Medical Imaging, Proceedings*, vol. 5636, J. L. Prince, D. L. Pham, and K. J. Myers, Eds. Berlin: Springer-Verlag Berlin, 2009, pp. 300–313.

[12] M. Nieuwenhuis, N. E. M. van Haren, H. E. H. Pol, W. Cahn, R. S. Kahn, and H. G. Schnack, "Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples," *Neuroimage*, vol. 61, no. 3, pp. 606–612, Jul. 2012.

[13] P. Dluhos, D. Schwarz, and T. Kasparek, "Wavelet features for recognition of first episode of schizophrenia from MRI brain images,“ *Radioengineering*, vol. 23, pp. 274–281, Apr. 2014.

[14] Y. X. Liu, L. Teverovskiy, O. Carmichael, R. Kikinis, M. Shenton, C. S. Carter, V. A. Stenger, S. Davis, H. Aizenstein, J. T. Becker, O. L. Lopez, and C. C. Meltzer, "Discriminative MR image feature analysis for automatic schizophrenia and Alzheimer's disease classification," in *Medical Image Computing and Computer-Assisted Intervention - Miccai 2004*, vol. 3216, C. Barillot, D. R. Haynor, and P. Hellier, Eds. Berlin: Springer-Verlag Berlin, 2004, pp. 393–401.

[15] P. Wang and R. Verma, "On Classifying Disease-Induced Patterns in the Brain Using Diffusion Tensor Images," in *Medical Image Computing and Computer-Assisted Intervention - Miccai 2008*, vol. 5241, D. Metaxas, L. Axel, G. Fichtinger, and G. Szekely, Eds. Berlin: Springer-Verlag Berlin, 2008, pp. 908–916.

[16] D. Schwarz and T. Kasparek, "Brain morphometry of MR images for automated classification of first-episode schizophrenia,“ *Information Fusion*, vol. 19, pp. 97–102, Sep. 2014.

[17] F. Shi, Y. Liu, T. Jiang, Y. Zhou, W. Zhu, J. Jiang, H. Liu, and Z. Liu, "Regional homogeneity and anatomical parcellation for fMRI image classification: Application to schizophrenia and normal controls," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2007*, vol. 4792, N. Ayache, S. Ourdelin, and A. Maeder, Eds. Berlin: Springer-Verlag Berlin, 2007, pp. 136–143.

[18] O. Demirci, V. P. Clark, and V. D. Calhoun, "A projection pursuit algorithm to classify individuals using fMRI data: Application to schizophrenia," *Neuroimage*, vol. 39, no. 4, pp. 1774–1782, Feb. 2008.

[19] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "COMPARE: Classification of morphological patterns using adaptive regional elements," *IEEE Trans. Med. Imaging*, vol. 26, no. 1, pp. 93–105, Jan. 2007.

[20] M. Vounou, E. Janousova, R. Wolz, J. L. Stein, P. M. Thompson, D. Rueckert, and G. Montana, "Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease," *Neuroimage*, vol. 60, no. 1, pp. 700–716, Mar. 2012.

[21] D. Schwarz, T. Kasparek, I. Provaznik, and J. Jarkovsky, "A deformable registration method for automated morphometry of MRI brain images in neuropsychiatric research," *IEEE Trans. Med. Imaging*, vol. 26, no. 4, pp. 452–461, Apr. 2007.

[22] P. Rogelj and S. Kovacic, "Spatial deformation models for non-rigid image registration," in *Proceedings of the 9th Computer Vision Winter Workshop (CVWW'04)*, Piran (Slovenia), 2004, pp. 79–88.

[23] D. M. Witten and R. Tibshirani, "Penalized classification using Fisher's linear discriminant," *J. R. Stat. Soc. Ser. B - Stat. Methodol.*, vol. 73, pp. 753–772, 2011.

[24] N. Meinshausen and P. Bühlmann, "Stability selection," *J. R. Stat. Soc. Ser. B - Stat. Methodol.*, vol. 72, pp. 417–473, 2010.

[25] T. Kasparek, C. E. Thomaz, J. R. Sato, D. Schwarz, E. Janousova, R. Marecek, R. Prikryl, J. Vanicek, A. Fujita, and E. Ceskova, "Maximum-uncertainty linear discrimination analysis of first-episode schizophrenia subjects," *Psychiatry Res.-Neuroimaging*, vol. 191, no. 3, pp. 174–181, Mar. 2011.

[26] I. Ellison-Wright, D. C. Glahn, A. R. Laird, S. M. Thelen, and E. Bullmore, "The anatomy of first-episode and chronic schizophrenia: An anatomical likelihood estimation meta-analysis," *Am. J. Psychiatry*, vol. 165, no. 8, pp. 1015–1023, Aug. 2008.

[27] H. Nasrallah, S. Schwarzkopf, S. Olson, and J. Coffman, "Gender Differences in Schizophrenia on Mri Brain-Scans," *Schizophr. Bull.*, vol. 16, no. 2, pp. 205–209, 1990.

DOI: 10.15439/2015F276
ACSIS, Vol. 5

# Computer based quantification of normal and pathological vocal folds phonatory processes from laryngovideostroboscopy

Bartosz Kopczyński
Institute of Electronics, Lodz
University of Technology,
ul. Wolczanska 211/215, 90-924
Lodz, Poland, e-mail:
bartosz.michal.k@gmail.com

Paweł Strumiłło
Institute of Electronics, Lodz
University of Technology,
ul. Wolczanska 211/215, 90-924
Lodz, Poland,
e-mail: pawel.strumillo@p.lodz.pl

Ewa Niebudek-Bogusz
Department of Audiology and
Phoniatrics, The Nofer Institute
of Occupational Medicine,
ul. Teresy 8, 91-348 Lodz, Poland
e-mail: ebogusz@imp.lodz.pl

*Abstract* — **Medical imaging techniques offer novel visualization and analysis methods of the vocal folds during phonation and automatic computation of indices aiding the phoniatrist in a more precise diagnosis of voice disorders. The aim of this study is to apply *computer vision algorithms for qualitative and quantitative analysis of vocal folds' vibrations. Videostroboscopic examinations of the larynx were carried out for 30 patients. Image pre-processing and image segmentation algorithms were applied to compute the glottis area during phonation.* The glottovibrograms which are spatio-temporal visualizations of the vibrating vocal folds were also built. The proposed indices allow for a quantitative and comparative analysis of normal and disordered phonatory processes. The conducted pilot study has confirmed the validity of the computer aided imaging methods for the qualitative and quantitative analysis of the videostroboscopic images of the phonatory motions of the vocal folds.**

## I. Introduction

Early diagnosis of occupational voice disorders is becoming one of the priorities in public health in Poland and in other countries of the European Union. Currently, European standards emphasize the need for a comprehensive assessment of voice disorders, including the assessment of the larynx function during the phonatory tests [1]. The test that allows the specialist to accurately assess the condition of the voice organ and is recognized as the gold standard is the laryngovideostroboscopy (LVS) [2]. The aim of our study is to apply computer image processing and analysis methods for quantification of vocal folds' phonatory activity by examining sequences of LVS images. In this communication we focus on examining videostroboscopic images of normophonic individuals with healthy vocal folds and patients with diagnosed nodules.

## II. Related Works

The standardization of automatic segmentation of vocal folds videostroboscopic images remains an unsolved problem since 1995 [3]. There have been several automatic segmentation methods of the vocal folds images proposed [4]-[6]. Many of the developed methods and algorithms are designed for specific image recording conditions and work properly only for local databases containing videos collected in a particular institute, hospital or health center. It has turned out to be a very difficult problem to work out an algorithm which would give satisfactory results for every given video presenting the vibrating vocal folds. Authors in [7] proposed a method based on supervised thresholding methods by applying the Fourier descriptors. This method is additionally combined with a glottal neighborhood descriptor which specifies distance–weighted color differences between the glottis and the surroundings tissues of the vocal folds. The additional knowledge of local color distributions increases the recognition quality.

The key task preceding the image analysis methods of the glottis is segmenting out the space between the vocal folds, termed the glottal area. The most popular and straightforward image segmentation method is thresholding. The threshold value is selected on the basis of the image histogram. Local minima of the multimodal histogram designate threshold values. In the simplest case the first histogram minimum assigns a threshold value that distinguishes the image regions belonging to the space between the vocal folds from the regions representing the background (glottis environment). More sophisticated threshold methods utilize the neighborhood information, adaptively adjusting the threshold value [8] or including wavelet transformation [9]. There are several methods which work properly for certain image classes i.e. region based, model based and their combinations. More complex methods are based on the so called active contour models. Active contours are energy minimizing flexible splines guided by hypothetical external forces influenced by the image content and internal forces arising from the curvature and continuity of the nodes forming the contour. Note, however, that the contour must be properly initialized near the object of interest [Marendic et al. 2001].

## III. Materials and methods

The general block diagram of the developed image processing and analysis algorithms for analysis of laryngostroboscopic images is shown in Fig. 1.

978-83-60810-66-8/$25.00©2015, IEEE

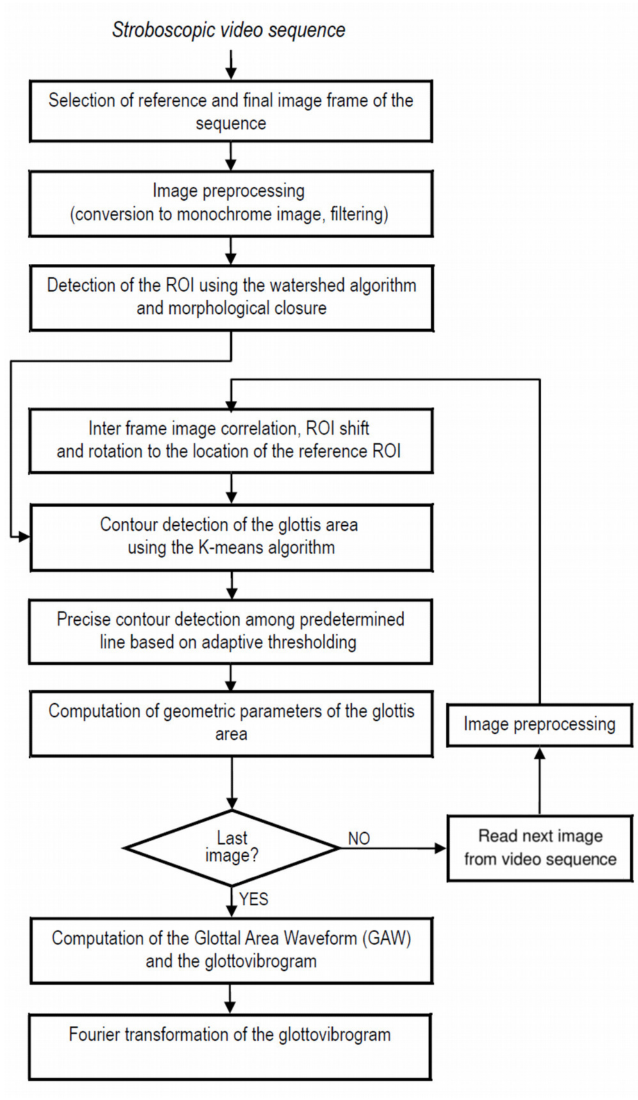

Stroboscopic video sequence
Selection of reference and final image frame of the sequence
Image preprocessing (conversion to monochrome image, filtering)
Detection of the ROI using the watershed algorithm and morphological closure
Inter frame image correlation, ROI shift and rotation to the location of the reference ROI
Contour detection of the glottis area using the K-means algorithm
Precise contour detection among predetermined line based on adaptive thresholding
Computation of geometric parameters of the glottis area
Image preprocessing
Last image?
NO
Read next image from video sequence
YES
Computation of the Glottal Area Waveform (GAW) and the glottovibrogram
Fourier transformation of the glottovibrogram

Fig 1. Schematic of the algorithm for qualitative and quantitative analysis of the vocal folds' phonation from videostroboscopy

The algorithm was developed after numerous consultations with phoniatrists working in the Department of Audiology and Phoniatrics, The Nofer Institute of Occupational Medicine in Lodz, Poland. The algorithm allows the user to select the analyzed film and indicate the phonation region of the glottis. Then the algorithm proceeds with automatic computations of the following quantities and representations:

- fundamental frequency of the vocal folds' vibration and the vibration frequency at each level of the total glottis length,
- geometric and kinetic parameters of the glottal area (the space between the vocal folds) during phonation,
- relations between the opening and closing time of the glottis,
- the glottal area waveform (GAW), illustrating time variations of the glottal area,
- the glottovibrogram which is a time-space representation of the glottal gap (local distances between the vocal folds at different levels of the glottis) imaged as a grayscale image (see Figs. 4, 5),
- the Fourier transform of the glottovibrogram.

Videostroboscopic examinations of the larynx were carried out for 30 patients, i.e. for 15 individuals with no voice disorders (see an example in Fig. 2) and 15 patients with diagnosed nodules (see Fig. 3). At the outset of the analysis the phoniatrist is asked to select the reference sequence of images for analysis i.e. the starting and ending frame from the videostroboscopic film. This is the only user-dependent step in the proposed image analysis procedure. After defining the reference sequence the program determines the region of interest and adjusts processing parameters to the selected regions of the analyzed video and reduces the distortions due to movements of the camera versus the glottis.

The aim of the second step is to apply general image processing methods for image enhancement. Color images in the LVS sequence are converted from an RGB to a monochrome image format, by applying the following weighted average of the color components:

$$I = 0.299\,R + 0.587\,G + 0.114\,B \qquad (1)$$

in which $I$ is the level of intensity in the monochrome image and $R$, $G$, $B$ are the values of the red, green and blue components of the color image correspondingly. Digital image filtering is then applied to reduce noise and artifacts due to unwanted reflections and compression distortions. This step allows the residual adaptation of parameters for the varying shape and location of the region of interest (ROI) which undergoes the following distortions: non-uniform and time-varying illumination, loss of sharpness due to the vapor covering the camera lens, varying distance relative to the vocal folds and loss of camera focus. Preprocessing of the region of interest additionally improves results of the thresholding step.

Then, an automatic analysis of the LVS images starts with the detection of the ROI, i.e. a rectangular region containing the examined part of the glottis. Similarly to the approach adopted in [10] this was improved with morphological operations (erosion and closure) and basic filtering methods comprising median and lowpass filtering. Finally, the watershed image segmentation algorithm is employed to roughly determine the glottal area ROI. The watershed transformation enables extraction of areas in which there is continuity in terms of image features such as brightness or color. This transformation method is considered as one of the most effective methods of image segmentation [11]. It yields well designated areas of the image with clearly defined outlines. The behavior of this method lies in the distribution of the image gradient which separates areas of the image assigned to the seeds along the thresholded value of the gradient.

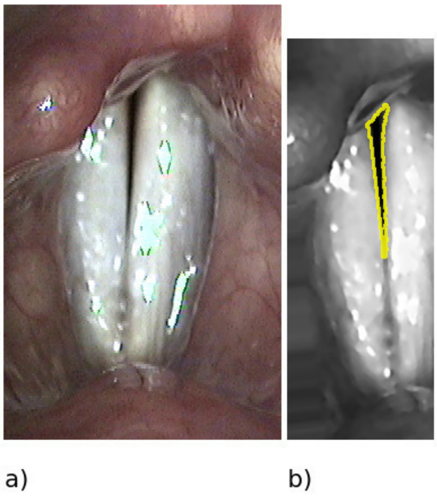


Fig 2. Stroboscopic images of the normophonic vocal folds (i.e. no diagnosed abnormalities): input image a), segmentation result b)

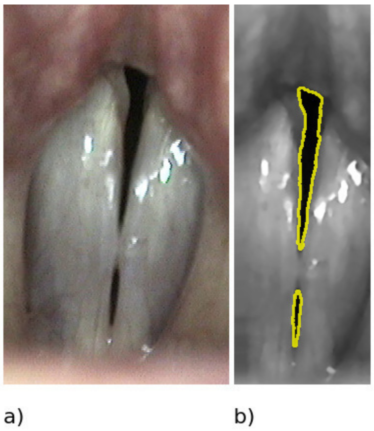


Fig 3. Stroboscopic images of the vocal folds with nodules: input image a), segmentation result b)

While being a robust segmentation method, the watershed algorithm features poor precision in determination of the glottis boundary. Thus, this segmentation method is adopted for the first frame of the image sequence only and serves as a reference location of the ROI (see block diagram in Fig. 1) After defining the coarse position of the ROI obtained by the watersheds, the precise contour, i.e. the boundaries of the vocal folds, is being extracted with the use of the *K*-means algorithm. The *K*-means algorithm is an iterative data clustering method in which the image content is partitioned into *K*-separate regions, in which each pixel of the image is assigned to the region with the nearest distance to the centroid of the region, i.e. according to the criterion:

$$minimum(S)=\sum_{i=1}^{K}\sum_{x\in S_i}||x-u_i||^2 \qquad (2)$$

where: $K$ – is the number of segmentation regions, $x$ – is grey level value of a pixel, $S_i$ – is the $i$-th segmentation region, and $u_i$ – is the centroid of pixels within $S_i$ [12]. The main advantage of the *K*-means algorithm is its computational simplicity yielding such a partitioning result that the shapes of the clusters are maximally compact.

In our application, at the first step we define two (*K = 2*) clusters of which the one with a lower mean value (darker) is assigned to the glottal area and the other to the remaining regions of the larynx image.

In videos taken from patients with vocal folds nodules there may occur a circumstance in which the glottal area is separated into two regions (see an image of vocal fold with nodules in Fig. 3b).

Further, the method is improved by applying an adaptive thresholding method for subregions of the ROI that contain the glottal area. Firstly an ellipse is fitted to the so far detected area that is based on a method which automatically sets a reference line along the hitherto designated area. Ellipse fitting is achieved by applying a built in OpenCV library function named “fitEllipse” [13]. After defining the center point of the fitted ellipse and its angle, a reference semi-major axis line is determined.

During the videostroboscopic examination of the larynx there is a continuous movement of the camera versus the glottis. An efficient way to eliminate unwanted motion from the video is to move each ROI of the image by a vector calculated for each video frame. ROI displacement vector can be determined by computing the matrix of correlation coefficients between adjacent image frames in the videostroboscopic film. In all subsequent image frames the detected glottal area is shifted to this reference location by the estimated displacement vector. The criterion of the maximum of the two-dimensional correlation function is applied for the best positioning of the consecutive image frames versus the first reference frame. By these means, in the preprocessed image sequence, the glottal area is positioned in a fixed location which significantly facilitates further image analysis procedures. The cross-correlation of adjacent images was determined by using a built-in OpenCV library function which calculates the matrix of correlation coefficients [14]:

$$r_{x,y}=\frac{\sum_{x,y}[P(x',y')-\overline{P}]\cdot[N(x+x',y+y')-\overline{N}]}{\sqrt{\sum_{x,y}[P(x',y')-\overline{P}]^2\cdot\sum_{x,y}[N(x+x',y+y')-\overline{N}]^2}}$$

(3)

where:

- $r_{x,y}$ - correlation matrix,
- $P(x,y)$ - ROI from the current video frame,
- $\overline{P}$ - mean value of pixel intensities in an image,
- $N(x,y)$ - next video frame image,
- $\overline{N}$ - mean value of pixel intensities in an image.

The correlation matrix indicates the level of the linear relationship between $P$ corresponding pixels in images and $N$.

After applying the multi-stage segmentation methods (the watershed algorithm followed by the *K*-means and adaptive thresholding) and removing image distortions due to movement of vocal folds the GAW and the glottovibrogram can be built (see Figs. 4, 5 and a more detailed explanation in Fig. 8).

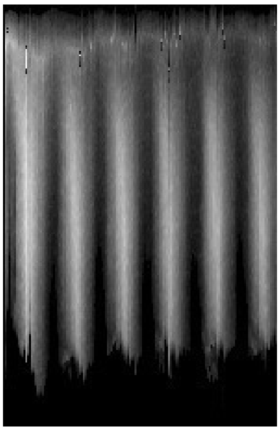


Fig 5. Glottovibrogram of the vibrating vocal folds of a healthy individual

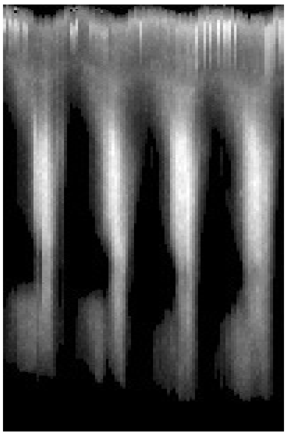


Fig 4. Glottovibrogram of the vibrating vocal folds of the patient with diagnosed vocal nodules

Such a representation termed, the glottovibrogram is a space-time image of the video sequence presenting vocal folds movements. The horizontal axis in the glottovibrogram is time (*t*) and the vertical axis represents the level (*l*) along the glottis (from the anterior commissure – bottom of the image to the posterior commissure – top of the image); each pixel value *g*(*t, l*), i.e. its brightness in the glottovibrogram, represents the value of width *g* of the space between the vocal folds (glottal gap) computed at level *l* of the glottis for time instance *t*. The GAW illustrates time changes of the instantaneous values of the glottal area.

A single phonatory cycle shown in the glottovibrogram or the associated GAW corresponds on average from 20 to 30 images taken by the stroboscopic camera. It is essential to gather at least 20 images for one cycle to determine the kinetic parameters of the vocal folds (opening, closing time and their ratios). Approximately 4 cycles are analyzed for each video sequence, so the glottovibrogram has an average length of 100 pixels. The stroboscopic films are recorded with a sampling rate of 25 frames per second, so the average length of the recording is about 4 seconds. The fundamental frequency of the vocal fold oscillation is provided by the videostroboscopic apparatus. This information enables to correctly scale the frequency axis of the Fourier transform of the glottovibrogram and visualize oscillation frequencies of the vibrating vocal folds (Fig. 6-7).

Thanks to the strobing frequency the information obtained from the videostroboscopic camera we can quantitatively estimate the frequencies taking part in the phonation. It is noteworthy that we analyze “virtual frequencies” which are aliases of the true frequencies of vocal folds' vibration, typically in the range of 150Hz (for men) and 250Hz (for women), reaching peak values of 450 Hz for singers.

The Fourier transformation of the glottovibrogram was calculated by applying the following equation:

$$G(s,l)=\frac{1}{N}\sum_{l=0}^{N-1} g(t,l)e^{-j\frac{2\pi l}{N}s} \qquad (4)$$

$$|G(s,l)|=\sqrt{\Re^2(G(s,l))+\Im^2(G(s,l))} \qquad (5)$$

where:

- $g(t,l)$ is the $l-th$ row of the glottovibrogram (y-axis, i.e. glottis level), $t$ determines the column (x-axis, i.e. discrete time),

- $G(s,l)$ is the Fourier transformed glottovibrogram, determining the frequencies,

- $N$ is the length of the glottovibrogram,

- $j$ is the imaginary unit.

After calculating the magnitudes of the transformation Eq. (5) the phoniatrist is able to gain access to the distribution of frequencies taking part of the vocal folds phonation. In case of the vocal nodules the pathology manifests itself in the form of a second harmonic frequency, significantly blurring the first amplitude peak.

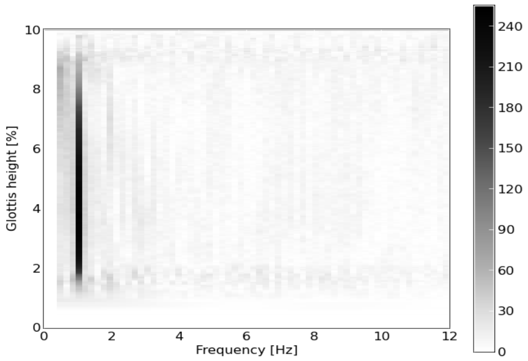


Fig 6. Fourier amplitude spectrum of the glottovibrogram of a healthy patient

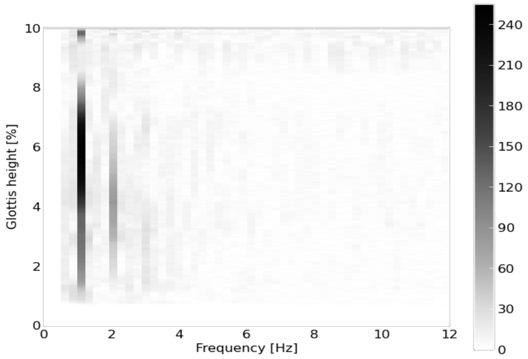


Fig 7. Fourier amplitude spectrum of the glottovibrogram of a patient with vocal nodules

## IV. Phonatory indices

Even after the correct segmentation and transformation process there is still an ambiguity in the relevance of different visualizations representing the phonation process. In this paper we propose a number of indices for quantifying the phonatory processes of the healthy vocal folds and folds with diagnosed nodules.

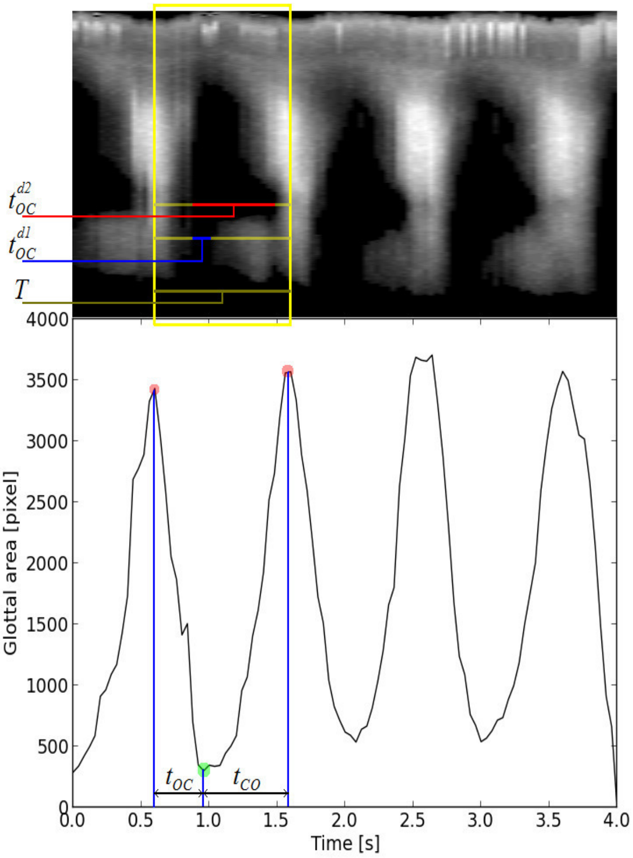


Fig 8. The glottovibrogram (upper panel) and the glottal area waveform (lower panel) with the indicated time intervals used in calculations of the phonatory indices (eq. (6, 7))

Table I.

Comparison of phonatory indices computed for the two studied groups of patients

| | d1 | d2 | d3 | CI | SI |
|---|---|---|---|---|---|
| Control group | 0.00 ± 0.00 | 0.88 ± 0.99 | 2.25 ± 2.65 | 0.10 ± 0.09 | -0.08 ± 0.30 |
| Vocal nodules | 2.10 ± 2.10 | 1.85 ± 1.96 | 4.42 ± 2.72 | -0.48 ± 0.26 | -0.14 ± 0.13 |
| Significance P (ANOVA) | 0.000 | 0.000 | 0.026 | 0.000 | 0.390 |

The *CI* (the closing index) is defined as follows (see also Fig. 8):

$$CI=\frac{t_{OC}^{d2}-t_{OC}^{d1}}{T} \tag{6}$$

where:

$T$ – total cycle time,

$t_{OC}^{d2}$ – is the closed time for the glottis gap measured at the level equal to 50% of the total length in case of normophonic patients, in case of patients with vocal nodules the value is measured for the level of the maximal closure (apparently the place where the vocal nodules occur),

$t_{OC}^{d1}$ – is the closed time of the glottis gap measured at 25% level of the glottis length for normophonic patients, and at the level placed between the maximal closure point and the posterior part of the glottis.

The speed index *SI* is defined by the following equation:

$$SI=\frac{t_{CO}-t_{OC}}{t_O} \tag{7}$$

where:

$t_O$ – opened interval,

$t_{CO}$ – vocal fold closed phase,

$t_{OC}$ – vocal fold closing phase.

In Table I the computed values of the indices proposed in this study are listed, where *d1*, *d2*, *d3* are the levels along the glottis length at which the indices were computed. Location of these levels are at the particular percentages of the total glottis length, correspondingly: 25%, 50% and 75%. The *SI* value in most patients assumes negative values indicating that the opening time of the folds is shorter than the closing time. However, this index yielded poor statistical significance values ($p < 0.36$ from the ANOVA analysis of variance) to use it as a differentiation index for the pathology in question. On the other hand, the computed values of the *CI* index clearly differentiate the patients with diagnosed nodules from the healthy individuals ($p < 0.026$) The *SI* value in most patients reaches negative values which indicates that the opening time is slower than the closing time. Comparison of the calculated parameters reveal that there are pronounced between-group differences and comparable within-group values.

## V. Conclusion

The main purpose of this work is to support the phoniatrist in diagnosis of the vocal folds. The computer image processing methods applied to laryngovideostroboscopic images allow for quantitative analysis of the vocal folds vibrations during phonation. It was shown that the proposed *CI* parameter proved to be viable in differentiation and quantification of the studied glottis pathology.

The program, although, in an early development stage allows for automation of the analysis process of the laryngovideostroboscopic films. Larger scale trials are required before a wider introduction of these computed image analysis techniques into the clinical practice.

## References

[1] B. Kopczyński, P. Strumiłło, E. Niebudek-Bogusz. Assessment of vocal folds phonation by means of computer analysis of laryngovideostroboscopic images – a pilot study. *Otorynolaryngologia – przegląd kliniczny* (in Polish), vol. 13, no. 3 2014 pp. 139–146.

[2] P. Woo. Stroboscopy. Plural Publishing, United Kingdom 2010.

[3] T. Wittenberg, U. Eysholdt. Estimation of Vocal Fold Vibrations Using Image Segmentation, Mustererkennung 1995, pp. 145-152.

[4] S-Z. Karakozoglou, N. Henrich, C. d'Alessandro, Y Stylianou. A Segmentation Scheme Based on Rayleigh Distribution Model for Extracting Glottal Waveform from High-speed Laryngeal Images.

[5] S. Z. Karakozoglou, N. Henrich, C. d'Alessandro, Y. Stylianou. Automatic glottal segmentation using local-based active contours and application to glottovibrography, Speech Communication 2012, 54: pp. 641-654.

[6] A. Méndez, E.M.Ismaili Alaoui, B. Garcia, E. Ibn-Elhaj, I. Ruiz, Glottal space segmentation from motion estimation and gabor filtering. 31st Annual International Conference of the IEEE EMBS. Minneapolis, Minnesota, USA, September 2-6, 2009.

[7] O. Gloger, B. Lehnert, A. Schrade, H. Volzke. Fully Automated Glottis Segmentation in Endoscopic Videos Using Local Color and Shape Features of Glottal Regions. IEEE Trans. Biomed. Eng. 2015, pp. 795-806.

[8] Information and Communication Technologies (WICT), 2011 World Congress on 11-14 Dec. 2011, pp. 313 - 318.

[9] Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on 30-31 March 2012 Nagapattinam, Tamil Nadu 161 - 166.

[10] V. Osma-Ruiz, JI Godino-Llorente, N. Saenz-Lechon, R. Fraile Segmentation of the glottal space from laryngeal images using the watershed transform. Comput Med Imaging Graph. 2008; 32(3): pp. 193-201.

[11] J.M. Gutierrez-Arriola, V. Osma-Ruiz, N. Saenz-Lechon, J.I. Godino-Llorente, R. Fraile, J.D. Arias-Londono, Segmentation of the glottal space from laryngeal images using the watershed transform. *Computerized Medical Imaging and Graphics*, 2008 pp.193-201.

[12] L. Dongju Liu, J. Yu. Otsu Method and K-means. Hybrid Intelligent Systems, 2009. HIS '09. Ninth International Conference, 12-14 Aug. 2009, pp. 344-349.

[13] A. W. Fitzgibbon, R.B. Fisher. A Buyer's Guide to Conic Fitting. Proc.5th British Machine Vision Conference, Birmingham, 1995, pp. 513-522.

[14] OpenCV image processing library: http://docs.opencv.org, accessed May 2015.

DOI: 10.15439/2015F304
ACSIS, Vol. 5

# Utilization of Single Whole Body Vibration Training Unit in Rehabilitation of Elderly Patients with Neurological Disorders

Jaroslav Majerník
Pavol Jozef Šafárik University in Košice, Trieda SNP 1, 040 11 Košice, Slovakia
Technical University of Ostrava, 17. listopadu 15, Ostrava – Poruba, Czech Republic
Email: jaroslav.majernik@upjs.sk

Miriam Dziaková
Louis Pasteur University Hospital in Košice, Trieda SNP 1, 040 11 Košice, Slovakia
Email: miriam.dziakova@unlp.sk

Jozef Živčák
Technical University in Košice, Letná 9, 042 00 Košice, Slovakia
Email: jozef.zivcak@tuke.sk

***Abstract*—Whole body vibration (WBV) techniques are increasingly applied in rehabilitation processes to improve patients' functional mobility. Therapeutic usage of WBV reported enhanced muscular strength, power or even bone density. Therefore, the purpose of our study was to investigate immediate response of single WBV training unit on mobility of elderly patients with neurological disorders. Nineteen patients (66.74 ± 3.65 years, 6 males and 13 females) were assessed before WBV (10 min, 30 Hz, vertical 2 mm) and then 1 min afterwards. Individual changes in gait kinematics indicated positive effects of WBV, while kinematics of the gait was more symmetric considering right and left side. Individually, depending on the disease's severity, these changes were more or less significant. The improvements in gait kinematics convinced us that WBV can be carefully used in patients' therapy and it may be used together with individually planed rehabilitation processes to bring more satisfying results.**

## I. Introduction

Ageing of the population brings growing demands on social and health care systems due to the increased risks of various chronic diseases in older individuals. Health related quality of human life and loss of independence can be also affected by natural weakening of human body and its functional systems. This leads to more frequent falls and consecutive health and mobility problems [1-3].

Here, the disorders like sensory loss, vestibular dysfunction, impaired vision, muscular weakness, bone rarefaction or gait disorders are multi-factorial disorders that contribute to vulnerability and frailty of elderly persons. Moreover, elderly falls are usually reasons for additional medical interventions. Motor impairment syndromes are also associated with neurological disorders [4]. This group of clinically heterogeneous diseases causes severe problems in coordination, gait, balance, voluntary muscle control, power and strength [5, 6].

Strategies to prevent negative consequences of functional mobility diseases have to be widely discussed, assessed and applied in health care of elderly patients. Such strategies include also rehabilitation techniques and physical exercises. Here, due to the published results the WBV can be considered as reliable and effective tool in rehabilitation and sport medicine [7, 8].

Continuing in our previous pilot study [9] we hypothesize that the WBV may stimulate muscle activation in elderly patient with neurological disorders and that application of WBV will result in improvement of quality of their gait. Therefore, the purpose of this study was to investigate whether the single WBV session has any positive effects on gait kinematics.

## II. Whole Body vibration

Whole body vibration (WBV) has been extensively studied for its dangerous effects on humans, especially when exposed as occupational vibration at high amplitudes and specific frequencies [10-12]. However, WBV is also the concept that was already applied in many studies to confirm benefits for astronauts, athletes, wellness of healthy population, but also patients with various diseases [13-15].

Recent clinical works suggest that low amplitude and low frequency of mechanical stimulation of the human body is a safe and effective way to exercise musculoskeletal structures. The studies realized during past decade indicate that WBV may increase muscle strength, neuromuscular function, bone mass and mineral density [16-19], can be useful in improving physical capacity, cardiorespiratory functions, hormonal production, proprioception, and balance [20-22].

Despite of WBV positive effects presented in almost all related research studies, the authors interpret their results with caution. Also, the underlying mechanisms by which WBV enhance neuromuscular performance vary between studies and are still unclear. Inconsistency in presented results is caused by various training protocols and heterogeneity in study designs. However, several mechanisms should be considered as crucial. Vibration stimulates skeletal muscle by activating a stretch reflex analogous to the tonic vibration reflex. Thus the activated muscle spindle activity causes involuntary reflexive muscle contraction in an attempt to control the vibration imposed muscle length change. WBV also evoke changes in postural control strategy where the vibration stimulus generates postural instability and the body activates particular muscles to maintain body balance. Another mechanism is generated

978-83-60810-66-8/$25.00©2015, IEEE

when oscillations induce a muscle tuning response in an attempt to minimize the propagation of the vibrations throughout the body. Then, the muscles might activate to prevent further transmission of the vibration stimulus through the body.

There are various vibrating platforms available as commercial products used by many wellness centers and rehabilitation clinics. The main differences are in the type of vibration stimulation, frequency range, and amplitude of vibration [23, 24]. Most of the vibrating platforms vibrate sinusoidally. The subjects stand on the platform which oscillates only vertically or side alternating. Other types of vibrating devices uses two separate platforms to produce vibrations for each foot independently. While commercial usage promotes WBV as attractive and effective alternative to resistance training, the therapeutic applications should carefully consider duration of WBV exposure, main vibration characteristics as well as their effects on neuromuscular system respecting patients' physical capabilities. The rest periods between vibration trainings may also play a significant role in final WBV effects. The effect of vibration may be also tested using various modeling techniques [25].

## III. Material and Methods

### A. Patients

A group of nineteen elderly patients (age 66.74 ± 3.65 years, 6 males and 13 females) with various neurological disorders were included in the study. The inclusion criteria were ambulatory patients that are able to walk independently, had no cardiovascular disease or epilepsy, and had no prior experience in WBV training. All the patients were informed about the WBV training, about the tests to be realized and also about possible risks and benefits of the research. Prior to participation they gave written informed consent approved together with the study design by the University Hospital Ethics Committee.

All participants attended a familiarization session before the study and before the tests were performed. No other physical treatment or intervention was realized at least 24 hour before WBV training session. The anthropometric measures were also taken and registered in patients' experimental protocols. As for the aim of the study, the anthropometric characteristics of lower extremities were preferred, including thigh length (right: 44.89 ± 2.26 cm, left: 44.42 ± 2.14 cm), calf length (right: 41.89 ± 3.80 cm, left: 41.58 ± 3.91 cm) and foot length (right: 25.74 ± 2.82 cm, left: 25.89 ± 2.55 cm).

### B. Experiment

The experimental protocol was designed to discover potential immediate response of a single WBV training unit to the quality of gait kinematics in elderly patients. Training sessions were supervised by rehabilitation specialist and measurements as well as WBV sessions were conducted in the same thermally neutral room intended for physical training. All subjects did not engage in any therapeutic or rehabilitation procedures before testing.

The training session started with physical examination and short warming-up walk. Then, the patient's gait was captured and analyzed before WBV exposure. Participants were asked to walk at their natural waking speed along the 6 m long path. After reaching the end point of the path, they were asked to turn back (180°), i.e. to change the direction of gait, and to walk back to the starting point. Then, they turned back again and walked to the end point of the path, where the last turn back was realized and the patients finished walking in starting point of the path. In that sense, the subjects passed the length of walking path four times. WBV session followed one minute after this control gait was captured and analyzed. Here, each participant stood in static position on the vibration platform (VibroGym inSPORTline) with no shoes and socks and holding on the device handle. Erected posture with slightly bended knees was required during vibration test. The patients were asked to stop the training in the case of any pain responses to vibration. Duration of one WBV training unit was set to 10 min. Sinusoidal vertical vibration frequency was set to 30 Hz with amplitude of 2 mm. 1 min rest interval followed after this WBV exposure. Then, the patients walked again and the kinematics of their gait was captured and analyzed in the same way as it was before WBV session. Finally, the kinematics obtained before and after WBV training unit was compared to evaluate its effects on patients' gait.

### C. Parameters

Gait assessment was performed using our marker-free motion analysis system MAFRAN [26]. The system is based on motion tracking method in video sequence with no passive or active markers attached to the patients' bodies tries to minimize disadvantages of currently available marker–free systems, to bring new of motion analysis and to offer clinicians cheaper, but not worse alternative preserving all advantages of motion analysis.

Here, the patient's gait is captured in sagittal plane using common commercial video camera. Then, the raw record is used in the system to automatically reconstruct motion trajectories of human body anatomical landmarks, i.e. the trajectories of all lower extremity joints and adjacent segments. These trajectories are consequently used to calculate other kinematical parameters for detailed description of patient's gait. Here analyzed parameters include positions, velocities and accelerations of individual joints, hip flexion/extension, knee flexion/extension, and ankle plantar/dorsal flexion angles, gait cycle length, gait cycle time, gait cycle velocity, cadence (cycles per minute), stance phase and swing phase of the gait cycle.

### D. Statistics

All acquired kinematical parameters were analyzed individually within the patient and then within the group of here included patients as well. The kinematical characteristics of the patients were evaluated as differences between right and left side. The hypothesis was based on assumption that these differences should be smaller after WBV training comparing values obtained before WBV training. Otherwise, the WBV will probably have no immediate benefits for gait kinematics. Statistical methods included descriptive statistics and Student's paired t-test and were used to ascertain specific and significant differences. The significance level was set to P < 0.05.

## IV. Results

WBV session was well accepted by all elderly patients included in this continuing study. No one felt any pain or expressed any problems during WBV exposure. A first analysis was performed with anatomical joint angles of lower extremities in sagittal plane. Individually, no of the patients had the same curves of all tree joints comparing before and after WBV values. At least one joint's trajectory was changed either in positive or negative direction. An example of anatomical joint angles changes in 65 years old male patient with right hemiparesis is shown on following figures.

Figure 1 shows hip flexion/extension angle before and after single WBV training unit. A paired t-test determined that the mean decrease of differences (M=0.47, SD=4,29, N=51) was not significantly greater than zero, t(50)=0.78, two-tail p=0.441 (95% CL=1.208), providing evidence that the WBV was not effective in reduction of differences between right and left hip flexion/extension angle.

Figure 2 shows knee flexion/extension angle before and after single WBV training unit. Knee flexion/extension angle of the same patient showed that the mean differences between right and left side before and after WBV were decreased, but the mean decrease of differences (M=0.54, SD=8,207, N=51) was not significantly greater than zero, t(50)=0,47, two-tail p=0.643 (95% CL=2.308), providing evidence that the WBV was not effective in reduction of differences between right and left knee flexion/extension angle in this patient.

Figure 3 shows ankle plantar/dorsal flexion angle before and after single WBV training unit. The mean decrease of the right and the left side differences in ankle plantar/dorsal flexion angle of the same 65 years old male patient (M=3.79, SD=4,089, N=51) was significantly greater than zero, t(50)=6.63, two-tail p=0.000 (95% CL=1.150) provided evidence that the WBV was effective in reduction of differences between right and left ankle plantar/dorsal flexion angle.

Anatomical joint angles were analyzed in all participants of this study in the same way. The summary of right and lefts side differences in anatomical joint angles in elderly patients showed that the most significant changes were registered in the ankle plantar/dorsal flexion angle (89.47%), followed by hip flexion/extension angle (73.68%) and knee flexion/extension angle (57.89%). However, these significant changes include both the positive and the negative changes. The only significant positive changes were chiefly in hip flexion/extension angle (63.16%) followed by the knee flexion/extension angle (47.37%) and ankle plantar/dorsal flexion angle (47.37%).

Individually, there were thirteen patients (68.42%) who had at least two significant positive changes of these kinematical parameters or they had no significant negative

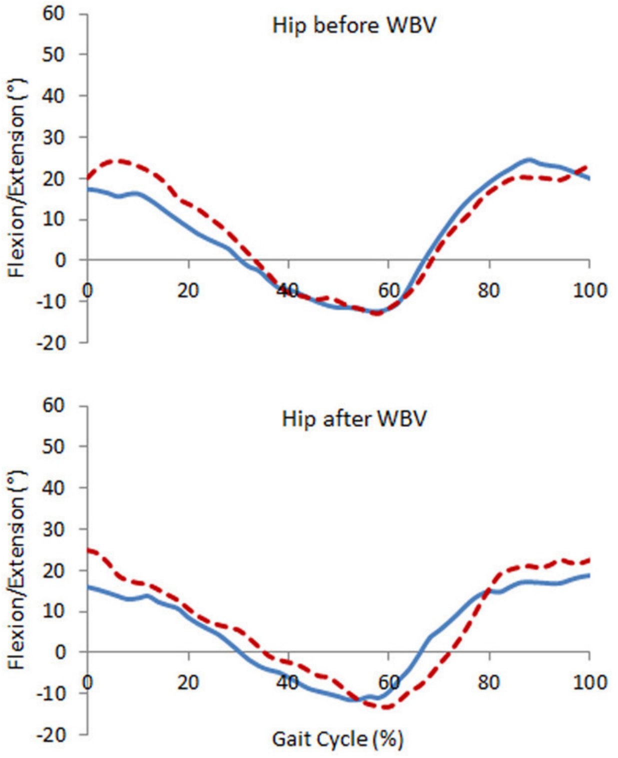


Fig. 1 Hip flexion/extension angle in 65 years old male patient with right hemiparesis before and after WBV exposure (solid line – right leg, dashed line – left leg).

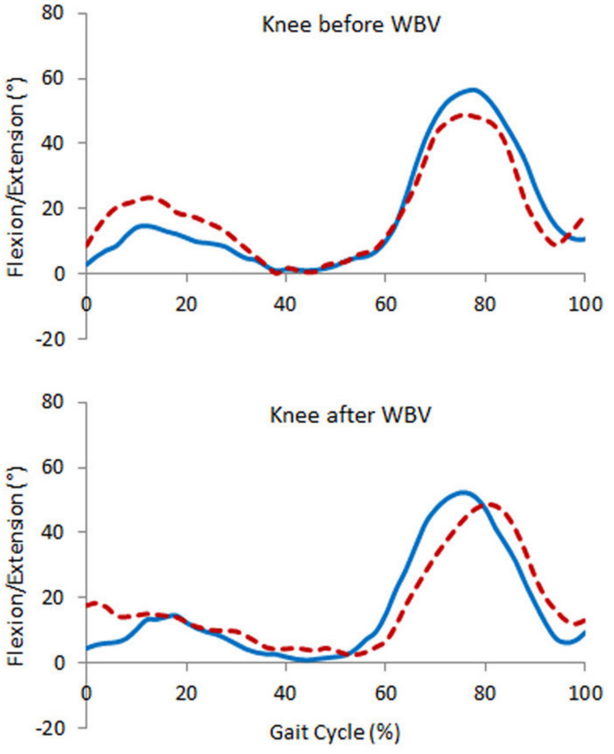


Fig. 2 Knee flexion/extension angle in 65 years old male patient with right hemiparesis before and after WBV exposure (solid line – right leg, dashed line – left leg).

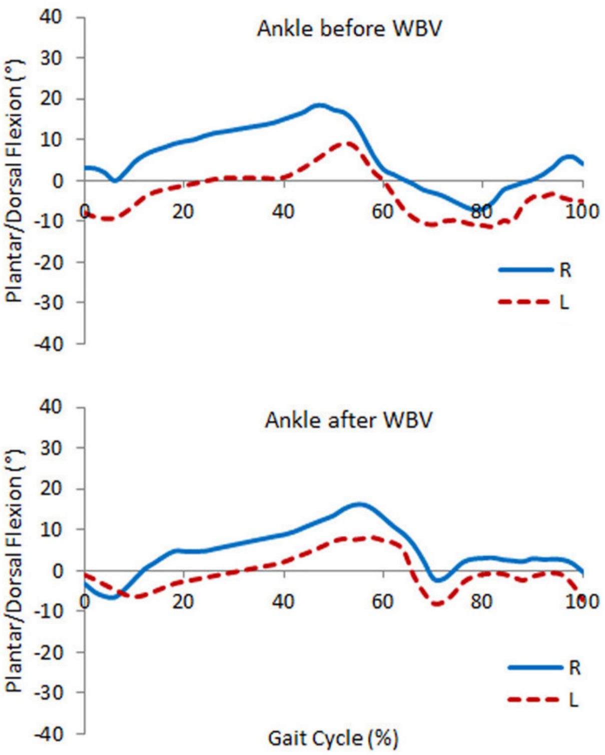


Fig. 3 Ankle plantar/dorsal flexion angle in 65 years old male patient with right hemiparesis before and after WBV exposure (solid line – right leg, dashed line – left leg).

changes (subjects 1, 2, 3, 5, 8, 9, 10, 13, 14, 15, 17, 18 and 19). Two patients (10.53%) had no beneficial improvements resulting from applied WBV exposure (subjects 6 and 12) and four patients (21.05%) of the elderly patient subgroup registered worsening because of no significantly positive or only significantly negative changes (subjects 4, 7, 11 and 16).

The second analysis was performed in spatial-temporal parameters. Here, the symmetry of all characteristics was examined and summarized. The mean differences between right and left sides and standard deviations of these parameters are listed in the table 1.

No of here analyzed parameters had significantly neither positive nor negative changes. Nevertheless, some of the parameters had positive and another negative tendency. The positive trends were shown in decreasing differences between right and left side in gait cycle length (1.470 ± 6.417) and velocity (0.016 ± 0.056). On the other side small increase was registered in cadence (-0.064 ± 1.507) and gait cycle stance (-0.885 ± 4.951 and swing (-0.907 ± 4.932) phase. Gait cycle time remained almost unchanged (0.002 ± 0.043).

## V. Conclusion

In this study, the effect of single WBV training unit was tested in elderly patients with neurological disorders including multiple sclerosis, Parkinson disease, cerebral palsy and radiculoneuritis. The significant improvement was confirmed mainly in anatomical joint angles rather than in spatial-temporal parameters. As we expected, the single whole body vibration training unit had only short-time effect as six of the beneficial patients were analyzed one week after the experiment and their kinematics was similar to the pre- exposure status. Utilization of WBV training in patients with neurological disorders may result in benefits for kinematics of human motion, but its significance and mechanism still remains unclear and undiscovered.

The results of this pilot study provided invaluable data for rehabilitation specialists as well as for development of further research programs in physiotherapy. Based on the results we obtained, it was also confirmed that an individual WBV training and supervised functional treatment should be specified for particular patient. Further research should be realized to clarify WBV specific benefits. The various training approaches including standing in different static positions, sitting on chair with legs on vibrating platform or performing exercises on the platform during therapy sessions should be investigated as well.

## Acknowledgment

This work has been elaborated in the framework of the project ”Support research and development in the Moravian–Silesian Region 2014 DT 1 – International research teams” (RRC/07/2014), financed from the budget of the Moravian–Silesian Region (50%) and in the framework of the project KEGA 020UK-4/2014 (50%).

Table I.
Spatial-temporal characteristics of patients' gait obtained before (Pre) and after (Post) WBV exposure (N=19).

| Parameter | Pre WBV | Post WBV | Delta pre/post |
|---|---|---|---|
| GC length (cm) | 5.751 ± 4.170 | 4.281 ± 4.708 | 1.470 ± 6.417 |
| GC time (s) | 0.086 ± 0.075 | 0.084 ± 0.077 | 0.002 ±0.043 |
| GC velocity (m/s) | 0.067 ± 0.038 | 0.051 ± 0.048 | 0.016 ± 0.056 |
| Cadence (GC/min) | 2.014 ± 1.284 | 2.078 ± 1.286 | -0.064 ± 1.507 |
| Stance phase (%) | 3.556 ± 2.944 | 4.441 ± 4.984 | -0.885 ± 4.951 |
| Swing phase (%) | 3.562 ± 2.942 | 4.469 ± 4.963 | -0.907 ± 4.932 |

## References

[1] S. Rogan, R. Hilfiker, K. Herren, L. Radlinger, E.D. De Bruin, “Effects of whole-body vibration on postural control in elderly: A systematic review and meta-analysis”, *BMC Geriatric*, 2011, 11, art. no. 72, DOI: 10.1186/1471-2318-11-72.

[2] M. Sitjà-Rabert, M.J. Martínez-Zapata, A. Fort-Vanmeerhaeghe, F. Rey-Abella, D. Romero-Rodríguez, X. Bonfill, X. “Whole body vibration for older persons: an open randomized, multicentre, parallel, clinical trial”, *BMC geriatrics*, 2011, 11, p.89, DOI: 10.1186/1471-2318-11-89.

[3] N. Shibata, K. Ishimatsu, S. Maeda, S. “Gender difference in subjective response to whole-body vibration under standing posture”, *International Archives of Occupational and Environmental Health*, 2012, 85 (2), pp. 171-179, DOI: 10.1007/s00420-011-0657-0.

[4] B.D. Pozo-Cruz, J.C. Adsuar, J.A. Parraca, J.D. Pozo-Cruz, P.R. Olivares, N. Gusi, N. “Using whole-body vibration training in patients affected with common neurological diseases: A systematic literature review”, *Journal of Alternative and Complementary Medicine*, 2012, 18 (1), pp. 29-41, DOI: 10.1089/acm.2010.0691.

[5] R.W.K. Lau, T. Teo, F. Yu, R.C.K. Chung, M.Y.C. Pang, “Effects of whole-body vibration on sensorimotor performance in people with parkinson disease: A systematic review”, *Physical Therapy*, 2011, 91 (2), pp. 198-209, DOI: 10.2522/ptj.20100071.

[6] R.D. Pollock, S. Provan, F.C. Martin, D.J. Newham, “The effects of whole body vibration on balance, joint position sense and cutaneous sensation”, *European Journal of Applied Physiology*, 2011, 111 (12), pp. 3069-3077, DOI: 10.1007/s00421-011-1943-y.

[7] S.A. Moussavi-Najarkola, A. Khavanin, R. Mirzaei, M. Salehnia, M. Akbari, “Effects of whole body vibration on outer hair cells' hearing response to distortion product otoacoustic emissions”, *In Vitro Cellular and Developmental Biology - Animal*, 2012, 48 (5), pp. 276-283, DOI: 10.1007/s11626-012-9490-3.

[8] T. Furness, N. Bate, L. Welsh, G. Naughton, C. Lorenzen, C. “Efficacy of a whole-body vibration intervention to effect exercise tolerance and functional performance of the lower limbs of people with chronic obstructive pulmonary disease”, *BMC Pulmonary Medicine*, 2012, 12, art. no. 71, DOI: 10.1186/1471-2466-12-71.

[9] J. Majernik, J. Zivcak, “Effect of Whole Body Vibration on Functional Mobility in Elderly Patients”, *Acta Mechanica Slovaca*, Vol. 17 (3), 2013, ISSN 1335-2393, pp. 64-69.

[10] R. Mani, S. Milosavljevic, S. J. Sullivan, “The effect of occupational whole-body vibration on standing balance: A systematic review”, *Int. Journal of Industrial Ergonomics*, 2010, 40 (6), pp. 698-709, DOI: 10.1016/j.ergon.2010.05.009.

[11] S. Maeda, N. J. Mansfield, N. Shibata, “Evaluation of subjective responses to whole-body vibration exposure: Effect of frequency content”, *International Journal of Industrial Ergonomics*, 2008, 38 (5-6), pp. 509–515, DOI: 10.1016/j.ergon.2007.08.013.

[12] B. R. Santos, Ch. Lariviere, A. Delisle, A. Plamondon, P.-E. Boileau, D. Imbeau, “A laboratory study to quantify the biomechanical responses to whole-body vibration: The influence on balance, reflex response, muscular activity and fatigue”, *International Journal of Industrial Ergonomics*, 2008, 38 (7-8), pp. 626–639, DOI: 10.1016/j.ergon.2008.01.015.

[13] M. Cardinale, J. Wakeling, “Whole body vibration exercise: are vibrations good for you?”, *Br J Sports Med*, 2005, 39 (9), pp. 585–589, DOI:10.1136/bjsm.2005.016857.

[14] K. H. Madou, J. B. Cronin, “The effects of whole body vibration on physical and physiological capability in special populations”, *Hong Kong Physiotherapy Journal*, 2008, 26, pp. 24-38, DOI: 10.1016/S1013-7025(09)70005-3.

[15] R.. D. Prisby, M. Lafage-Proust, L. Malaval, A. Belli, L. Vico, “Effects of whole body vibration on the skeleton and other organ systems in man and animal models: What we know and what we need to know”, *Ageing Research Reviews*, 2008, 7 (4), pp. 319–329, DOI: 10.1016/j.arr.2008.07.004.

[16] S.F. Baumbach, M. Fasser, H. Polzer, M. Sieb, M., Regauer, W. Mutschler, M. Schieker, M. Blauth, “Study protocol: The effect of whole body vibration on acute unilateral unstable lateral ankle sprain- a biphasic randomized controlled trial”, *BMC Musculoskeletal Disorders*, 2013, p22, DOI: 10.1186/1471-2474-14-22.

[17] L. Slatkovska, S.M.H.. Alibhai, J. Beyene, H. Hu, A. Demaras, A.M. Cheung, “Effect of 12 months of whole-body vibration therapy on bone density and structure in postmenopausal women: A randomized trial”, *Annals of Internal Medicine*, 2011, 155 (10), pp. 668-679, DOI: 10.7326/0003-4819-155-10-201111150-00005.

[18] M. Cerny and M. Penhaker, ”Wireless Body Sensor Network in Health Maintenance Systems“, *Elektronika Ir Elektrotechnika*, 2011, (9), pp. 113-116, DOI: 10.5755/j01.eee.115.9.762.

[19] J.O. Totosy de Zepetnek, L.M. Giangregorio, B.C. Craven, “Whole-body vibration as potential intervention for people with low bone mineral density and osteoporosis: A review”, *Journal of Rehabilitation Research and Development*, 2009, 46 (4), pp. 529-542, DOI: 10.1682/JRRD.2008.09.0136.

[20] D. Macura, A. Macurova, A. “Bounded solutions of the nonlinear differential systems”, *International Journal of Pure and Applied Mathematics*, 2011, 70 (5), pp. 755-760.

[21] M. Claerbout, B. Gebara, S. Ilsbroukx, S. Verschueren, K. Peers, P. Van Asch, P. Feys, “Effects of 3 weeks' whole body vibration training on muscle strength and functional mobility in hospitalized persons with multiple sclerosis”, *Multiple Sclerosis*, 2012, 18 (4), pp. 498-505, DOI: 10.1177/1352458511423267.

[22] E.G. Artero, J.C. Espada-Fuentes, J. Argüelles-Cienfuegos, A. Román, P.J. Gómez-López, A. Gutiérrez, “Effect of whole-body vibration and resistance training on knee extensors muscular performance”, *European Journal of Applied Physiology*, 2012, 112 (4), pp. 1371-1378, DOI: 10.1007/s00421-011-2091-0.

[23] V. Kasik, M. Penhaker, V. Novak, R. Bridzik, J. Krawiec, “User Interactive Biomedical Data Web Services Application”, *E-Technologies and Networks for Development*. vol. 171, J. J. Yonazi, E. Sedoyeka, E. Ariwa, and E. ElQawasmeh, Eds., ed, 2011, pp. 223-237, DOI: 10.1007/978-3-642-22729-5_19.

[24] C. Milanese, F. Piscitelli, C. Simoni, R. Pugliarello, C. Zancanaro, “Effects of whole-body vibration with or without localized radiofrequency on anthropometry, body composition, and motor performance in young nonobese women”, *Journal of Alternative and Complementary Medicine*, 2012, 18 (1), pp. 69-75, DOI: 10.1089/acm.2010.0324.

[25] A. Sonza, Ch. Maurer, M. Achaval, M. A. Zaro, B. M. Nigg, “Human cutaneous sensors on the sole of the foot: Altered sensitivity and recovery time after whole body vibration”, *Neuroscience Letters*, 2013, 533 (1), pp. 81– 85, DOI: 10.1016/j.neulet.2012.11.036.

[26] J. Majerník, “Reconstruction of Human Motion Trajectories to Support Human Gait Analysis in Free Moving Subjects”, In: Pancerz,K. and Zaitseva,E.: Computational Intelligence, Medicine and Biology; Studies in Computational Intelligence, 2015, 600, pp. 57-77, DOI: 10.1007/978-3-319-16844-9_4.

DOI: 10.15439/2015F361
ACSIS, Vol. 5

# Medical diagnosis support and accuracy improvement by application of total scoring from feature selection approach

Wiesław Paja
Faculty of Mathematics and Natural Sciences, University of Rzeszów,
1 Prof. S. Pigonia Street, 35-310 Rzeszów, Poland
Email: wpaja@ur.edu.pl

*Abstract*— **Melanoma is the most deadly form of skin cancer. Early detection and successful treatment of this disease often is possible. The main goal of this paper is to present results of application of feature selection method to find the most important or all important features that characterize melanocytic spots on the skin and in this way defining of a new Total Dermatoscopy Score formula. Thus, it is possible to decrease dimensionality of that problem. Results gathered during research focus on about six from thirteen descriptive attributes which are the most relevant and are stated as core attributes. Based on these attributes a simple total scoring method could be applied to improve prediction (diagnosis) results, additionally also reducing complexity of problem. Results were acquired by application of six different machine learning algorithms and estimated using several evaluation measures.**

## I. Introduction

Melanoma is the most deadly form of skin cancer. The World Health Organization estimates that more than 65000 people a year worldwide die from too much sun, mostly from malignant skin cancer [1]. It is an increasingly common tumour, it is  the cutaneous tumour with the worst prognosis and its incidence is growing, because most melanomas arise on areas of skin that can be easily examined. Early detection and successful treatment often is possible, most dermatologists can accurately diagnose melanoma in about 80% of cases according to well-known ABCD process [2]. ABCD formula devotes to Asymmetry of lesions, their Border, Color and Diversity of structures (or Diameter in other approach), and in some cases the Evolving over time (the ABCDE formula). Based on these features dermatologists could prepare diagnosis by simply observation of investigated lesion.

Meanwhile the incorporation of dermatoscopic techniques, reflectance confocal microscopy and multispectral digital dermatoscopy have greatly enhanced the diagnosis of this cutaneous melanoma. While these devices and techniques could give dermatologists a closer look at suspicious skin lesions. This, in turn, can help dermatologists find suspicious lesions earlier than before and better determine whether a biopsy is needed. None of these devices can confirm that a suspicious lesion is melanoma. It is, however, not yet possible to tell if a patient has melanoma or any type of skin cancer without a biopsy. It is important to combine the classically ABCDs and biopsy to prevention and diagnosis of melanoma.

The five-year survival rate for people whose melanoma is detected and treated before it spreads to the lymph nodes is 99 percent. Five-year survival rates for regional and distant stage melanomas are 65 percent and 15 percent, respectively [3]. Thus the curability of this type of skin cancer depends essentially on its early diagnosis and excision. For that reason the ABCD (asymmetry, border, color and diversity of structure) clinical rule is commonly used by dermatologists in visual examination and detection of early melanoma. It is also used in development of diagnosis platforms such as DERMA [4] or IMDLS systems [5].

Previous research [5-8] focused on using of data mining and image mining techniques to provide early support to diagnosis of melanocytic lesions. Now, it is proposed to apply feature selection methods to find interesting features inside investigated melanocytic datasets. Thus, we could try to recognize the minimal set of important (relevant) features, but on the other hand we can calculate the importance of each feature used in ABCD formula in the domain of melanoma classification. According to Kohavi and John [9] feature X could be defined to be strongly relevant when removal of X alone from the data always results in deterioration of the prediction accuracy of the ideal Bayes classifier. Feature X is weakly relevant if it is not strongly relevant and there exists a subset of features S, such that the performance of ideal Bayes classifier on S is worse than the performance on S $\cup$ {X}. A feature is irrelevant if it is neither strongly nor weakly relevant. Improving the performance of machine learning classifiers for diagnosis based on feature selection is often applied [10,11]. In this paper additional application of FS methods is investigated.

## II. Dataset used during experiments

The medical dataset which was used in this research concerns melanocytic skin lesions that are a very serious skin and lethal cancer. It is a disease of contemporary time,

This work was partially supported by The Fund of Dean of the Faculty of Mathematics and Natural Sciences, University of Rzeszów

978-83-60810-66-8/$25.00©2015, IEEE

the number of melanoma cases is constantly increasing, due to, among other factors, sun exposure and a thinning layer of ozone over the Earth. Statistical details on this data are given in [12]. Investigated data consist of 326 case of *Benign nevus* and *Blue nevus*, 108 cases of *Suspicious nevus* and 114 cases of *Melanoma malignant*, a total of 548 cases. Descriptive attributes of the data were divided into four categories:

- *Asymmetry*, has three different values: *symmetric spot, one-axial asymmetry* and *two-axial asymmetry,*
- *Border*, is a numerical attribute with values from *0* to *8*,
- *Color* group, has six possible types: *Black, Blue, Dark brown, Light brown, Red* and *White*,
- *Diversity of structures* group, has five possible types: *Pigment dots, Pigment globules, Pigment network, Structureless areas* and *Branched streaks*;

Each of these 11 types of Color and Diversity have values 0 or 1, that is 0 means lack of the corresponding property and 1 means the occurrence of the property. In dermatology this set of features is known as ABCD formula and is also applied to calculated the so-called *Total Dermatoscopy Score (TDS)* [13,14]. The ABCD formula of dermoscopy was the first dermoscopy algorithm created to help differentiate benign from malignant tumors [14]. This algorithm was developed to quantitatively address the crucial question in dermoscopy of whether a melanocytic skin lesion under investigation is benign, suspicious (borderline), or malignant. Based only on four dermatoscopic criteria this method is relatively easy to learn and to apply. The ABCD method has been extensively studied and it has been shown that it improves the diagnostic performance of clinicians evaluating pigmented skin lesions.

The goal was to use selected machine learning methods to estimate hierarchy of importance of melanocytic symptoms. These symptoms are part of well-known parameter *TDS* (*Total Dermatoscopy Score*) that is a useful diagnostic tool for melanoma. The *TDS* is computed using the following formula (known as the ABCD formula):

$$\begin{aligned} TDS &= 1.3 * Asymmetry + 0.1 * Border + \\ &+ 0.5 * \sum Colors + 0.5 * \sum Diversity \end{aligned} \tag{1}$$

where A is a description of lesion’s asymmetry, B is a description of lesion’s border, C is a description of colors appearing in considered lesion, and D is a specification of lesion’s diversity.

## III. Methods of experiments

During research a following general procedure was applied:

*1. Selection of dataset and features for investigation*

*(a) Application of set of ranking measures to calculate rank of importance for each feature*

*(i) With set of contrast features*

*(ii) Without contrast features*

*(b) Definition (selection) of the most important feature subset*

*2. TDS calculation for all original features*

*3. New TDS calculation based on selected most important features*

*4. Application of different machine learning algorithms for classification of unseen objects using 10-fold cross validation method*

*(a) Using all descriptive features*

*(b) Using only selected, most important features*

*(c) Using all descriptive features with TDS added*

*(d) Using only selected, most important features with NewTDS added*

*5. Comparison of gathered results using different evaluation measures*

In the first step, dataset and features for investigation were defined. Then, different ranking measures were applied to estimate importance of each feature. In order to check specificity of the feature selection, the dataset was extended by contrast variables. It means that each original variable was duplicated and it’s values were randomly permuted between all objects. Hence a set of non-informative by design shadow variables was added to original variables. The number times when the shadow variables were selected as important gives estimate of the expected level of false discovery. These variables that were selected as important significantly more often than random, were examined further, using different test. To define level of feature importance six well-known ranking measures were applied: *ReliefF*, *Information Gain*, *Gain Ratio*, *Gini Index*, *SVM weight* and *RandomForest*. Additionally, a new parameter, called *RuleQualityFS* (see Table 1), were introduced. It is based on frequency of presence of different feature in rule model generated from dataset and also takes into consideration quality of the rules in which there is. Rank quality of $i^{th}$ attribute could be presented as follow:

$$Q_{A_i} = \sum_{j=1}^{n} Q_{R_j} \{A_i\} \tag{2}$$

where $n$ is a number of rules inside the model, $Q_{Rj}$ defines classification quality of rule $R_j$ and $\{A_i\}$ describe the presence of $i^{th}$ attribute, usually *0* or *1*.

In turn, quality of rule is defined as follow:

$$Q_{R_j} = \frac{E_{corr}}{E_{corr} + E_{incorr}} \tag{3}$$

where $E_{corr}$ depicts number of correctly matched learning examples by $j^{th}$ rule and $E_{incorr}$ depicts number of incorrectly matched learning examples by this rule.

In the second step, the standard *TDS* calculation were performed based on original values of attributes and using formula (1). It is standard procedure utilized by medical specialists.

However, in my research, the third step is crucial. In this point a *NewTDS* value is defined and calculated (see formula

4). According to acquired factors from the first step of experiments (Table 1), a new formula for *TDS* calculation were introduced:

$$\begin{aligned} NewTDS &= 38.72 * Border + \\ &+ 31.82 * Asymmetry + \\ &+ 23.22 * PigmentNetwork + \\ &+ 20.00 * BlueColor + \\ &+ 16.60 * BranchedStreaks + \\ &+ 15.50 * WhiteColor \end{aligned} \tag{4}$$

Six selected attributes were used according to Table 1. For each of them corresponding factor from this table (*RuleQualityFS* column) were inserted. In this way, new attribute which connects others into one value were added to original dataset.

During the fourth step test probing the importance of variables was performed by analyzing the influence of variables used for model building on the prediction quality. Four different combination of attributes were applied.

Six different machine learning model were applied to build different predictors: *Classification Tree (CT), Random Forest (RF), CN2 decision rules algorithm (CN2), Naïve Bayes (NB), k Nearest Neighbors (kNN)* and Support *Vector Machine (SVM).* During this step a 10-fold cross validation paradigm were used. Ten known evaluation measures were utilized in each predictor: *Classification Accuracy (CA), Sensitivity, Specificity, Area Under ROC curve (AUC), Information Score (IS), F1 score (F1), Precision, Brier measure, Matthew Coefficient Correlation (MCC)* parameter and finally *Informadness ratio* [11].

TABLE I.
RANKING OF FEATURES USING SEVEN DIFFERENT MEASURES

| Attribute | ReliefF | Inf. gain | Gain Ratio | Gini | SVM weight | RF | RuleQuality FS |
|---|---|---|---|---|---|---|---|
| Border | 0.03 | **0.17** | **0.09** | **0.03** | **4.93** | **3.74** | **38.72** |
| Asymmetry | **0.25** | **0.46** | **0.34** | **0.07** | **7.34** | **10.99** | **31.82** |
| Pigment network | **0.19** | **0.18** | **0.18** | **0.02** | **1.90** | **3.82** | **23.22** |
| Blue color | **0.16** | **0.41** | **0.58** | **0.06** | **13.79** | **10.17** | **20.00** |
| Branched streaks | **0.13** | **0.23** | **0.23** | **0.02** | **2.22** | **3.51** | **16.60** |
| White color | 0.03 | **0.06** | **0.07** | **0.01** | **1.64** | **1.12** | **15.50** |
| Border (contrast) | -0.06 | 0.01 | 0.01 | 0.00 | 0.08 | -0.12 | 12.50 |
| Black color | -0.05 | **0.11** | **0.11** | **0.01** | **2.35** | **2.02** | 11.00 |
| Light brown color | -0.02 | **0.05** | **0.06** | **0.01** | **1.24** | **1.06** | 11.00 |
| Pigment dots | 0.08 | **0.09** | **0.10** | **0.01** | **1.26** | **1.08** | 10.80 |
| Asymmetry (contrast) | 0.01 | 0.01 | 0.01 | 0.00 | 1.16 | 0.01 | 10.52 |
| Structureless areas | 0.00 | **0.04** | **0.07** | **0.01** | **1.24** | **0.48** | 9.00 |
| Red color | 0.02 | **0.08** | **0.08** | **0.01** | 1.13 | **1.58** | 6.50 |
| Black color (contrast) | 0.01 | 0.00 | 0.00 | 0.00 | 0.08 | -0.01 | 5.80 |
| Pigment network (contrast) | -0.08 | 0.00 | 0.00 | 0.00 | 0.03 | -0.05 | 5.60 |
| Light brown color (contrast) | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | -0.06 | 5.50 |
| White color (contrast) | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.24 | 5.00 |
| Dark brown color | 0.05 | **0.06** | **0.07** | **0.01** | 0.97 | **0.90** | 4.80 |
| Pigment dots (contrast) | -0.06 | 0.00 | 0.00 | 0.00 | 0.04 | -0.12 | 4.80 |
| Branched streaks (contrast) | 0.08 | 0.00 | 0.00 | 0.00 | 0.11 | -0.03 | 4.80 |
| Dark brown color (contrast) | 0.03 | 0.01 | 0.01 | 0.00 | 0.08 | 0.18 | 4.00 |
| Blue color (contrast) | 0.03 | 0.00 | 0.00 | 0.00 | 0.01 | 0.04 | 3.00 |
| Red color (contrast) | 0.05 | 0.00 | 0.00 | 0.00 | 0.03 | -0.08 | 3.00 |
| Pigment globules | 0.02 | **0.05** | **0.09** | **0.01** | **1.26** | **0.73** | 3.00 |
| Pigment globules (contrast) | -0.02 | 0.00 | 0.00 | 0.00 | 0.42 | 0.02 | 3.00 |
| Structureless areas (contrast) | -0.01 | 0.00 | 0.01 | 0.00 | 0.14 | -0.08 | 3.00 |

## IV. Results of experiments

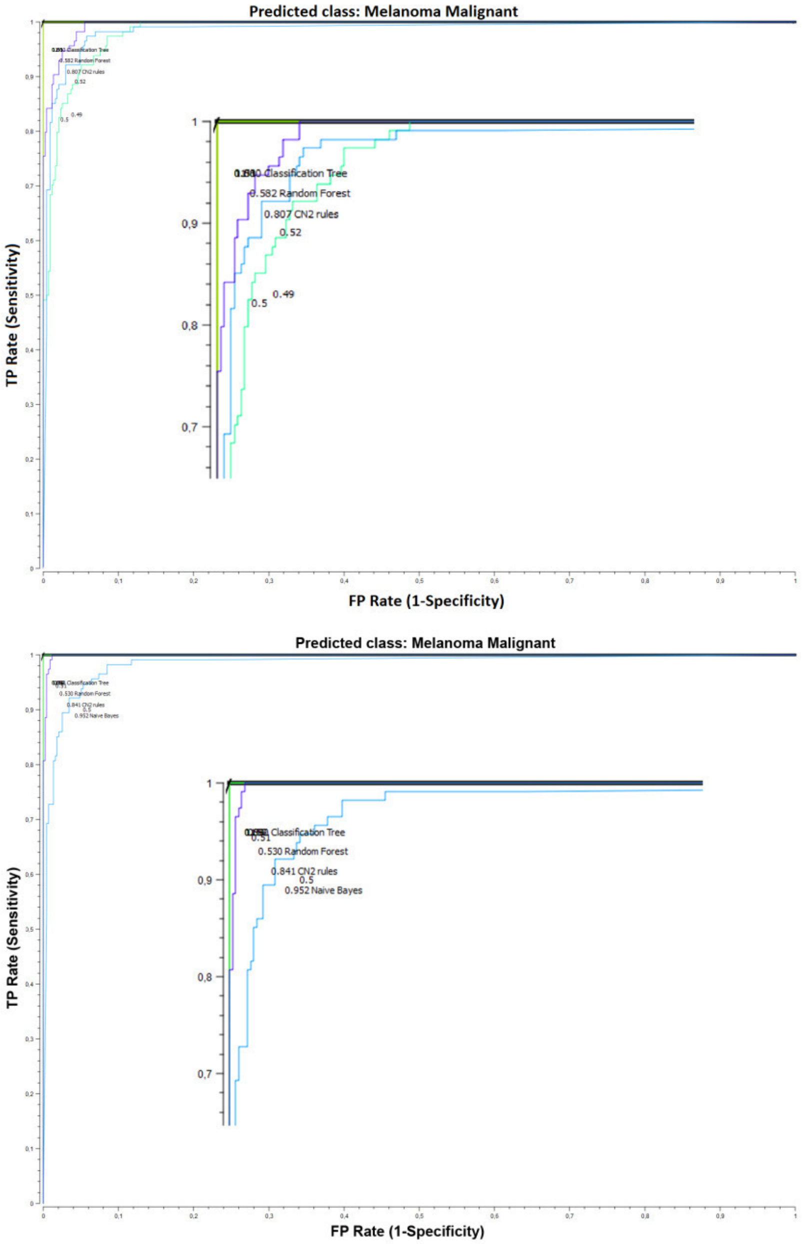
The first experiment revealed six variables, called *core features*, that were indicated as important by all, or nearly all, ranking measures, see Table 1. In this table, we can observe that *Border, Asymmetry, Pigment network*, *Blue color*, *Branched streaks* and *White color* features create stable and core set of features which have the highest values of seven measures of importance, particularly using *RuleQualityFS* measure, introduced in this investigation. In the same table, comparison with importance of contrast values (grey rows colored and *contrast* index) is also presented. The most important contrast feature is *Border (contrast)* for which *RuleQualityFS* measure, defined in earlier section, is equal to *12.50*. In this way, he is also treated as a threshold that separates the *core set* of attributes from all contrast features and other less informative attributes. Most of the measures used in this approach focused that selected core set of features has higher values of these parameters than gathered threshold attribute value. These values are denoted in bold style in Table 1. Hereby,

Table II.
Average classification results gathered using different classification quality measures applied to six machine learning models for four investigated sets of features combination

| Model | CA | Sens | Spec | AUC | IS | F1 | Prec | Brier | MCC | Informadness |
|---|---|---|---|---|---|---|---|---|---|---|
| **All original feature set** | | | | | | | | | | |
| **CT** | 0.79 | 0.78 | 0.92 | 0.92 | 1.30 | 0.78 | 0.78 | 0.34 | 0.70 | 0.70 |
| **RF** | 0.83 | 0.79 | 0.93 | 0.97 | 1.11 | 0.80 | 0.85 | 0.27 | 0.75 | 0.72 |
| **CN2** | 0.82 | 0.79 | 0.93 | 0.94 | 1.32 | 0.81 | 0.84 | 0.27 | 0.75 | 0.72 |
| **NB** | 0.78 | 0.77 | 0.92 | 0.96 | 1.24 | 0.78 | 0.80 | 0.27 | 0.71 | 0.69 |
| **kNN** | 0.81 | 0.82 | 0.93 | 0.94 | 1.40 | 0.82 | 0.81 | 0.29 | 0.75 | 0.76 |
| **SVM** | 0.84 | 0.83 | 0.94 | 0.97 | 1.37 | 0.84 | 0.85 | 0.21 | 0.78 | 0.78 |
| AVG | **0.81** | **0.80** | **0.93** | **0.95** | **1.29** | **0.80** | **0.82** | **0.28** | **0.74** | **0.73** |
| **Selected core feature set** | | | | | | | | | | |
| **CT** | 0.77 | 0.73 | 0.91 | 0.91 | 1.19 | 0.73 | 0.75 | 0.34 | 0.65 | 0.64 |
| **RF** | 0.75 | 0.69 | 0.90 | 0.94 | 0.99 | 0.68 | 0.72 | 0.33 | 0.61 | 0.58 |
| **CN2** | 0.76 | 0.70 | 0.90 | 0.92 | 1.09 | 0.72 | 0.78 | 0.34 | 0.64 | 0.60 |
| **NB** | 0.73 | 0.70 | 0.90 | 0.94 | 1.12 | 0.71 | 0.73 | 0.33 | 0.61 | 0.59 |
| **kNN** | 0.77 | 0.75 | 0.91 | 0.92 | 1.26 | 0.75 | 0.76 | 0.34 | 0.67 | 0.66 |
| **SVM** | 0.75 | 0.71 | 0.90 | 0.93 | 1.09 | 0.72 | 0.74 | 0.33 | 0.63 | 0.61 |
| AVG | **0.75** | **0.71** | **0.90** | **0.93** | **1.12** | **0.72** | **0.75** | **0.33** | **0.64** | **0.61** |
| **All original feature set with TDS parameter** | | | | | | | | | | |
| **CT** | 1.00 | 1.00 | 1.00 | 1.00 | 1.85 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 |
| **RF** | 0.99 | 0.98 | 0.99 | 1.00 | 1.59 | 0.98 | 0.99 | 0.06 | 0.98 | 0.98 |
| **CN2** | 1.00 | 1.00 | 1.00 | 1.00 | 1.62 | 1.00 | 1.00 | 0.03 | 1.00 | 1.00 |
| **NB** | 0.92 | 0.90 | 0.97 | 0.99 | 1.53 | 0.90 | 0.91 | 0.13 | 0.88 | 0.87 |
| **kNN** | 0.86 | 0.87 | 0.95 | 0.96 | 1.51 | 0.86 | 0.86 | 0.21 | 0.81 | 0.82 |
| **SVM** | 0.94 | 0.92 | 0.98 | 1.00 | 1.64 | 0.93 | 0.93 | 0.08 | 0.91 | 0.90 |
| AVG | **0.95** | **0.94** | **0.98** | **0.99** | **1.62** | **0.95** | **0.95** | **0.09** | **0.93** | **0.93** |
| **Selected core feature set with NewTDS parameter** | | | | | | | | | | |
| **CT** | 1.00 | 1.00 | 1.00 | 1.00 | 1.85 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 |
| **RF** | 1.00 | 1.00 | 1.00 | 1.00 | 1.61 | 1.00 | 1.00 | 0.04 | 1.00 | 1.00 |
| **CN2** | 1.00 | 1.00 | 1.00 | 1.00 | 1.59 | 1.00 | 1.00 | 0.04 | 1.00 | 1.00 |
| **NB** | 0.99 | 0.98 | 1.00 | 1.00 | 1.80 | 0.98 | 0.98 | 0.02 | 0.98 | 0.98 |
| **kNN** | 0.94 | 0.92 | 0.98 | 0.99 | 1.68 | 0.93 | 0.93 | 0.10 | 0.91 | 0.90 |
| **SVM** | 0.98 | 0.98 | 0.99 | 1.00 | 1.74 | 0.98 | 0.98 | 0.04 | 0.97 | 0.97 |
| AVG | **0.98** | **0.98** | **1.00** | **1.00** | **1.71** | **0.98** | **0.98** | **0.04** | **0.98** | **0.98** |

we can observe that different measures give different threshold, and it also shows that some other measures than *RuleQualityFS* include all original variables in core sets, e.g. *Information Gain, Gain Ratio, Gini index* and *Random Forest* application. Thus, we cannot extract smaller set of relevant attributes than the original one.

The second part of experiments focused on calculation of standard *TDS* and *NewTDS* defined earlier. Based on formula 1 and formula 4, these two values were obtained. Then, two datasets that include *TDS* and *NewTDS* respectively were investigated in next part of experiment.



The third part of experiment devoted to estimation of prediction quality of utilized machine learning algorithms described in section III. During this step six different algorithms were applied using 10-fold cross validation method. Average results are collected in Table 2. Procedure were utilized to four specified sets:

*(i)* *original set containing all descriptive features,*

*(ii)* *only selected core feature set based on its importance calculated in the first step,*

*(iii)* *original set containing all descriptive features with added standard TDS parameter,*

*(iv)* *core feature set with added NewTDS parameter.*

Additionally, to compare results, average values (*AVG*) of all evaluation measures were calculated.

Fig. 1 Comparison of ROC curves gathered for Melanoma malignant class using six learning algorithms by investigation of original dataset (top chart) and selected core features with added *NewTDS* attribute (bottom chart)

Based on acquired results (see Table 2), it could be stressed that core set of features which contains only 6 from 13 attributes has very similar prediction quality as it was observed with all original 13 attributes. For instance, popular measure in data analysis AUC decreased on average only from 0.95 to 0.93. However, average Classification Accuracy decreased rather significantly from 0.81 to 0.75, and also Informadness, which in itself connects Sensitivity and Specificity, decrease on average from 0.73 to 0.61. Next, if we tried to add calculated standard TDS values it is observed that AUC reached better average value, 0.99. This outcome could be also observed in form of Receiver Operating Characteristic curve. Comparison of ROC curves for original and with NewTDS feature set generated only for Melanoma malignant class is presented on figure 1. In turn, all other measures also increased significantly. By adding TDS values to dataset Informadness measure increased on average from 0.73 to 0.93. Thus, it could be said that this approach could be positively applied in other, different medical issues.

The last step of experiment shows that the feature space could be probably significantly reduced. It means, that we can use only six from thirteen descriptive attributes in connection with new total score parameter could be successfully applied. In Table 2, the average value of all evaluation measure increased significantly reaching almost limits. For example AUC and Specificity reached 1.0, in turn Informadness achieve 0.98, what is very good result. According to this results it could be stressed that this methodology improves prediction of learning models and additionally simplifies space of problems by reducing its dimensionality.

## References

[1] R. Lucas, A. McMichael, B. Armstrong, and W. Smith, "Estimating the global disease burden due to ultraviolet radiation exposure.," *Int. J. Epidemiol.*, vol. 37, pp. 6546–67, 2008.

[2] F. R. Rigel D.S., Russak J., "The evolution of melanoma diagnosis: 25 years beyond the ABCDs," *CA. Cancer J. Clin.*, vol. 60, no. 5, pp. 301–316, 2010.

[3] American Cancer Society, "Cancer Facts & Figures," *Cancer Facts Fig.*, 2014.

[4] R. Nicolas, A. Fornells, E. Golobardes, G. Corral, S. Puig, and J. Malvehy, "DERMA: A melanoma diagnosis platform based on collaborative multilabel analog reasoning," *Sci. World J.*, vol. 2014, 2014.

[5] J. W. Grzymala-Busse, Z. S. Hippe, M. Knap, and W. Paja, "Infoscience technology: the impact of internet accessible melanoid data on health issues," *Data Sci. J.*, vol. 4, pp. 77–81, 2005.

[6] P. Cudek, W. Paja, and M. Wrzesien, "Automatic System for Classification of Melanocytic Skin Lesions Based on Images Recognition," in *Man-Machine Interactions 2, Proceedings of the 2nd International Conference on Man-Machine Interactions, ICMMI 2011, The Beskids, Poland, October 6-9, 2011*, vol. 103, pp. 189–196.

[7] P. Cudek, W. Paja, and M. Wrzesien, "Image Recognition System for Diagnosis Support of Melanoma Skin Lesion," in *Security and Intelligent Information Systems - International Joint Conferences, SIIS 2011, Warsaw, Poland, June 13-14, 2011, Revised Selected Papers*, 2011, vol. 7053, pp. 217–225.

[8] W. Paja and M. Wrzesien, "Medical Datasets Analysis: A Constructive Induction Approach," in *Advances in Data Mining. Applications and Theoretical Aspects, 10th Industrial Conference, ICDM 2010, Berlin, Germany, July 12-14, 2010. Proceedings*, 2010, vol. 6171, pp. 442–449.

[9] R. Kohavi and R. Kohavi, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.

[10] A. Wosiak and D. Zakrzewska, "Feature Selection for Classification Incorporating Less Meaningful Attributes in Medical Diagnostics," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, 2014, vol. 2, pp. 235–240.

[11] N. Pérez, M. A. Guevara, A. Silva, I. Ramos, and J. Loureiro, "Improving the performance of machine learning classifiers for Breast Cancer diagnosis based on feature selection," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, 2014, vol. 2, pp. 209–217.

[12] Z. S. Hippe, S. Bajcar, P. Blajdo, J. P. Grzymala-Busse, J. W. Grzymala-Busse, M. Knap, W. Paja, and M. Wrzesien, "Diagnosing Skin Melanoma: Current versus Future Directions," *TASK Q.*, vol. 7, no. 2, pp. 289–293, 2003.

[13] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig, "The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions.," *J. Am. Acad. Dermatol.*, vol. 30, no. 4, pp. 551–559, 1994.

[14] U. Weigert, W. H. C. Burgdorf, and W. Stolz, "ABCD rule," in *An Atlas of Dermoscopy, Second Edition*, A. A. Marghoob, J. Malvehy, and R. P. Braun, Eds. CRC Press, 2012, pp. 113–117.

DOI: 10.15439/2015F246
ACSIS, Vol. 5

# Anthropometric Predictors and Artificial Neural Networks in the diagnosis of Hypertension

Krzysztof Pytel
University of Lodz,
Faculty of Physics and
Applied Informatics,
Lodz, Poland
Email: kpytel@uni.lodz.pl

Tadeusz Nawarycz,
Lidia Ostrowska-Nawarycz
Medical University of Lodz,
Department of Biophysics,
Lodz, Poland

Wojciech Drygas
Medical University of Lodz,
Department of Epidemiology,
Lodz, Poland

*Abstract*—**Artificial Neural Networks (ANNs) play a vital role in the medical field in solving various health problems like estimating the risk of cardiovascular diseases. The article concerns the process of developing ANNs for estimating the risk of arterial hypertension. ANNs proposed in this article use anthropometrical predictors, easy to control for everybody at home without special equipment. In the article we analyze four different models of ANNs and try to find out which model and set of anthropometrical predictors estimates the risk the most accurately. We use dataset of 2485 real cases of patients from the city of Lodz. The experiment was done in the Matlab environment. The performance of the proposed method in terms of accuracy and facility of use shows that ANNs can be effective tools for preliminary tests of arterial hypertension.**

## I. Introduction

ARTERIAL hypertension (HT) as a major risk factor for the development of cardiovascular diseases (CVD) constitutes an important problem of public health in Poland and other countries around the world [1]. Arterial HT is a disease that affects a wide range of the population, particularly the elderly after the age of 55. In compliance with the data given by multicenter Polish population health status study-WOBASZ, arterial hypertension appears among about 42.1% of men and 32.9% of women in Poland [2].

The diagnostics of HT as a sickness is based on:

a) the showing of increased values of blood pressure (based on their repeated measurement),
b) the estimation of the degree of the disease,
c) the differentiating of its etiology and results.

The value 140 mmHg for systolic blood pressure (SBP) and/or 90 mmHg the diastolic blood pressure (DBP) have been accepted for grown-ups as a border-line of hypertension [3]. The same classification is used for young, middle-aged and elderly subjects, whereas different criteria, based on percentilage for boys and girls according to their age and height [4]. The state of high risk of development of HT, the so-called pre-hypertension state (PHT), is defined when SBP amounts 120 - 139 mmHg and/or DBP amounts 80 - 89 mmHg. The patient with PHT needs the estimation of other cardiovascular risk and factors and modification of the lifestyle in the first instance.

Overweight and obesity are the most widespread occurrent environmental factor, that can cause the development of arterial hypertension (HT). The increased prevalence of obesity (especially visceral obesity) and other cardiovascular risk factors are closely associated with the rising incidence of CVD and type 2 diabetes mellitus. The enlarged quantity of the adipose tissue ties in with hyperinsulinemia and insulin resistance, which contributes to the level of blood pressure. Hormones and cytokines produced in adipocytes (eg. leptin, resistin, adiponektion, the interleukin-6, tumor necrosis factor-alpha and others) play a key role, and emitted to the circulation of blood system, regulate lipid and glucose metabolism [5]. In Framingham research 70% of men and 60% of women with HT and with coexisting overweight or obesity, had their systaltic blood pressure increased about 4.5 mmHg on every 5 kg of overweight [6].

Similar relations between the Body Mass Index (BMI) and arterial blood pressure were shown also in INTERSALT research [7]. With people having the similar body height, a 10 kg difference caused the increase of systaltic/diastolic pressure by 3/2.2 mmHg respectively.

In Poland, the result of the IDEA (The International Day for Evaluation of Abdominal Obesity) research on grown-ups aged from 18 to 80, showed that the frequency of appearing of both abdominal overweight and arterial hypertension is one of the highest in Europe [8].

Considerably more often than among our European neighbours, both from the North and West, one ascertained abdominal overweight (in Poland: 54% of women and 38% of men; in north-west Europe: 45% of women and 33% of men). Besides, it was observed that arterial hypertension more often appears among people with overweight or obesity when compared to their slim peers of the same age. Stout people more often suffered from diseases of the cardiovascular (CVD) system, diabetes and dyslipidemia. Very often these diseases coexisted with overweight or obesity. The investigation the IDEA concluded that both among men and women, abdominal obesity and the BMI tie in with CVD independently.

In spite, of the well-known correlation between HT and overweight, controversies about which anthropometrical factors (general and abdominal overweight and other such pa-

978-83-60810-66-8/$25.00©2015, IEEE

rameters as age and sex determine the most effective model of HT prediction) exist all the time.

In our work, we used ANNs, and the data from the research adults aged from 20 to 80 from the city Lodz, and we executed estimations of the efficiency of HT diagnostic for different models and anthropometrical data.

## II. Dataset for Artificial Neural Networks

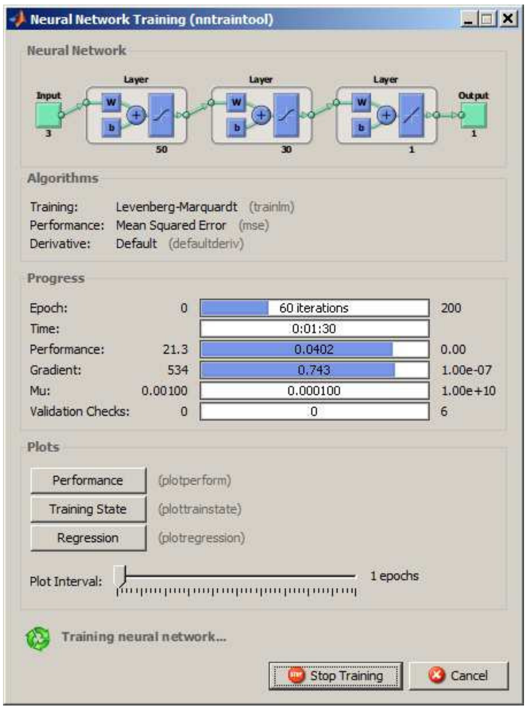
The Artificial Neural Networks (ANNs) are constructed using various predictors, to be trained, tested and validated using the respective data sets. In experiments we used the set of real data concerning patients from the city of Lodz, aged from 20 to 80. The set of data embraced 2485 cases of patients, 1197 men and 1288 women. Patients who did not have hypertension is the largest group (1370 cases, whereof 624 is men and 746 is women). In the group of patients with the 1-st stage of hypertension, there were 569 cases (315 men and 254 women). The group of patients with the 2-nd stage hypertension embraced 464 cases (221 men and 243 women). The least numerous group of patients was the one with the 3-rd stage of hypertension and if consisted of 82 cases (37 men and 45 women).

The dataset was divided into a training set and a test set. The training set was used for neural network training, that is for adjustment of weights between neurons. The test set was used after the network was trained, to test ANNs accuracy on new data (not used for training).

## III. Models of the Artificial Neural Networks

Artificial Neural Networks (ANNs) are algorithms inspired by function of the human brain. ANNs are usually presented as systems of interconnected "neurons" divided into a few layers. Each neuron can compute values from inputs and is capable of machine learning. Each neuron has a certain number of inputs, a real number associated with each connection (the weight of the connection), and has its own transfer function. The behaviour of an ANN depends on the weights and the transfer function. Most of the algorithms used in training ANN employ some form of backpropagation method. In supervised learning, a given set of input/output pairs is presented to the network. A learning algorithm tries to adjust connection weights to minimize the average squared error between the network's output and the target value for all the example pairs. ANNs can produce the values between 0 and 1 for any input vector. The advantage of ANNs is their ability to learn from the observed data. They can be used to infer a function from observations. This is useful in complex applications or tasks, where discovering such a function by hand is impractical or impossible.

All the experiments were done in the Matlab environment. In the experiments we used three-layer artificial neural networks. The number of neurons in input, hidden and output layers was fixed to 50, 30 and 1 respectively. The proposed structure of the network was obtained as a result of a series of initial tests during which we were testing different structures. The hyperbolic tangent sigmoid transfer function (tansig) was used in the input and hidden layers, and the linear function (purelin) in the output layer. A Levenberg-Marquardt method with 200 cycles was used for teaching the networks. In Fig. 1 we show an example of the network teaching in the Matlab environment.



Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

```
Epoch:                    0   60 iterations   200
Time:                         0:01:30
Performance:           21.3   0.0402          0.00
Gradient:               534   0.743           1.00e-07
Mu:                 0.00100   0.000100        1.00e+10
Validation Checks:        0   0               6
```

(plotperform)
(plottrainstate)
(plotregression)

Plot Interval: 1 epochs

Training neural network...

Fig. 1. An example of a three-layer artificial neural network teaching with the Levenberg-Marquardt method

The result of the classification was represented on the output layer in the form of a real number. We accepted that output values from 0 to 0.17 represent the lack of hypertension, values from 0.17 to 0.5 represent the 1st degree hypertension, values from 0.5 to 0.83 represent 2nd degree hypertension, and values above 0.83 represent the 3rd degree hypertension. In experiments we tested four models of the network with different input values:

- model 1: we accepted 3 inputs: the Body Mass Index (BMI), the Waist Circumference (WC) and the Age. On the output of the network we received the real number representing the result of the classification;
- model 2: we accepted 3 inputs: the Body Mass Index (BMI), the Waist Circumference normalized in relation to the BMI (WCBMI) and the Age. On the output of the network we received the real number representing the result of the classification;
- model 3: we accepted 4 inputs: the Body Mass Index (BMI), the Waist Circumference (WC), the Sex and the Age. On the output of the network we received the real number representing the result of the classification;
- model 4: we accepted 4 inputs: the Body Mass Index (BMI), the Waist Circumference normalized in relation to the BMI (WCBMI), the Sex and the Age. On the output

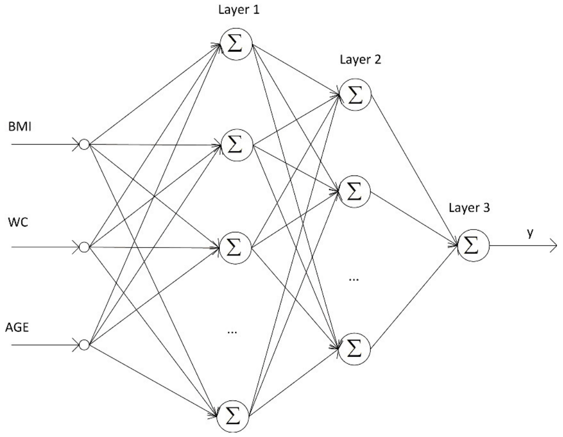


Fig. 2. Artificial Neural Network Architecture

TABLE I
THE RESULTS OBTAINED BY EACH MODEL OF NEURAL NETWORK (THE AVERAGE NUMBER OF INCORRECTLY RECOGNIZED CASES IN PERCENTAGES)

| | HT 0 | HT 1 | HT 2 | HT 3 |
|---|---|---|---|---|
| model 1 | 7.96 | 10.42 | 17.76 | 21.18 |
| model 2 | 6.68 | 11.94 | 16.6 | 17.3 |
| model 3 | 6.52 | 9.82 | 15.18 | 14.86 |
| model 4 | 7.00 | 11.92 | 16.9 | 19.92 |

of the network we received the real number representing the result of the classification.

In models 3 and 4, a non-numerical attribute (Sex) was used. The value of this attribute was converted into number: 0 for women and 1 for men.

In Fig. 2 we present the Architecture of Artificial Neural Network.

## IV. Experimental results

For the test we used all four proposed models of ANNs. Every network was run five times, embracing the learning process, and then, the test of the efficiency of the classification. The obtained result is the average from all five runs. The results obtained by each neural network are collected in table 1. The values in the table are in percent ages and represent the average number of incorrectly recognized cases with the lack of hypertension (HT0), hypertension of the first stage (HT1), hypertension of the second stage (HT2) and hypertension of the third stage (HT3). During the experiments, we used only the results of anthropometrical measurement and systaltic (SBP) and diastolic (DBP) blood pressure. The other factors, such as medicines, family conditionings and other, were not taken into account.

The comparison of model 1 and model 2, let on to check which input value (WC or WCBMI) better determines the results of the classification. The graph in Fig. 3 illustrat the differences in the average number of incorrectly diagnosed cases by neural networks with the use of models 1 and 2.

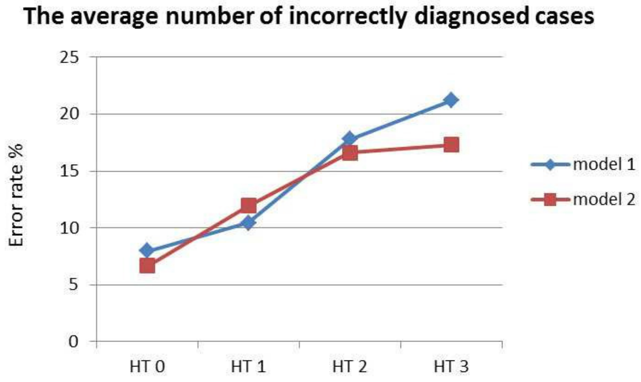


Fig. 3. Differences in the average number of incorrectly diagnosed cases by neural networks with the use of models 1 and 2.

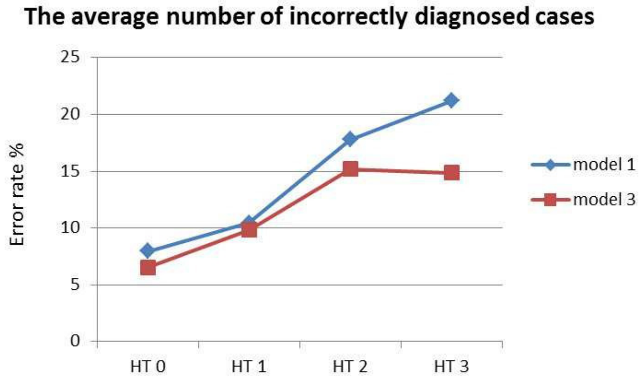


Fig. 4. Differences in the average number of incorrectly diagnosed cases by neural networks with the use of models 1 and 3.

The graph in Fig. 4 illustrates differences in the average number of incorrectly diagnosed cases by neural networks with the use of models 1 and 3.

Models 3 and 4 let on to check if the additional parameter (Sex) will improve exactitude of the classification and how it influences the value of the hypertension. The graph in Fig. 5 illustrates the average number of incorrectly diagnosed cases by neural networks with the use of all four models.

## V. Conclusions

Overweight and obesity leading to arterial hypertension are a serious problem of the population of Poland and other countries around the world. Artificial Neural Networks, basing on anthropometrical factors, which are easy to measure at home, let on to make diagnostic reconnaissances towards the risk of arterial hypertension.

The neural network in which we used the Waist Circumference normalized in relation to the BMI (WCBMI), as the input (model 1) obtained a greater accuracy than the network in which we used the Waist Circumference without normalization (model 2).

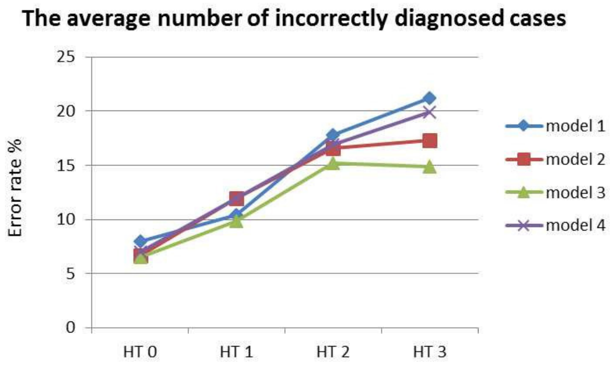


Fig. 5. Average number of incorrectly diagnosed cases by neural networks with the use of all four models.

The introduction of the additional input parameter to the network (the Sex), brought about the enlargement of the accuracy of the network. The greatest difference can be observed in the third stage of hypertension, though in the test set in this category the least cases appeared.

The results of our study may be a starting point for the construction of a more complex ANN system for the assessment of global arterial hypertension risk. Taking into account the growing scale of overweight and obesity epidemic around the world, the Neural Networks systems should be more widely used for diagnosis of this problem.

## References

[1] Hajjar I, Kotchen JM, Kotchen TA. *Hypertension: trends in prevalence, incidence, and control.* Annu Rev Public Health. 2006;27:465-90.

[2] Tykarski A., Posadzy-Malaczynska A., Wyrzykowski B. i wsp.: *Rozpowszechnienie nadcisnienia tetniczego oraz skutecznosc jego leczenia u doroslych mieszkancow naszego kraju. Wyniki programu WOBASZ. Kardiol.* Pol. 2005; 63 (supl. 4): S614-619 (in Polish).

[3] 2013 ESH/ESC Guidelines for the management of arterial hypertension, Journal of Hypertension 2013, 31:1281-1357

[4] Lurbe E, Cifkova R, Cruickshank JK, et al. *Management of high blood pressure in children and adolescents: recommendations of the European Society of Hypertension.* J Hypertens 2009; 27:1719-1742

[5] 5. Scherer PE, Williams S, Fogliano M, et al. *A novel serum protein similar to C1q, produced exclusively in adipocytes.* J Biol Chem. 1995; 270:26746-26749

[6] Franklin SS, Gustin WIV, Wong ND, et al. *Haemodynamic patterns of age-related changes in blood pressure. The Framingham Heart Study.* Circulation 1997; 96:308-315.

[7] Stamler J. *The INTERSALT Study: background, methods, findings, and implications.* Am J Clin Nutr February 1997 vol. 65 no. 2 626S-642S

[8] *Nadcisnienie tetnicze u osob w wieku podeszlym.* (red)T. Grodzicki, J.Kocemba, B. Gryglewska (in Polish)

[9] B. Sumathi, Dr. A. Santhakumaran *Pre-Diagnosis of Hypertension Using Artificial Neural Network*

[10] Kaur A., Bhardwaj A. *Artificial Intelligence in Hypertension Diagnosis: A Review.* International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2633-2635

[11] Samant, Rahul, and Srikantha Rao. *Evaluation of Artificial Neural Networks in Prediction of Essential Hypertension.* International Journal of Computer Applications 2013

[12] Shehu, N., S. U. Gulumbe, and H. M. Liman. *Comparative study between conventional statistical methods and neural networks in predicting hypertension status.* Advances in Agriculture, Sciences and Engineering Research 2013

[13] Ture, Mevlut, et al. *Comparing classification techniques for predicting essential hypertension.* Expert Systems with Applications, Elsevier 2005

[14] Djam, X. Y., and Y. H. Kimbi. *Fuzzy expert system for the management of hypertension.* The Pacific Journal of Science and Technology 2011.

[15] Zurada, J. M. *Introduction to artificial neural systems.* West Publishing Co. 1992.

[16] Nawarycz T., Pytel K., Gazicki-Lipman M., Drygas W., Ostrowska-Nawarycz L., *A Fuzzy Logic Approach to The Evaluation of Health Risks Associated with Obesity.* FedCSIS 2013: 231-234

DOI: 10.15439/2015F263
ACSIS, Vol. 5

# Short survey: adaptive threshold methods used to segment immunonegative cells from simulated images of follicular lymphoma stained with 3,3'-Diaminobenzidine&Haematoxylin

Lukasz Roszkowiak, Anna Korzynska, Dorota Pijanowska
Nalecz Institute of Biocybernetics and Biomedical Engineering
Ks. Trojdena 4 Str., 02-109 Warsaw, Poland
Email: lroszkowiak@ibib.waw.pl

***Abstract*—We perform a short survey of image thresholding methods for very specific task, and assess their performance comparison. We analyse performance of adaptive thresholding methods concerning segmentation of immunonegative cells of follicular lymphoma tissue samples stained with 3,3'-Diaminobenzidine&Haematoxylin. We use artificial images based on experimental images that greatly simulates real samples and simplifies process of evaluation. We chose 8 methods of adaptive threshold segmentation, with different approach. They were applied to 6 different monochromatic images derived from original RGB images, by splitting layers, conversion to Lab colour space and colour deconvolution. Evaluation of the results was performed with basic statistical measures as sensitivity and specificity along with Jaccard's coefficient. We identify the thresholding algorithms with superior performance. Collected results will be used to design the new better method based on this approach.**

## I. Introduction

TISSUE samples stained immunohistochemically are commonly used by pathologists to distinguish various types of cancer [1]. One of them is follicular lymphoma which is the second most common form of non-Hodgkin's lymphoma [2]. Nowadays, pathologist doesn't have to evaluate samples via microscope but it can be done by examination of digital images of samples. Unfortunately, even with this aid the human evaluation is irreproducible and prone to error [3]. Moreover, it tends to change from one expert to another, as well as in time for one expert. As analysis is mostly based on counting of immunopositive and immunonegative cells, computer evaluation would give many advantages, like acceleration and reproducibility of the process. Unfortunately, segmentation of such images is not an easy task. Most computer-based procedures for immunohistochemistry image analysis [4]–[7] have limited applicability due to numerous drawbacks.

For test images in this study we decided to use simulated (artificial) images based on experimental images of follicular lymphoma tissue sections stained with 3,3-diaminobenzidine (DAB) and contra-stained with haematoxylin (H). The main reason for this decision was having 'gold standard' reference image that can be used to evaluate results of segmentation methods.

Experimental as well as artificial images consist of brown objects among blue ones with bright background that sometimes has slight blue tint. Most of the objects have minor but visible texture. Unfortunately, each and every sample differs in many characteristics. There is a huge variability of shape (from round to elongated), size and colour of objects of interest as well as in the image generally; in colour intensity, range of colour and tone throughout image plane.

For this survey we chose 8 methods of adaptive threshold segmentation to test and evaluate their ability to segment immunonegative cells. We decided to apply segmentation algorithms to 6 different monochromatic images derived from original RGB images. Apart from simple dividing images into single layer images of separate RGB layers we used conversion to Lab colour space, as it was proved to be useful during segmentation in other applications [8]. Since b axis of Lab colour space represents yellow and blue at two ends of the axis, it could be very useful to find proper threshold value to segment blue objects of interest. Also we try to use L (luminescence) layer of Lab colour space as the objects should have contrasting value from the background. Another method we applied in this research is colour deconvolution [9]. We use Haematoxylin layer obtained with predefined colour vector in colour deconvolution algorithm. In summary, we compare results of segmentation performed on: (1) red layer of RGB; (2) green layer of RGB; (3) blue layer of RGB; (4) L layer of Lab colour space; (5) b layer of Lab colour space; (6) Haematoxylin layer after colour deconvolution.

To sum up, in this investigation, we present preliminary study concerning segmentation of immunonegative cells of follicular lymphoma tissue stained with DAB&H. It will lead to development of improved method of segmentation. In cooperation with our previous research [10] on segmentation of immunopositive cells we hope to achieve fully capable software to analyse digitized samples of follicular lymphoma tissue sections.

We acknowledge the financial support of the Polish National Science Centre grant, PRELUDIUM, 2013/11/N/ST7/02797.

## II. Methods

For this study, we decided to use previously created artificial images. Thus, we obtain the 'gold standard' images to evaluate the results. To process the images we have used previously tested methods with changed parameters appropriately for the new type of segmentation (presented in Table I). Evaluation of the results was performed with basic statistical measures as sensitivity, specificity along with Jaccard's coefficient. All necessary computations were performed in MATLAB.

### A. *Artificial images synthesis*

The process of creating and the usefulness confirmation of the artificial images is fully described in our previous work [10]. In short, it is done using the adjusted version of SIMCEP software [11], [12] and Camera Raw 4.1 module of Photoshop CS5. Lehmussola and co-workers developed SIMCEP to synthesize the full colour fluorescent microscopic images of nuclei or cells' culture. With our modification it creates images of transmission light microscopy.

Our artificial images were created based on the model images experimentally collected in Hospital Verge de la Cinta in Tortosa, Spain. They were created to resemble their experimental counterparts as much as possible. The number of cells, their size, shape and distribution have been adjusted. We tried to simulate signal degradation, typical for the microscope and camera technical limitations, such as noise, vinietting and blurring.

The great advantage of this technique is that along with the artificial image we obtain the reference image of 'true' objects of interest. This lets us not only compare number of segmented objects and their approximate position but properly evaluate every pixel of the image.

### B. *Methods of segmentation*

Since samples tend to show wide range of characteristics like colour, tone and intensity as well as contrast fluctuations, the locally adaptive thresholding methods seems to be most appropriate. Local threshold is calculated for every pixel with sliding window image processing. Threshold value is based on the intensity of the analysed pixel and its neighbourhood. All methods used in this study are chosen based on the survey [13], and are fully described in previous publication [10], therefore in the following part we present them briefly. All values of parameters used in segmentation algorithms are chosen experimentally and are presented in Table I.

Niblack [14] is the most basic adaptive threshold method; based on local variance.

Sauvola [15] is another local variance method and can be treated as modified version of Niblack's method.

Bernsen [16] is based on local contrast. Threshold value is calculated as a mean of the minimum and maximum value in neighbourhood of the analysed pixel if the contrast value was high enough.

White [17], basically, if the pixel is considerably (depending on the bias value parameter) darker than its surrounding, it is considered as an object.

Palumbo [18] is using centre-surround scheme. The treated area is divided into near neighbourhood and 4 diagonal windows are far neighbourhood. The tested pixel is supposed to be treated as object when the central window contains the foreground object and the neighbouring windows are filled with background.

Yasuda [19] is local contrast method and consists of four steps. First two are preprocessing; increasing dynamic range in the image, followed by nonlinear smoothing. Then primary thresholding is done, with course marking of background based on local contrast. Finally, precise segmentation to classify rest of the pixels is performed.

Hybrids of Niblack and Sauvola methods are described in previous publication [10]. As the basic methods seemed unsatisfactory in terms of sensitivity and specificity we modified them by adding the contrast condition.

TABLE I
Values of parameters used in segmentation methods

| | $w$ | $k$ | $R$ | $bias$ | $T_c$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ |
|---|---|---|---|---|---|---|---|---|---|
| Niblack | 51 | -0.2 | | | | | | | |
| Hyb.Nib. | 51 | -0.2 | | | 40 | | | | |
| Sauvola | 51 | 0.5 | 128 | | | | | | |
| Hyb.Sau. | 51 | 0.5 | 128 | | 40 | | | | |
| White | 51 | | | 1.05 | | | | | |
| Bernsen | 51 | | | | 40 | | | | |
| Palumbo | 21 | | | | | 100 | 0.85 | | |
| Yasauda | 51 | | | | | 0.19 | 0.39 | 0.8 | 0.05 |

### C. *Methods of evaluation*

To perform proper evaluation exact position of every object should be known. Fortunately, while we use artificial images we can use the black and white reference image of 'true' objects of interest as the 'gold standard'. Taking into account the result of each segmentation method and 'gold standard' image, following measurements are possible: true positive (TP), true negative (TN), false positive (FP), false negative (FN). Based on these parameters, statistical measurement of the performance of segmentation methods can be calculated, such as sensitivity, specificity, and Jaccard's coefficient.

## III. Results

Segmentation was performed on 6 different layers derived from one image. We tried performing segmentation on all of RGB layers. We also tested performance on the Haematoxylin layer acquired by applying colour deconvolution. Additionally, we transformed images from RGB to Lab colour space, and performed segmentation on layers $L$ and $b$.

Performing segmentation on RGB layers is computationally most simple approach that does not involve any transformations. The mean values of sensitivity and specificity for every segmentation method on all images are presented in Table II. Comparing mean results of RGB channels only red and green layers give tolerable results. Segmentation of blue channel cannot give good results as values of object and background

TABLE II
MEAN VALUES OF SENSITIVITY AND SPECIFICITY FOR ALL IMAGES APPLIED TO 6 DIFFERENT MONOCHROMATIC IMAGES DERIVED FROM THE ORIGINAL RGB IMAGES

| layer | sensitivity | specificity |
|---|---|---|
| R (RGB) | $0.895 \pm 0.043$ | $0.914 \pm 0.044$ |
| G (RGB) | $0.827 \pm 0.097$ | $0.912 \pm 0.048$ |
| B (RGB) | $0.501 \pm 0.219$ | $0.878 \pm 0.080$ |
| H (deconv.) | $0.919 \pm 0.047$ | $0.929 \pm 0.050$ |
| L (Lab) | $0.804 \pm 0.129$ | $0.915 \pm 0.049$ |
| b (Lab) | $0.839 \pm 0.071$ | $0.930 \pm 0.079$ |

are very similar for that layer, as can be seen in Figure 1. Hence, we discard blue channel from further analysis. The best of those three channels is segmentation of red channel. Transformation from RGB to Lab colour space is complex computational task but it may become helpful as results of segmentation on $L$ and $b$ layers are better in comparison to those of RGB layers. Objects of interest (blue immunonegative cells) in Luminescence layer ($L$) are significantly contrasted from the background. Unfortunately, so are other objects (brown immunopositive cells). Better results are achieved for $b$ layer of Lab colour space. It seems that blue and brown objects can be well separated in this layer since the representation of $b$ axis is yellow-blue. It is understandable because model brown colour consist only of red and green values (in RGB) as well as yellow which has the same components, while on the other end of the axis is blue colour. Mean value of sensitivity for these four layers is between 0.804 and 0.895 while specificity is between 0.912 and 0.930. Haematoxylin layer acquired by colour deconvolution algorithm stands out as best for performing segmentation. It has mean value of sensitivity $0.919 \pm 0.047$ and $0.915 \pm 0.049$ specificity, what is better than any other tested layer.

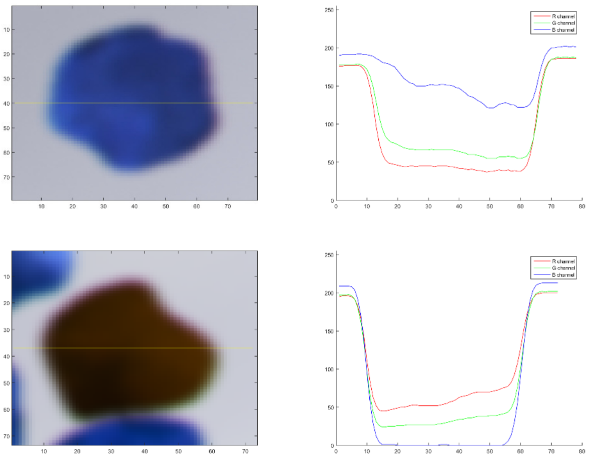


Fig. 1. Comparison of blue and brown objects in RGB colour space. Magnified blue object from artificial image (top-left) and its line profile (top-right); brown object (bottom-left) and its line profile (bottom-right).

In Table III are presented values of sensitivity, specificity, and Jaccard's coefficient for all tested methods on all proposed layers (channels) as mean performance on all test images.

As we stated before segmentation of red channel gives better results than green or blue channel. This also is confirmed while analysing separate methods instead of their mean value. For red channel best sensitivity is achieved by Sauvola method which has also worst specificity value. Contradictory is Palumbo method with best specificity and lowest sensitivity. White method has best value of Jaccad's coefficient and quite good sensitivity and specificity values, both above 0.90. Another to consider is Bernsen method with second highest Jaccard's coefficient and values of sensitivity and specificity above 0.91.

Layer $L$ from Lab colour space presents information about Luminescence in the image. For this layer Sauvola and Palumbo methods have similar results as for red channel. Highest sensitivity and lowest specificity by Sauvola; highest specificity and lowest sensitivity by Palumbo. Best value of Jaccard's coefficient is achieved by HybridSauvola method but it has lower value than for red channel of RGB.

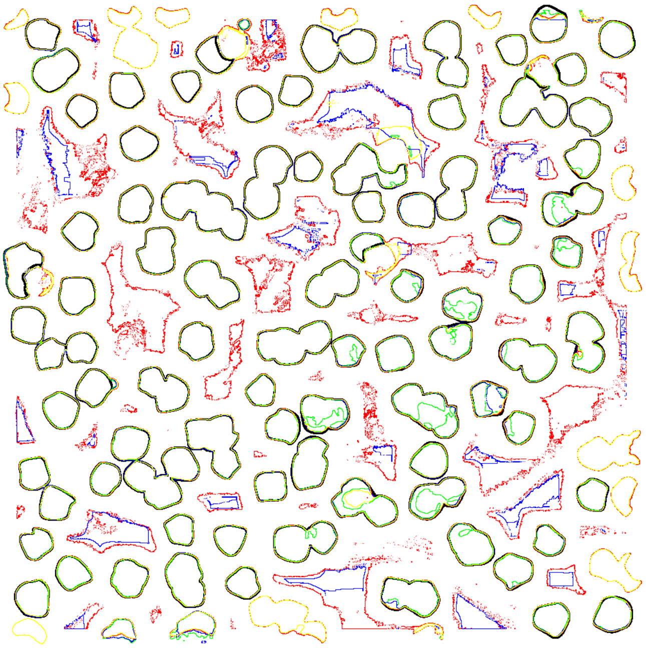


Fig. 2. Comparison of 4 methods performance on Haematoxylin layer. Sauvola—red; Palumbo—green; Bernsen—blue; White—yellow.

Layer $b$ from Lab colour space shows slightly different results. Best value of sensitivity has once again Sauvola method, and also worst specificity. HybridSauvola method seems to increase the specificity of Sauvola method to satisfactory level (0.937) but unfortunately simultaneously lowers the sensitivity value. White method has highest value of specificity but also lowest sensitivity. Palumbo method has acceptable results of sensitivity and specificity, 0.886 and 0.970 respectively, with best value of Jaccard's coefficient.

Last and possibly most interesting tested layer was Haematoxylin created with colour deconvolution. As it was for other

TABLE III
RESULTS OF IMAGE SEGMENTATION.

| Method | Sensitivity | Specificity | $r_J$ | | Sensitivity | Specificity | $r_J$ | | Sensitivity | Specificity | $r_J$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| layer | R from RGB | | | | G from RGB | | | | B from RGB | | |
| Niblack | 0.8869 | 0.8948 | 0.7153 | | 0.8462 | 0.8910 | 0.6775 | | 0.6100 | 0.8386 | 0.4505 |
| Hyb.Nib. | 0.8869 | 0.9421 | 0.7845 | | 0.8460 | 0.9385 | 0.7437 | | 0.6078 | 0.9078 | 0.5079 |
| Sauvola | 0.9357 | 0.8144 | 0.6571 | | 0.9064 | 0.8043 | 0.6264 | | 0.6937 | 0.7054 | 0.4169 |
| Hyb.Sau. | 0.9357 | 0.9111 | 0.7787 | | 0.9061 | 0.9034 | 0.7452 | | 0.6912 | 0.8649 | 0.5351 |
| Bernsen | 0.9148 | 0.9372 | 0.7989 | | 0.8626 | 0.9377 | 0.7541 | | 0.5248 | 0.8826 | 0.4226 |
| White | 0.9024 | 0.9433 | 0.8000 | | 0.8494 | 0.9444 | 0.7541 | | 0.5283 | 0.9370 | 0.4634 |
| Palumbo | 0.7988 | 0.9469 | 0.7090 | | 0.6053 | 0.9478 | 0.5359 | | 0.0611 | 0.9502 | 0.0549 |
| Yasuda | 0.8951 | 0.9222 | 0.7595 | | 0.7937 | 0.9251 | 0.6741 | | 0.2912 | 0.9379 | 0.2550 |
| | | | | | | | | | | | |
| layer | b from Lab | | | | L from Lab | | | | H from deconv. | | |
| Niblack | 0.9027 | 0.8957 | 0.7306 | | 0.8388 | 0.8924 | 0.6735 | | 0.9213 | 0.9043 | 0.7562 |
| Hyb.Nib. | 0.7888 | 0.9731 | 0.7458 | | 0.8380 | 0.9427 | 0.7423 | | 0.9213 | 0.9536 | 0.8337 |
| Sauvola | 0.9568 | 0.7466 | 0.6089 | | 0.9021 | 0.8039 | 0.6230 | | 0.9586 | 0.8246 | 0.6852 |
| Hyb.Sau. | 0.8309 | 0.9366 | 0.7301 | | 0.9012 | 0.9128 | 0.7524 | | 0.9586 | 0.9106 | 0.7967 |
| Bernsen | 0.8071 | 0.9625 | 0.7531 | | 0.8467 | 0.9411 | 0.7455 | | 0.9500 | 0.9435 | 0.8411 |
| White | 0.8067 | 0.9881 | 0.7859 | | 0.8365 | 0.9460 | 0.7451 | | 0.9307 | 0.9637 | 0.8596 |
| Palumbo | 0.8862 | 0.9695 | 0.8379 | | 0.5018 | 0.9503 | 0.4463 | | 0.8142 | 0.9851 | 0.7863 |
| Yasuda | 0.7356 | 0.9655 | 0.6849 | | 0.7652 | 0.9276 | 0.6524 | | 0.8959 | 0.9434 | 0.7911 |

where: $r_J$ - Jaccard's coefficient; deconv. - colour deconvolution.

layers Sauvola has highest sensitivity and lowest specificity, while Palumbo has highest specificity and lowest sensitivity. The best results according to Jaccard's coefficient are achieved by White method, and it has relatively good results of sensitivity and specificity, 0.930 and 0.964 respectively. Other methods worth considering using on this layer are Bernsen and HybridNiblack with second and third best values of Jaccard's coefficient. Figure 2 shows comparison of 4 most efficient methods that perform segmentation on Haematoxylin layer. Black thick outline of 'gold standard' objects is covered with 4 colour overlays representing 4 methods of segmentation; Sauvola—red, Palumbo—green, Bernsen—blue, White—yellow.

Overall, best results are achieved for *b* layer of Lab colour space and Haematoxylin layer of colour deconvolution. Slightly higher values of specificity can be observed as results of segmentation *b* layer of Lab colour space. Also, slightly higher values of sensitivity are achieved for Haematoxylin layer. On the whole, according to Jaccard's coefficient best of all methods is White method segmentation performed on Haematoxylin layer.

## IV. DISCUSSION AND CONCLUSION

The main aim of this investigation is to analyse performance of adaptive thresholding methods and to gather knowledge how to design the new better method based on this approach. We evaluated performance of 8 methods of segmentation applied to: separate channels of RGB, L and b layer of Lab colour space, and Haematoxylin layer after colour deconvolution. Best results were achieved for b layer of Lab colour space and Haematoxylin layer of colour deconvolution, so only these two should be taken into further consideration. Comparing results of all 8 methods, we concluded that only 4 have future potential: Sauvola, Palumbo, Bernsen, and White. Hybrid methods are also worth mentioning, they tend to have better overall results than their unmodified counterparts based on the Jaccard's coefficient.

Sauvola has best overall sensitivity, what means that there are not many false negative pixels, so there are least pixels representing 'true' objects not included into segmented image. The area of segmented objects of interest seems to be properly segmented; it does not change shape of segmented objects significantly. This method has problem with vast areas of clear background as there are less intensity variations. The extra segmented objects could be discarded from the results during validation phase or if compared to other segmentation method.

Quite the opposite, Palumbo in general has best specificity. The false positive pixels are minimal for this method. Unfortunately it does not segment full area of objects of interest and they have holes as a result. The main advantage of this method is that there are very little extra segmented pixels.

Bernsen has moderate results. It does not affect object size and the error in area detection is regularly located on border part of object. With too large window the method works slow and tends to misclassify small objects with low contrast, especially when they are located near objects with better contrast.

White method decreases the size of segmented objects. The local threshold level in the White method is dependent on bias which increase intensity of analysed pixel causes that the method perform well in images with high contrast between objects and background. Furthermore, in this investigation we adjusted the window size and values of the parameters especially for the tested images. The idea is to find criteria for automatic adjustment of parameters to fit to processed image. We should consider developing appropriate preprocessing phase that could equalize images [20] before they are segmented as it may improve results and make it easier to adjust parameter values.

As a result of this investigation we believe that there are three ways to improve segmentation of tested methods. First, globally, results of different methods could be merged using logical operands and probability analysis. For example, the results of the most accurate in object localization method could be used as seed points for the other methods with precise local area selection. Regrettably, the computational cost and time of computation could be quite high. Second, it seems object oriented analysis could give good results. Separate objects segmented by the best method could be evaluated by shape or colour features. Unfortunately, the immunonegative cells have wide range of these features, shape form round to elongated and colour from light blue to very dark. Also background has often blue tint that could distort the evaluation. Third, development of new method that would merge advantages of best tested methods. Since all four best methods are based on different factors, combining the benefits of variance method, centre-surround scheme, contrast based method, and bias method would not be easy but should give best results. This methodology could be most efficient regarding time and computation cost. As sliding window operations of adaptive thresholding methods tend to be slow it would be best to limit the number of whole image processing as much as possible.

### A. Future works

In future work, we plan to apply tested methods of segmentation to experimental images. If the results will not be satisfactory we plan to develop new method of segmentation based on the advantages of best of tested methods.

## Acknowledgment

We would like to thank Molecular Biology and Research Section, Hospital de Tortosa Verge de la Cinta, Institut dInvestigaci Sanitria Pere Virgili (IISPV), URV, Spain and Pathology Department of the same hospital for their cooperation and generous sharing of samples.

## References

[1] S. Swerdlow, I. A. for Research on Cancer, and W. H. Organization, *WHO classification of tumours of haematopoietic and lymphoid tissues*, ser. World Health Organization classification of tumours. International Agency for Research on Cancer, 2008, iSBN-13: 9789283224310; ISBN-10: 9283224310.

[2] D. Sandeep S., G. Wright, B. Tan, A. Rosenwald, R. D. Gascoyne, W. C. Chan, R. I. Fisher, R. M. Braziel, L. M. Rimsza, T. M. Grogan, T. P. Miller, M. LeBlanc, T. C. Greiner, D. D. Weisenburger, J. C. Lynch, J. Vose, J. O. Armitage, E. B. Smeland, S. Kvaloy, H. Holte, J. Delabie, J. M. Connors, P. M. Lansdorp, Q. Ouyang, T. A. Lister, A. J. Davies, A. J. Norton, H. K. Muller-Hermelink, G. Ott, E. Campo, E. Montserrat, W. H. Wilson, E. S. Jaffe, R. Simon, L. Yang, J. Powell, H. Zhao, N. Goldschmidt, M. Chiorazzi, and L. M. Staudt, "Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells," *New England Journal of Medicine*, vol. 351, no. 21, pp. 2159–2169, 2004. doi: 10.1056/NEJMoa041869 PMID: 15548776.

[3] T. Seidal, A. J. Balaton, and H. Battifora, "Interpretation and quantification of immunostains," *The American Journal of Surgical Pathology*, vol. 25, no. 9, pp. 1204–1207, 2001. doi: 10.1097/00000478-200109000-00013

[4] G. Bueno, R. Gonzalez, O. Deniz, M. Garcia-Rojo, J. Gonzalez-Garcia, M. Fernandez-Carrobles, N. Vallez, and J. Salido, "A parallel solution for high resolution histological image analysis," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 1, pp. 388 – 401, 2012. doi: 10.1016/j.cmpb.2012.03.007

[5] K. Kayser, D. Radziszowski, P. Bzdyl, R. Sommer, and G. Kayser, "Towards an automated virtual slide screening: theoretical considerations and practical experiences of automated tissue-based virtual diagnosis to be implemented in the internet," *Diagnostic Pathology*, vol. 1, pp. 1–8, 2006. doi: 10.1186/1746-1596-1-10

[6] C. Lopez, M. Lejeune, M. T. Salvado, P. Escriva, R. Bosch, L. Pons, T. Alvaro, J. Roig, X. Cugat, J. Baucells, and J. Jaen, "Automated quantification of nuclear immunohistochemical markers with different complexity," *Histochemistry and Cell Biology*, vol. 129, pp. 379–387, 2008. doi: 10.1007/s00418-007-0368-5

[7] C. Lopez, M. Lejeune, R. Bosch, A. Korzynska, M. Garcia-Rojo, M.-T. Salvado, T. Alvaro, C. Callau, A. Roso, and J. Jaen, "Digital image analysis in breast cancer: an example of an automated methodology and the effects of image compression." *Studies in health technology and informatics*, vol. 179, pp. 155–171, 2011. doi: 10.3233/978-1-61499-086-4-155

[8] M. Apro, D. Novakovic, S. Pal, S. Dedijer, and N. Milic, "Colour space selection for entropy-based image segmentation of folded substrate images," *Acta Polytechnica Hungarica*, vol. 10, no. 1, pp. 43–62, 2013. doi: 10.12700/aph.10.01.2013.1.3. ISSN: 1785-8860.

[9] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution." *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291–299, Aug 2001. [Online]. Available: http://europepmc.org/abstract/MED/11531144

[10] A. Korzynska, L. Roszkowiak, C. Lopez, R. Bosch, L. Witkowski, and M. Lejeune, "Validation of various adaptive threshold methods of segmentation applied to follicular lymphoma digital images stained with 3,3'-diaminobenzidine&haematoxylin," *Diagnostic Pathology*, vol. 8, no. 1, p. 48, 2013. doi: 10.1186/1746-1596-8-48

[11] A. Lehmussola, P. Ruusuvuori, J. Selinummi, H. Huttunen, and O. Yli-Harja, "Computational framework for simulating fluorescence microscope images with cell populations," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 7, pp. 1010 –1016, july 2007. doi: 10.1109/TMI.2007.896925

[12] A. Lehmussola, P. Ruusuvuori, J. Selinummi, T. Rajala, and O. Yli-Harja, "Synthetic images of high-throughput microscopy for validation of image analysis methods," *Proceedings of the IEEE*, vol. 96, no. 8, pp. 1348 –1360, aug. 2008. doi: 10.1109/JPROC.2008.925490

[13] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–168, 2004. doi: 10.1117/1.1631315

[14] W. Niblack, *An introduction to image processing*. Englewood Cliffs, NJ: Prentice-Hall International, 1986.

[15] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225 – 236, 2000. doi: 10.1016/S0031-3203(99)00055-2

[16] J. Bernsen, "Dynamic thresholding of gray-level images," in *ICPR'86: International Conference on Pattern Recognition*, 1986.

[17] J. M. White and G. D. Rohrer, "Image thresholding for optical character recognition and other applications requiring character image extraction," *IBM Journal of Research and Development*, vol. 27, no. 4, pp. 400 –411, july 1983. doi: 10.1147/rd.274.0400

[18] P. W. Palumbo, P. Swaminathan, and S. N. Srihari, "Document image binarization: Evaluation of algorithms," *Proc. SPIE, Applications of Digital Image Processing IX*, vol. 697, no. 278, pp. 278–285, December 1986. doi: 10.1117/12.976229

[19] Y. Yasuda, M. Dubois, and T. Huang, "Data compression for check processing machines," *Proceedings of the IEEE*, vol. 68, no. 7, pp. 874 – 885, july 1980. doi: 10.1109/PROC.1980.11753

[20] U. Neuman, A. Korzynska, C. Lopez, M. Lejeune, L. Roszkowiak, and R. Bosch, "Equalisation of archival microscopic images from immunohistochemically stained tissue sections," *Biocybernetics and Biomedical Engineering*, vol. 33, no. 1, pp. 63–76, 2013. doi: 10.1016/S0208-5216(13)70056-1

DOI: 10.15439/2015F312
ACSIS, Vol. 5

# Exploring Medical Curricula Using Social Network Analysis Methods

Martin Víta
NLP Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
Email: 333617@mail.muni.cz

Martin Komenda, Andrea Pokorná
Institute of Biostatistics and Analyses,
Faculty of Medicine, Masaryk University
Kamenice 126/3, 625 00 Brno, Czech Republic
Email: {komenda, pokorna}@iba.muni.cz

*Abstract*—**This contribution demonstrates how to apply concepts of social network analysis on educational data. The main aim of this approach is to provide a deeper insight into the structure of courses and/or other learning units that belong to a given curriculum in order to improve the learning process.**

**The presented work can help us discover communities of similar study disciplines (based on the similarity measures of textual descriptions of their contents), as well as identify important courses strongly linked to others, and also find more independent and less important parts of the curriculum using centrality measures arising from the graph theory and social network analysis.**

## I. Introduction

NOWADAYS, the process of improving educational curricula requires more sophisticated analyses of relevant datasets than in previous decades. Using various powerful virtual learning environments (see [7], [9], [14], [10], [5]) authorial teams consisting of curriculum designers and guarantors are able to construct a detailed description of each lecture, seminar and practice. The current curricula usually represent a huge amount of data records (thousands of standard pages in total), which cover all necessary requirements on the graduates based on a predefined structure in an online environment including formal and semantic verification. The main objective of this paper is to present an innovative approach of exploring the (medical) curricula in a transparent way. This approach is based on a proven concept of social network analysis (SNA), since SNA provides a natural way of dealing with graph notions such as centrality, and with various models of importance. The primary motivation for these investigations is to create a suitable tool/methodology for anyone who is involved in the complicated process of curriculum design at institutions of higher education. The research can help us when answering – among others – the following questions:

- Are there any content overlaps and/or gaps in the curriculum? (The overlaps might be desirable in some cases and undesirable in others).
- Which disciplines/courses/learning units are similar in terms of contents? And which are "self-standing"?
- Are there some significant communities (clusters) of disciplines/courses/learning units (such that a change of one may influence others)?
- Which disciplines/courses/learning units are probably the "central" parts of the curriculum?

The results of our analysis are intended to be used by the (human) experts responsible for the development and further evaluation of the curricula and institutional management, nevertheless they could also be interesting for students who want to study efficiently and focus mainly on the important parts of the study plan. We are going to introduce a methodology/data mining process that is fully compatible with the process of standardised knowledge discovery in databases (KDD). The data mining process was inspired by Trigo and Brazdil's works [16] and [2].

Our basic requirement on the selected approach is the simplicity and re-usability in practice. We use only a commonly available and free software (R), during the modelling stage, so the process can be repeated without any notable expenses. The possibility of straightforward visualisation is one of the advantages of this approach: we are able to show these content-based relations among textual descriptions of medical education (namely disciplines), not only the formal or organisational relations. Such visualisation can provide a quick and smart overview of a study plan and provide comprehensive information for the subsequent global in-depth curriculum inspection, which could be used for the future planning and changes in the curricula.

## II. Methodology

Data exploration was done in accordance with a proven KDD background, namely the standardised methodology CRISP-DM (CRoss-Industry Standard Process for Data Mining) [1]. This process is entirely independent of selected modeling tools and consists of a cycle that involves six stages (Fig. 1). Each stage represents an independent issue which must be completed in order to move forward. This methodology is used in a wide range of applications including biosciences, industry, even finance [12]. All of the CRISP-DM steps are described below for our setting.

We have already proposed a complex curriculum planning model supporting the outcome-based paradigm [8], which promotes a clear communication between the involved stakeholders (teachers, guarantors, curriculum designers, supervisors and faculty management) [11]. Based on a robust web-oriented

978-83-60810-66-8/$25.00©2015, IEEE

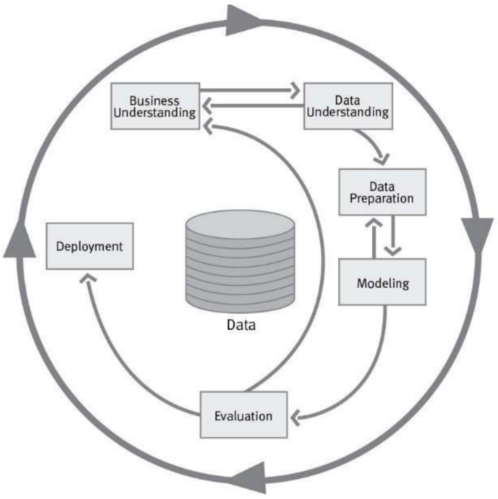


Fig. 1. Diagram of CRISP-DM.

platform for complex curriculum management which provides an effective tools for creating, transparent browsing, and reviewing the curriculum, a correctly compiled and balanced description of the General Medicine study field was defined by the authorial team consisting of 384 guarantors and teachers of the Faculty of Medicine at Masaryk University. It has covered more than 1300 learning units and 7100 learning outcomes, i.e. approximately 2500 standard pages of text. With respect to human cognition abilities, it is not possible to carefully read and verify the content of all learning units with all their linkages and co-dependencies. We have decided to apply the CRISP-DM methodology as a general framework to obtain a deeper knowledge from the medical educational data.

### A. Business understanding

Business understanding is an initial stage, which is focused on understanding the objectives and the problem definition – in our case, in terms of the in-depth exploration of medical curriculum in-depth exploration. The proposed goal is to detect outlying and overlapping areas in the General Medicine study field as well as investigate the mutual similarities of the courses involved. The whole study field is split into four individual modules (Diagnostic Sciences and Neurosciences, Internal Medicine, Surgical Sciences and Theoretical Sciences). These modules consist of 44 medical disciplines (such as Anatomy, Stomatology, Neurology, etc.). In total 144 courses are assigned to one or more disciplines are grouped together on a level of these disciplines for the purpose of this analysis. In this way, each discipline is generated from content-related courses across the entire study field (e.g. the discipline of Anatomy contains the following courses: Anatomy I – lecture, Anatomy I – seminar, Anatomy II – lecture, Anatomy II – practice).

### B. Data understanding

The follow-up stage – called data understanding – begins with the initial collection of data. Each course is described by a set of learning units using textual parameters and descriptors (see Table I). In our particular case, we have detected a set of descriptive attributes related to the learning outcomes, which were defined in accordance with the Bloom's taxonomy [13]. The Bloom's scheme provides a standardised classification of educational objectives that gives a commonly understood meaning to objectives classified in one of six main categories and many subcategories, thereby enhancing communication and achievement of more complex skills and abilities.

TABLE I
SELECTED ATTRIBUTES OF A LEARNING UNIT.

| Attribute | Type |
|---|---|
| learning_unit | varchar (255) |
| total_range | int (10) |
| meaning | text |
| annotation | text |
| mesh_keyword | varchar (255) |
| significant_term | varchar (255) |
| learning_outcome | varchar (500) |
| grouped_outcome | varchar (500) |
| primary_index | varchar (500) |
| secondary_index | varchar (500) |
| assessment_form | varchar (200) |

We have identified several attributes of learning outcomes mined from our curriculum management system, which can provide complete information about the coverage among various medical disciplines. All text data that we used was in English due to easier preprocessing steps.

*Teaser of textual data to be processed:*

```
Section:
-----
Surgical Sciences

Medical discipline:
-------------
Surgery

Group outcome:
---------
Abdominal aorta and its branches
Primary index: Arterial diseases
Secondary index: Closure, stenosis

Learning outcomes:
------------
- Student masters anatomy of arterial system.
- Student describes injuries to the abdominal aorta
and its branches.
- Student lists anatomy of abdominal aorta
and its branches.
```

### C. Data preparation

The stage of data preparation is a necessary prerequisite before applying any mathods of data mining and/or SNA. It consists of a sequence of procedures to create the final dataset from the initial raw data.

Each discipline was represented by a single plaintext file that contains merged contents of several fields from Table I, namely learning_outcome, grouped_outcome, primary_index, secondary_index.

The collection of these plaintext files was loaded as a corpus into the `R` system extended by a couple of packages `tm` and `lsa` – standard packages for text mining and latent semantic analysis. After tokenization at this stage, several standard preprocessing issues were performed – in the following sequence:

- transformation to lowercase,
- stemming (using Snowball),
- punctuation removal,
- numbers removal,
- stopwords removal,
- whitespace stripping.

Similarly to the work by [16] we have chosen a bag-of-words representation of the documents in the corpus and a document-term-matrix was generated. (*tf-idf* weighting [6] was used). Consequently, dissimilarity matrices were computed on the basis of cosine similarity. Values were rounded to two decimal digits and values lower than a certain threshold were replaced by zeros, since extremely low similarities were considered as irrelevant.

### *D. Modeling*

Generaly, various modeling techniques are selected and applied at this stage – SNA methods were applied in our case. Since some of the algorithms involved have several parameters, this stage also contains experiments with different values of these parameters.

*Application of social network analysis – centrality concepts on the similarity graph:* The dissimilarity matrix can be viewed as an adjacency matrix of a similarity graph – undirected graph with weighted edges. On this graph we are going to perform several calculations of centrality measures. Obviously, these values should be interpreted differently – with respect to the roles of disciplines/courses/learning units in the educational framework. We are going to deal with the following centrality measures:

*Closeness* – The closeness centrality of a node $v$ in a graph $G$ is defined by the inverse of the sum of the lengths of the shortest paths to/from all the other nodes in the graph $G$, i. e.:

$$c(v) = \frac{1}{\sum_{i \in V(G), i \neq v} d(i, v)},$$

where $d(i, v)$ is the length of the shortest path from node $i$ to node $v$ and $V(G)$ is the set of all vertices of the graph $G$.

If there is no path between a couple of nodes then the total number of nodes of the graph is used instead of the path length. By the mentioned calculation we obtain the so-called *raw closeness* of the node. To get normalised closeness of the node $v$, we multiply the raw closeness by $n - 1$, where $|V(G)| = n$: we are going to use this normalised version. Nodes (disciplines) with low value of closeness are those disciplines with their content distant from other ones, thus, roughly speaking, they are independent on the others.

*Betweenness centrality* – In the simplest case (without edge weighting), the raw betweenness centrality of a node $v$ corresponds with the number of shortest paths from all nodes to all others that go through a considered node, i. e.:

$$b(v) = \sum\nolimits_{i,j,v \in V(G), i \neq j, i \neq v, j \neq v} \frac{g_{ivj}}{g_{ij}},$$

where $g_{ij}$ is the total number of shortest paths going from node $i$ to $j$ and $g_{ivj}$ is the total number of all shortest from node $i$ to node $j$ going through $v$. To get normalized betweenness $b_n(v)$ of the node $v$, we calculate $b_n(v) = \frac{2b(v)}{(n-1)(n-2)}$, where again, $|V(G)| = n$.

This definition can be extended for weighted networks. Nodes (disciplines) with a high betweennes centrality are those that are the best for joining the students' knowledge from different collections of disciplines.

*Eigenvector centrality* – One of the methods of computing of approximate importance of a given node. The idea behind this measure is that centrality of each node is the sum of the centrality values of its neighbour nodes. More precisely, the eigenvector centrality values correspond to the values of the first eigenvector of the adjacency matrix. The eigenvector centrality in our case models identification of important disciplines of the curriculum.

All these computations are done using the `igraph` package [3] that contains several built-in functions covering the area of SNA. The values of these measures were normalised in a relevant way. According to our goal, our intention of this stage is to find out nodes (disciplines) with high values of proposed measures to identify the core and important parts of the curriculum and, in contrary, also the nodes (disciplines) with lowest values of centrality measures to identify those that are relatively independent of others.

Extremely low values of these attributes can also indicate a non-proper description of the discipline (e.g. missing important parts of the description). Disciplines with the most interesting results (means highest values of centrality measures attributes) are presented below.

*Ten disciplines with the highest betweeness centrality:*

1) Pathological physiology (0.265),
2) Physiology (0.164),
3) Surgery III (0.140),
4) Pathology (0.111),
5) Clinical examination in neurology (0.104),
6) Neuroscience (0.083),
7) Immunology (0.070),
8) Intensive care medicine (0.065),
9) Biology (0.065),
10) Preventive medicine (0.063).

*Ten disciplines with the highest closeness centrality:*

1) Pathological physiology (0.106),
2) Pathology (0.105),
3) Clinical examination in neurology (0.105),
4) Immunology (0.105),
5) Biology (0.105),
6) Surgery III (0.105),
7) Internal medicine – part 4 – Gastroenterology and haematology (0.105),
8) Intensive care medicine (0.105),
9) Clinical oncology (0.105),
10) Histology and embryology (0.105).

*Ten disciplines with the highest eigenvector centrality:*

1) Pediatrics II (1.000),
2) Pediatrics III (0.904),
3) Clinical oncology (0.867),
4) Surgery I-II (0.842),
5) Surgery III (0.819),
6) Pathology (0.782),
7) Internal medicine – part 4 – Gastroenterology and haematology (0.522),
8) Clinical examination in surgery (0.406),
9) Dermatovenerology (0.375),
10) Pathological physiology (0.358).

On the oposite side there are disciplines with low values.

*Disciplines with zero betweenness centrality:*

- Anatomy II,
- Basic medical terminology I,
- Clinical examination in internal medicine,
- Communication and selfexperience,
- Community medicine,
- Diagnostic imaging methods,
- Family medicine and geriatrics,
- Gynecology and obstetrics,
- Internal medicine – part 2 – Cardiology and angiology,
- Internal medicine – part 6 – Occupational medicine,
- Medical ethics I,
- Medical ethics II,
- Medical chemistry,
- Medical microbiology II,
- Medical psychology,
- Nursing,
- Ophthalmology,
- Orthopaedics,
- Pediatrics II,
- Pharmacology I,
- Stomatology.

*Ten disciplines with the lowest closeness centrality:*

1) Anatomy II (0.098),
2) Medical ethics I (0.017),
3) Medical ethics II (0.017),
4) Basic medical terminology I (0.017),
5) Communication and selfexperience (0.017),
6) Diagnostic imaging methods (0.017),
7) Family medicine and geriatrics (0.017),
8) Nursing (0.017),
9) Ophthalmology (0.017)
10) Stomatology (0.017).

*Ten disciplines with the lowest eigenvector centrality:*

1) Medical microbiology II (0.013),
2) Clinical examination in internal medicine (0.009),
3) Pharmacology II (0.006),
4) Medical psychology (0.004),
5) Anatomy III (0.004),
6) Anatomy II (0.001),
7) Pharmacology I (0.001),
8) Anatomy I (0.001),
9) Medical ethics II (0.000),
10) Medical ethics I (0.000).

*Interpretation:* In general, we show medical disciplines with extreme (the highest/the lowest) values of selected centrality measures. The achieved results represent novel and hopefully useful information about the structure of the General Medicine study field created in the curricula. For instance, Pathological physiology and Pathology appear in all top-ten lists in all centrality measures, Surgery III, Clinical examination in neurology, Immunology, Biology, Internal medicine – part 4 – Gastroenterology and haematology, Clinical oncology in two of them. It may indicate that these disciplines belong purposely to an essential part of the curriculum or, on the contrary, it may reveal an undesirable preference of mentioned disciplines, which in fact cannot be identified by a visual human inspection. So, logically the final interpretation has to be done under the supervision of responsible curriculum designers and senior guarantors, who are familiar with optimal composition and intersections of individual disciplines, courses and learning units.

*Application of social network analysis – community detection on the similarity graph:* The community detection is a common task when dealing with social networks. Communities, i.e. densely connected subgraphs have also an importance for exploring the curricula: communities correspond with subsets of mutually close disciplines (with respect to content similarity). In contrast to hard clustering we do not insist on the rule that each item (discipline) is contained in just one cluster (community).

For detecting communities, the Walktrap algorithm [15] was used. The main idea of this algorithm is that short random walks tend to stay in the same community, see [3]. The implementation of this algorithm is also contained within the `igraph` package in `R`, and for our purposes it was executed with a parameter of length of the random walk set to $k = 4$.

Fig. 2 provides a basic overview of communities that were found. The thickness of the edges corresponds with the similarity between nodes (disciplines). Communities are bounded shapes with a grey background color. As we can see, we have several "singleton" communities along with some bigger ones that establish the core of the curriculum.

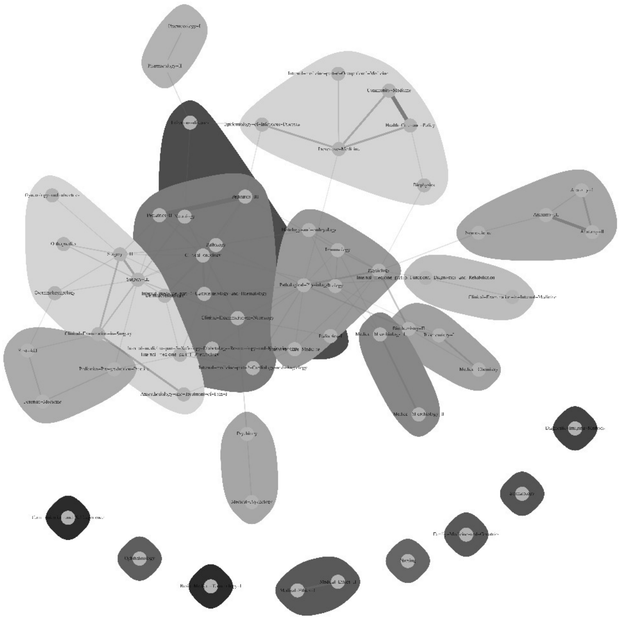

Pharmacology−I
Pharmacology−II
Internal−medicine−part−6−Occupational−Medicine
Community−Medicine
Infectious−diseases
Epidemiology−of−Infectious−Diseases
Health−Care−and−Policy
Preventive−Medicine
Biophysics
Gynecology−and−obstetrics
Pediatrics−III
Pediatrics−II
Neurology
Anatomy−I
Anatomy−III
Anatomy−II
Neuroscience
Histology−and−embryology
Orthopaedics
Surgery−II
Pathology
Clinical−oncology
Immunology
Physiology
Internal−medicine−part−5−Functional−Diagnostics−and−Rehabilitation
Surgery−III
Pathological−Physiology
Dermatovenerology
Internal−medicine−part−1−Gastroenterology−and−Hepatology
Clinical−Examination−in−Internal−Medicine
Clinical−Examination−in−Neurology
Clinical−Examination−in−Surgery
Pediatrics−I
Medical−Microbiology−I
Biochemistry−II
Biochemistry−I
First−Aid
Internal−medicine−part−3−Nephrology−Diabetology−Rheumatology−and−Endocrinology
Internal−medicine−part−4−Hematology
Intensive−Care−Medicine
Pediatrics−Pre−graduation−Practice
Internal−medicine−part−2−Cardiology−and−Angiology
Medical−Chemistry
Anaesthesiology−and−Treatment−of−Pain−I
Forensic−Medicine
Medical−Microbiology−II
Psychiatry
Diagnostic−Imaging−Methods
Medical−Psychology
Stomatology
Communication−and−Self−Experience
Family−Medicine−and−Geriatrics
Ophthalmology
Nursing
Medical−Ethics−II
Basic−Medical−Terminology−I
Medical−Ethics−I

Fig. 2. Overview of the curriculum with marked communities.

- Examples of larger communities: {Preventive medicine, Health care and policy, Community medicine, Biophysics, Internal medicine – part 6 – Occupational medicine, Epidemiology of infectious diseases}, {Medical chemistry, Biochemistry I, Biochemistry II }, . . .
- Examples of independent communities: {Nursing}, {Diagnostic-Imaging methods}, {Communication and self experience}, . . .

A detail of a community within the entire graph is shown in Fig. 3. There are also communities connecting just disciplines divided into two parts, such as Medical ethics I and Medical ethics II, Pharmacology I and II, etc. These results surely are not coincidental; quite the opposite, they confirm the reasonability of the method.

### *E. Evaluation and Deployment*

A checking procedure is performed in this stage in order to find the right meaning of analytical outputs. The obtained results were verified by representatives of the faculty management, in order to confirm the final interpretation. It may indicate either balanced or unbalanced representation of compulsory and optional courses intended for graduation to obtain professional qualification for employment as a physician. The curriculum visualisation based on the location of individual disciplines and their coloured marking is useful with regard to the possibility of a comprehensive evaluation of the curriculum structure. It makes it possible to identify weak points and shortcomings in terms of inconvenient interdisciplinary relations, as well as remote disciplines with hardly any relations to other disciplines. Additionally, the visualisation can very easily confirm the consistency of individual specialised disciplines, which leads to the idea of correctly specified requirements on the acquired knowledge and skills of future physicians. Currently, the final process of deployment is still ongoing. In the end, presented visualisation will be integrated directly into the curriculum management platform as an additional overview module.

## III. Conclusion and Future Work

In this work we have introduced a novel method for exploring general curricula that uses several concepts of social network analysis. The presented use-case provides an easy-to-understand visualisations of the entire medical curriculum and centrality models involved disciplines. The entire process of obtaining the knowledge is done according to an industrial standard in data mining (CRISP-DM). Our plans for further work are described below.

*1) Improving the similarity graph:* In this experimental stage, only a part of the accessible data was used

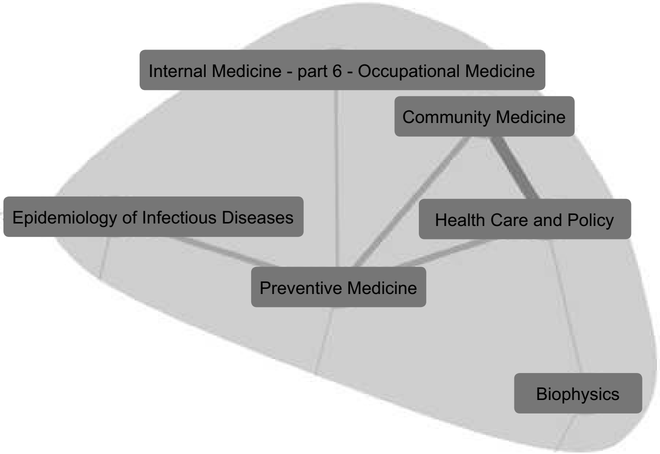

Internal Medicine - part 6 - Occupational Medicine
Community Medicine
Epidemiology of Infectious Diseases
Health Care and Policy
Preventive Medicine
Biophysics

Fig. 3. Detail of a given community of disciplines.

(learning_outcome, grouped_outcome, primary_index, secondary_index). Further development of this approach will be based on incorporating data such as the MeSH [4] terms. There is also an opportunity to experiment with weights: terms contained in certain controlled vocabularies, thesauri, ontologies etc. can be taken with boosted weights. The "NLP" stage will be improved in several ways, for example dealing with synonyms, hypernyms etc.

*2) Enriching the visualisation:* At this stage, we have only focused on the structure of the similarity graph and visualisation of obtained communities – that is, in fact, the core of the original work. Hence we have only used a small amount of attributes of the produced graph, i. e. the thickness of the edges (and vertices grouping). The remaining parameters of the visualisation can therefore carry additional formal, organisational or empirical data:

- The size of a node should correspond with the total number of teaching hours per discipline.
- The color of a concrete node can be chosen accordingly to a given classification of the discipline.
- The opacity should correspond with the importance of the node – the eigenvector centrality (or some other attribute of the discipline (for example, the ratio of students that fail the exams).
- Textual labels of the nodes can be used for information about the number of students attending the discipline.
- Edges on another layer (represented in a different color) can link a discipline having the same lecturers.

Both of these activities might lead to more information-rich visualisations that would provide a better insight to the specific curriculum which will be evaluated manually by experts in the near future.

## Acknowledgment

We would like to express our special thanks to the management of the Faculty of Medicine at Masaryk University, represented by Prof. Jaroslav Štěrba (Vice-Dean for Education in Clinical Branches) and Prof. Jiří Mayer (Dean of the Faculty of Medicine), and also to the Institute of Biostatistics and Analyses at Masaryk University represented by Assoc. Prof. Ladislav Dušek.

## References

[1] Azevedo, Ana Isabel Rojao Lourenço, 2008, KDD, SEMMA and CRISP-DM: a parallel overview. [online]. 2008. [Accessed 9 July 2013]. Available from: `http://recipp.ipp.pt/handle/10400.22/136`

[2] Brazdil, Pavel, Trigo, Luís, Cordeiro, Joao, Sarmento, Rui and Valizadeh, Mohammadraza, 2015, Affinity mining of documents sets via network analysis, keywords and summaries. Oslo Studies in Language, 7(1).

[3] Csardi, Gabor and Nepusz, Tamas, 2006, The igraph software package for complex network research. InterJournal, Complex Systems. 2006. Vol. 1695, no. 5, p. 1–9.

[4] Davis, Allan Peter, Wiegers, Thomas C., Rosenstein, Michael C. and Mattingly, Carolyn J., 2012, MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. Database. 2012. Vol. 2012, p. bar065.

[5] Erguzen, Atilla, Erel, Serafettin, Uzun, Ibrahim, Bilge, Hasan Sakir and Unver, Halil Murat, 2012, KUZEM LMS: A new learning management system for online education. Energy Education Science and Technology Part B-Social and Educational Studies. 2012. Vol. 4, no. 3, p. 1865-1878.

[6] Feldman, Ronen and Sanger, James, 2007, The text mining handbook: advanced approaches in analyzing unstructured data [online]. Cambridge University Press. [Accessed 2 May 2015].

[7] Frank, Jason R. and Danoff, Deborah, 2007, The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Medical teacher. 2007. Vol. 29, no. 7, p. 642-647.

[8] Harden, R. M., Crosby, J. R. and Davis, M. H., 1999, AMEE Guide No. 14: Outcome-based education: Part 1–An introduction to outcome-based education. Medical teacher. 1999. Vol. 21, no. 1, p. 7-14 (doi:10.1080/01421599979969).

[9] Kabicher, S. and Derntl, M., 2008, Visual Modelling for Design and Implementation of Modular Curricula. Zeitschrift für Hochschulentwicklung [online]. 2008. [Accessed 19 July 2012]. Available from: `http://www.zfhe.at/index.php/zfhe/article/view/64`

[10] Kerkiri, Tania Al and Papadakis, Spyros, 2012, Learning Outcomes Design Authoring Tool: The Educator is Not Alone! International Journal of e-Collaboration (IJeC). 2012. Vol. 8, no. 4, p. 22-34.

[11] Komenda, Martin, Schwarz, Daniel, Hřebíček, Jiří, Holčík, Jiří and Dušek, Ladislav, 2014, A Framework for Curriculum Management - The Use of Outcome-based Approach in Practice. In : Proceedings of the 6th International Conference on Computer Supported Education [online]. Barcelona: SCITEPRESS, p. 473-478. ISBN 978-989-758-020-8. Available from: `http://www.csedu.org`

[12] Korczak, Jerzy, et al. A-Trader-consulting agent platform for stock exchange gamblers. In: Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on. IEEE, 2012. p. 963-968.

[13] Krathwohl, David R., 2002, A revision of Bloom's taxonomy: An overview. Theory into practice. 2002. Vol. 41, no. 4, p. 212–218.

[14] Mong, Yu, Chan, Mangtang and Chan, Francis Kar Ho, 2008, Web-based outcome-based teaching and learning - An experience report. In: Advances in Web Based Learning—ICWL 2007. Berlin: Springer-Verlag Berlin. p. 475-483. ISBN 978-3-540-78138-7.

[15] Pons, Pascal and Latapy, Matthieu, 2005, Computing communities in large networks using random walks. In: Computer and Information Sciences-ISCIS 2005 [online]. Springer. p. 284–293.

[16] Trigo, Luís and Brazdil Pavel, 2014, Affinity Analysis between Researchers using Text Mining and Differential Analysis of Graphs, ECML/PKDD 2014 PhD session Proceedings, Nancy, France, p. 169–176.

[17] Uzunboylu, Höseyin, Bicen, Hüseyin and Cavus, Nadire, 2011, The efficient virtual learning environment: A case study of web 2.0 tools and Windows live spaces. Computers & Education. 2011. Vol. 56, no. 3, p. 720–726.

DOI: 10.15439/2015F151
ACSIS, Vol. 5

# On Integrating Clustering and Statistical Analysis for Supporting Cardiovascular Disease Diagnosis

Agnieszka Wosiak
Lodz University of Technology
Institute of Information Technology
ul. Wólczańska 215, 90-924 Łódź, Poland
Email: agnieszka.wosiak@p.lodz.pl

Danuta Zakrzewska
Lodz University of Technology
Institute of Information Technology
ul. Wólczańska 215, 90-924 Łódź, Poland
Email: danuta.zakrzewska@p.lodz.pl

***Abstract*—Statistical analysis of medical data plays significant role in medical diagnostics development. However in many cases the statistics is not effective enough. In the paper we consider combining statistical inference with clustering in the preprocessing phase of data analysis. The proposed methodology is checked on cardiovascular data and used for developing methods of early diagnosis of hypertension in children. Experiments, conducted on the real data, have demonstrated that the proposed hybrid approach allowed to discover relationships which have not been identified by using only the statistical methods. We have observed approximately 30% growth in the number of correlations between diagnosed attributes. Moreover all the obtained statistically significant dependencies were stronger in clusters rather than in the whole datasets.**

## I. Introduction

IN RECENT years medical progress as well as the equipment development make possible collecting the increasing amount of data. One can expect that their analysis will help medical practitioners in improving patient care, proposing new therapies or developing the existing ones. However, statistical analysis, which is commonly used to support medical diagnosis, in many cases, turns out to be not effective enough. Such situation takes place, when the correlations between parameters, which seem to be useful for medical inference, are not possible to obtain. For example, it might happen, when standard deviation in the dataset takes on the large value [1].

To improve the performance of statistical models, we propose the new approach, which consists in including clustering in the preprocessing phase. Both of the techniques have already been broadly investigated in medical applications, but their combination has not been examined as supporting tools for medical diagnosis so far. The presented approach enables to identify groups of similar instances, for which statistical models can be built effectively. Special attention is drawn to feature selection process. We assume that the set of attributes used in clustering and statistical analysis phases should be different, not correlated and consistent with the process of medical diagnosis as well as the state of art of the approaches to statistical analysis of the medical data (see [2] for example).

In the paper we focus on cardiovascular diseases, which are the leading causes of death in the majority of countries [3]. The research aims at developing methods for early diagnosis of hypertension (high blood pressure) in children. The proposed methodology was verified by experiments done on three different sets of real children cases. Experiment results showed that application of the cluster analysis effectively supports statistical inference for the diagnosis in the considered cardiovascular problem.

The remainder of the paper is organized as follows. In Section II relevant work is presented. Next, the medical issues of hypertension problems in cardiovascular diagnosis are introduced, then the proposed methodology is described. In the following section, the experiments conducted on real data are depicted. Finally, the results are discussed and some concluding remarks are presented.

## II. Related Work

In medicine most of the rules for diagnosis and treatment are based on statistical analysis. However, new challenges connected with medical data analysis impose application of more sophisticated methods such as data mining techniques which ensure the process improvement. Application of data mining techniques in biomedical and healthcare fields was discussed by Yoo et al. [4]. The authors stated that descriptive and predictive power of data mining could be widely used in these areas.

Cluster analysis has already been integrated with statistical methods for medical data in the research of Haldar [5]. However the goal of the study was not do discover new dependencies, but to define the phenotypes of clinical asthma. The research was proposed against other models of asthma classification and according to authors it might have played a supporting role for different phenotypes of heterogeneous asthma population. A survey of data mining methods that has been applied to traditional Chinese medicine (TCM) clinical data systems was provided by [6]. Cluster analysis, association rules, a latent structure model and a topic model were considered in the context of Chinese medicine.

In [1] different clinical decision support systems for heart disease prediction and diagnosis were compared. These systems were based on such data mining techniques as: a multiplayer perceptron, genetic algorithms, fuzzy rules, decision trees and Bayesian networks. As the result of investigations the authors stated that the considered techniques are not satisfactory and finally there is still lack of a solution for the

978-83-60810-66-8/$25.00©2015, IEEE

identification of treatment options for the patients with heart diseases.

In [7] a statistical inference of heart rate and blood pressure was examined. The authors considered three different approaches. The first one was based on examining correlation between raw data. Then since the measurements could be corrupted by noise, a filtration procedure was performed on data before correlating the signals. In the last approach least squares approximation was applied. The results of all of the techniques were similar. The obtained correlation coefficients seemed to be an unpredictable random numbers.

Meng et al. [8] compared the performance of logistic regression, artificial neural networks and decision tree models for predicting diabetes or prediabetes using common risk factors. The research proved the advantages of decision tree model comparing to the other considered techniques. The authors of [9] examined the performance of the alternate classification methods, such as bootstrap aggregation, boosting, random forests and support vector machines with conventional classification trees to classify patients with heart failure. Bashir et al. [10] proposed combination of three classifiers for intelligent heart disease diagnosis. Broad review of data mining techniques applied in this area was presented in [11].

## III. Hypertension Problems in Cardiovascular Diagnosis

Hypertension is the cardiovascular disease, which may have their onset in the young [12]. Hence arterial hypertension is a significant problem in pediatric practice. It is estimated that this pathology affects 3-5% of the total children population, while for teenagers the percentage of hypertension cases increases up to 10%. Therefore finding effective methods which support early diagnosis of hypertension and thus help in implementing an appropriate management to prevent the disease is currently the matter of interests of many researchers.

Hypertension is mainly defined on the basis of blood pressure measurements. However the initial cardiac data can be characterized by over 50 attributes. All patients undergo physical examination, manual arterial blood pressure measurements (RR SBP, RR DBP), ambulatory blood pressure monitoring (ABPM-S, ABPM-D), echocardiographic examination to evaluate cardiac function using standard parameters (ejection fraction, shortening fraction and myocardial performance index) and tissue Doppler examination (systolic mitral annular velocity profile and regional function parameters: velocity, strain, strain rate). A selection of attributes, which cardiologists use to diagnose arterial hypertension (see [13], [14] and [15]) is shown in Table I. The first two columns of the table contain names and descriptions of selected parameters, the third one presents ranges of attribute values.

To improve early detection of hypertension in children, the researchers look for new factors, which may indicate the high blood pressure appearance [16]. Medical data analysis helps in evaluating the characteristics of the variables in the data sets of healthy and diagnosed children and discovering the relationships between all the parameters. The detailed medical descriptions and statistical analysis of these issues were subjects of the research presented in [13], [14] and [15]. The authors stated that the high value of standard deviation in the dataset disabled obtaining some of the correlations between parameters, which could have been useful for medical inference. That fact motivated us to build methodology described in the presented paper.

TABLE I
The selection of parameters collected for arterial hypertension evaluation

| **Name** | **Description** | **Range** |
|---|---|---|
| HA | Arterial hypertension presence | Nom. (Yes/No) |
| BMI | Body mass index | 11.9 - 31.6 |
| BWT | Body weight | 14.6 - 88.0 |
| BSA | Body surface area | 1.4 - 2.2 |
| HC | Head circumference | 29.0 - 36.0 |
| PI | Ponderal index | 14.58 - 26.20 |
| RR SBP | Manual measurement of systolic blood pressure | 86 - 150 |
| RR DBP | Manual measurement of diastolic blood pressure | 44 - 87 |
| ABPM-S | Ambulatory systolic blood pressure monitoring | 19 - 87 |
| ABPM-D | Ambulatory diastolic blood pressure monitoring | 4 - 78 |
| IVSd | Interventricular septum | 5 - 14 |
| PWDs | Posterior wall thickness | 11 - 19 |
| TG | Triglyceride Level | 24 - 236 |
| E/A | Ratio of the early to the late mitral inflow velocities | 1.15 - 1.77 |
| DecT | Deceleration time - time interval of peak E-wave velocity to its extrapolation to the baseline | 126 - 325 |
| AEF | Atrial ejection force | 4570 - 9158 |

## IV. Materials and Methods

In medical research, an analysis of the results of observations plays the crucial role, as its effects are expected to be implemented into practical applications. The process of medical research is usually supported by statistical analysis but very often it is not effective enough. In many cases, dissimilarities or inconsistency within the data sets appear due to incorrect measurements or distortions. The presence of such deviations may lead to the rejection of true hypothesis in the case of small data sets. The use of clustering before conducting a statistical analysis allows to identify groups of similar cases, and thus to better evaluate respective parameters.

The proposed methodology of improving medical inference process from a medical data set consists of three main steps:

- feature selection, which enables choosing the set of attributes for building clusters,
- clustering based on the parameters indicated by the previous step to distinguish groups of similar characteristics,

- statistical analysis performed in clusters to find new dependencies between all the collected parameters.

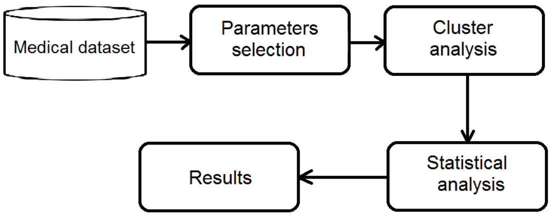


Fig. 1. System architecture for improved knowledge discovery using clustering techniques.

The general structure of our approach is shown on Fig. 1. We assume that clustering and statistical analysis are applied on the preprocessed and transformed data, and are preceded by feature selection process. The description of each step is presented in the following subsections.

### A. Feature Selection

Feature selection is a technique of choosing an appropriate subset of the available attributes, which ensures building the model with high classification accuracy. In medical data analysis there exist two main feature selection approaches: using of automatic feature selection mechanisms or selecting parameters as a result arising from the process of medical diagnosis. The first one was considered in [17]. The authors tried to provide a generic introduction to variable elimination, which can be applied to a wide range of machine learning problems. They considered filter, wrapper and embedded methods. However, they found out that comparison of feature selection methods can only be done on the data of the same characteristics.

Cheng, in turn, stated that in the case of cardiovascular diseases the feature subsets selected in the process of medical diagnosis improve the sensitivity of the analysis [18]. Such approach will be used in the proposed methodology, to obtain a set of attributes for further analysis.

### B. Clustering

Cluster analysis is one of the most commonly used data mining techniques, as it may be applied to classify complex data of many variables and many dimensions. Unlike discriminant analysis, no classification variables are inserted to divide the original data. What is more, in many cases, when the groups are detected, it is necessary to use other methods to discover the meaning of clustering [19]. Therefore, combination of cluster analysis and statistical inference seems to be the effective tool supporting medical diagnosis.

During investigations of the effectiveness of the proposed methodology, two different clustering approaches: deterministic and probabilistic were considered. As the presented technique aims at supporting physicians in making medical diagnosis, there were chosen simple comprehensible algorithms, because doctors should understand the tools they use. In the first group, k-means algorithm has been considered. In the case of medical data, this technique gives higher accuracy and lower root mean square error (RMSE) in comparison with other clustering methods, such as fuzzy C-means clustering, mountain clustering or subtractive clustering [20]. As the probabilistic method, expectation-maximization (EM) algorithm has been investigated. EM uses the finite Gaussian mixtures model to generate probabilistic descriptions of clusters in terms of means and standard deviations [19].

*1) The K-Means Algorithm:* The k-means algorithm is one of the most popular clustering method. The clusters in data set are defined by minimizing a distance (dissimilarity) function. In most of the cases Euclidean metric is considered as distance function [2],[21].

Let us consider the set of n data $X = \{x_i; i = 1, ..., n\}$ and the set $C$ of $k$ cluster centers $C = \{c_j; j = 1, ..., k\}$. For a given $k$, the goal of clustering is to find $C$ for which the function determined by (1) achieves its minimum.

$$min_j \left( \sum_{i=1}^{n} \|x_i - c_j\| \right) \tag{1}$$

The algorithm for k-means can be described as follows [21]:

1) Randomly choose $k$ data points from $X$ as the initial set $C$ of cluster centers. Denote them by $c_j, j = 1, ..., k$.
2) Reassign all $x_i \in X$ to the closest cluster mean $c_j$
3) Update all $c_j \in C$ as means of the points assigned to the corresponding clusters.
4) Repeat steps 2 and 3 until cluster assignments do not change.

Choosing $k$ initial centers at random, does not guarantee finding optimal clusters. To increase the chance of finding a global minimum in (1), it is usually suggested to run the algorithm several times with different initial choices and pick out the best final result - the one with the smallest total squared distance [21].

*2) The EM Algorithm:* The expectation - maximization (EM) algorithm is an iterative algorithm used to calculate maximum likelihood estimates in parametric models in the presence of missing data [22].

The goal of statistical models is to find the most likely set of clusters on the basis of training data and prior expectations. Expectation- Maximization algorithm (EM) uses the finite Gaussian mixtures model to generate probabilistic descriptions of clusters in terms of means and standard deviations [19]. The big advantage of EM algorithm is a possibility to select a number of clusters by cross validation techniques, what allows to obtain its optimal value [21]. That feature allows not to determine the number of clusters at the beginning. Similarly to k-means method, parameters are recomputed until the desired convergence value is achieved.

*3) Optimal Number of Clusters:* One of the most important issue connected with clustering is an identification of the optimal number of clusters. There exist different approaches to solve this problem.

In the case of k-means algorithm, the technique called elbow criterion has been considered. The elbow criterion says that one should choose a number of clusters, such that adding the next one does not increase quantity of information sufficiently [23]. When a graph for a validation measure calculated in the clusters is plotted against the number of clusters, at first amount of information is increasing, but at some point the gain starts decreasing, giving an angle in the graph, that is called the elbow.

In some cases, elbows cannot be unambiguously identified. Therefore it may be helpful to use another validation method for finding optimal number of clusters. As in the case of EM algorithm the optimal number of clusters can be determined by cross validation technique [21], the number of clusters indicated by elbow criterion can be confirmed by EM clustering.

It is worth mentioning that in medicine the number of clusters is very often equal to two as there exists common intention to split the whole data set into two groups [20]. Besides, when the number of considered instances is small, what very often takes place in medical applications, the bigger number of clusters would decrease the group sizes and as a consequence would make the medical inference less reliable as it is difficult to obtain sufficiently high power of statistical tests [2],[24].

### C. Statistical Analysis

Statistical data analysis usually begins with an assessment of measures of descriptive statistics, which allows to detect errors that were not identified during data preparation phase. The basic descriptors, for which the evaluation is indicated, include measures of central tendency (arithmetic mean, median and modal), measures of dispersion (range and standard deviation). Next statistical inference using a suitable test is carried out. The selection of the test is made on the basis of the type and the structure of the analyzed data. It depends on attribute types, scale type, number of experimental groups and their dependency, as well as the test power. The test selection should be done in accordance with the requirements of the USMLE (The United States Medical Licensing Examination). In the current research, we will consider the tests commonly used in medical diagnosis problems [1]:

- Kolmogorov-Smirnov test, which is used to test for normality of distribution of the attributes,
- Unpaired two-sample Student's *t*-test for the significance of a difference between two normally distributed values of attributes,
- Mann-Whitney U test, which is a non-parametric test for significant differences determination, where attributes were in nominal scales.

The impact of one variable measured in an interval or ratio scale to another variable in the same scale can be expressed using the Pearson's correlation coefficient $r_P(x, y)$. In the case where one or both of the variables are measured with an ordinal scale, or variables are expressed as an interval scale, but the relationship is not a linear one, the Spearman's correlation $r_S(x, y)$ test is used.

## V. Experimental Analysis and Results

The main objective of the experiments was to examine the performance of the proposed approach by comparing the results derived from statistical analysis carried out on clusters with the ones obtained for the whole datasets. The experiments were conducted on the real datasets, which were gathered for early diagnosis of arterial hypertension in children.

### A. Data Description

There have been considered three different datasets ("HEART", "ECHO", "IUGR") collected from children hospitalized in the University Hospital No 4, Department of Cardiology and Rheumatology, Medical University of Lodz. Each of the dataset was examined for the particular cardiovascular problem:

- "HEART" - to discover dependencies between arterial hypertension and left ventricle systolic functions,
- "ECHO" - to evaluate correlations between arterial hypertension and myocardial functions using tissue Doppler echocardiography,
- "IUGR" - to discover dependencies between abnormal blood pressure and being born as small for gestational age.

The "HEART" dataset consisted of 30 cases, the "ECHO" dataset of 66 instances and the "IUGR" dataset contained 50 specimens. There were no missing values within attributes.

### B. Cluster Analysis

In the first step of the experiments, the clusters for diagnosed children were created by using two clustering algorithms: k-means and EM implemented by WEKA Open Source software [21].

Clusters were built taking into account attributes according to feature selection method consistent with the process of medical diagnosis (see [2]). In the case of arterial hypertension, diagnosis performed by medical expert is mainly based on the blood pressure measurements (either manual or ambulatory monitored). The rest of the attributes are usually supportive for medical staff as each of them separately cannot indicate the disease and multivariate analysis is difficult to perform without any computer support. Therefore for "HEART" and "ECHO" datasets we considered 4 attributes: RR SBP, RR DBP, ABPM-S and ABPM-D in accordance with hypertension diagnosis (see Table I). In the case of "IUGR" dataset we used 3 attributes: birth weight, head circumference and ponderal index as consistent with the diagnosis of intrauterine growth restriction (being born as small for gestational age), and included 16 risk factors (i.a. hypertension in relatives, smoking during pregnancy) in a feature selection subset, which may have an impact on intrauterine growth restriction [28].

To choose the best number of clusters the elbow criterion has been applied. As validation measure, within cluster sum of squares has been considered. As the result, the charts for validation measures plotted against number of clusters indicated elbow points c=2 for "HEART" and "ECHO" datasets and c=3 for "IUGR" dataset (see Fig. 2).

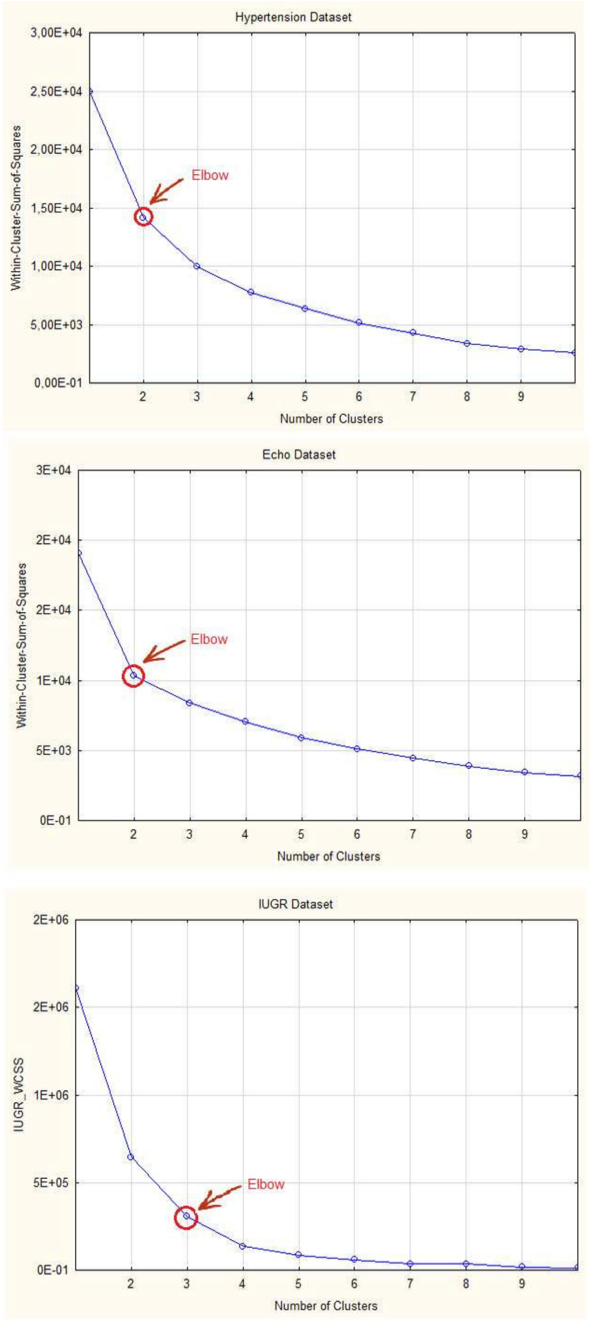


Fig. 2. Elbow charts for number of clusters determination.

Additionally to elbow criterion, EM algorithm which automatically generated number of clusters by using cross-validation [21] was implemented. The obtained results (c=2 for "HEART" and "ECHO" datasets and c=3 for "IUGR" dataset) confirmed the proper choice of the number of clusters.

It is worth mentioning that during clustering process, the subgroup characterized by higher mean values of parameters concerning arterial hypertension and lower standard deviations of those attributes or ,in the case of "IUGR" dataset, lower mean values of parameters concerning birth weight and size, was distinguished. In the case of "HEART" that subgroup consisted of 22 cases for k-means algorithm and 23 for EM method, "ECHO" subset included 44 and 35 instances respectively, "IUGR" subgroup contained 19 specimens for k-means algorithm and 25 for EM method. Sizes of the clusters for all the datasets are presented in Table II.

TABLE II
SIZES OF THE CLUSTERS

| Dataset | Numbers of instances in clusters | |
|---|---|---|
| | k-means clustering | EM clustering |
| HEART | 8, 22 | 7, 23 |
| ECHO | 22, 44 | 31, 35 |
| IUGR | 19, 14, 14 | 25, 15, 7 |

### C. Statistical Analysis

Analysis of correlations among datasets, in order to support medical knowledge acquisition and decision making, is still one of the most popular techniques in medical research [25]. Therefore next step of the experiments concerned indication of the attributes, which are not significantly correlated with the ones used for building clusters, and thus can be used in statistical analysis process. By insignificant correlation we mean values with correlation coefficient r<0.3 and p-value>0.05 ([26], [27]). Tables III, IV, V and Fig. 3 present correlation values of the features used in clustering and the ones indicated for statistical analysis.

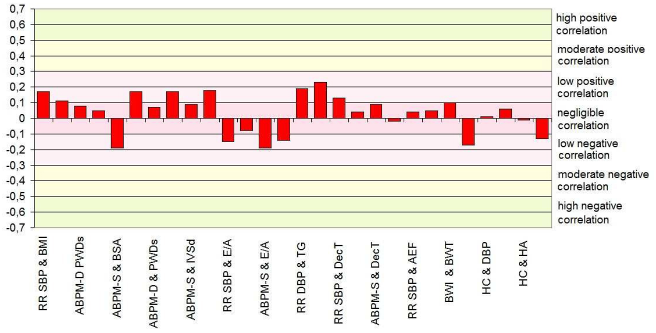


Fig. 3. Values of correlation coefficient between attributes used for clustering and statistical analysis.

Correlation values obtained for the clusters were compared to the ones got for the whole group of diagnosed children. Comparison of results confirmed effectiveness of the proposed methodology. For each dataset we obtained greater number of statistically significant correlations which may lead to improved medical diagnosis in the future. By significant correlations we mean values with correlation coefficient r>=0.3 and p-value<=0.05 ([26], [27]). The results of detected correlations are presented in Table VI, where the column (3) presents the numbers of discovered dependencies in clusters and the

TABLE III
VALUES OF CORRELATIONS BETWEEN FEATURES USED FOR CLUSTERING AND STATISTICAL ANALYSIS FOR "HEART" DATASET

| Attribute 1 | Attribute 2 | Corr. coeff. | p-value |
|---|---|---|---|
| RR SBP | BMI | 0.17 | 0.38 |
| ABPM-S | BMI | 0.11 | 0.55 |
| ABPM-D | BMI | 0.08 | 0.67 |
| RR DBP | BSA | 0.05 | 0.77 |
| ABPM-S | BSA | -0.19 | 0.31 |
| RR DBP | PWDs | 0.17 | 0.36 |
| ABPM-D | PWDs | 0.07 | 0.67 |
| RR SBP | IVSd | 0.22 | 0.23 |
| RR DBP | IVSd | 0.17 | 0.35 |
| ABPM-S | IVSd | 0.09 | 0.61 |
| ABPM-D | IVSd | 0.18 | 0.32 |

TABLE IV
VALUES OF CORRELATIONS BETWEEN FEATURES USED FOR CLUSTERING AND STATISTICAL ANALYSIS FOR "ECHO" DATASET.

| Attribute 1 | Attribute 2 | Corr. coeff. | p-value |
|---|---|---|---|
| RR SBP | E/A | -0.15 | 0.40 |
| RR DBP | E/A | -0.08 | 0.48 |
| ABPM-S | E/A | -0.19 | 0.12 |
| ABPM-D | E/A | -0.14 | 0.26 |
| RR SBP | TG | 0.24 | 0.10 |
| RR DBP | TG | 0.19 | 0.13 |
| ABPM-S | TG | 0.23 | 0.10 |
| RR SBP | DecT | 0.13 | 0.27 |
| RR DBP | DecT | 0.04 | 0.76 |
| ABPM-S | DecT | 0.09 | 0.47 |
| ABPM-D | DecT | -0.02 | 0.83 |
| RR SBP | AEF | 0.04 | 0.74 |
| RR DBP | AEF | 0.05 | 0.67 |

column (4) shows the percentage increase of correlations in comparison to the numbers of correlations for the whole dataset (column (2)).

Moreover the values of statistically significant correlations obtained after clustering were stronger than the corresponding values for the whole diagnosed group. Some of these correlations - as an overview of obtained results - are presented on Fig. 4 and in Table VII, where the column (2)- *"Correlation type"* contains the names of correlation parameters, the column (3) presents values of correlation coefficient and the last column (4) shows the p-value of significance level.

TABLE V
VALUES OF CORRELATIONS BETWEEN FEATURES USED FOR CLUSTERING AND STATISTICAL ANALYSIS FOR "IUGR" DATASET.

| Attribute 1 | Attribute 2 | Corr. coeff. | p-value |
|---|---|---|---|
| BWI | BWT | 0.10 | 0.49 |
| HC | BWT | 0.21 | 0.15 |
| PI | BWT | -0.17 | 0.23 |
| BWI | DBP | -0.25 | 0.10 |
| HC | DBP | <0.01 | 0.98 |
| PI | DBP | -0.23 | 0.11 |
| BWI | HA | 0.06 | 0.67 |
| HC | HA | -0.01 | 0.93 |
| PI | HA | -0.13 | 0.34 |

TABLE VI
NUMBER OF STATISTICALLY SIGNIFICANT CORRELATIONS DETECTED WITH AND WITHOUT CLUSTERING

| Dataset name (1) | Whole dataset (2) | Cluster (3) | Increase $[in\%]$ (4) |
|---|---|---|---|
| Clustering method: k-means | | | |
| HEART | 14 | 25 | 78% |
| ECHO | 13 | 16 | 23% |
| IUGR | 11 | 14 | 27% |
| Clustering method: EM | | | |
| HEART | 14 | 29 | 107% |
| ECHO | 13 | 17 | 31% |
| IUGR | 11 | 15 | 36% |

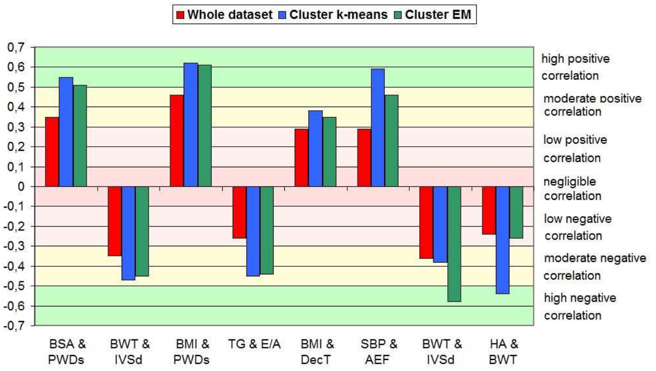


Fig. 4. Values of correlation coefficient for clusters in comparison to the whole dataset.

Concluding, the results of the experiments have shown that the proposed approach, which consists of supporting statistical inference by clustering, significantly improved the effectiveness of the medical data analysis for all the considered datasets.

## VI. CONCLUSIONS

Typically, during the process of computer-aided clinical and epidemiological studies only one of selected data analysis method is involved. In spite of the mostly used statistical analysis, in the paper, a hybrid methodology of medical data analysis has been proposed. The presented method consists of combination of clustering and statistical inference, where the first technique is used as a data preprocessing tool for the second one. The investigations have shown that supporting statistical analysis by clustering provides significant benefits. Experiments conducted on the real data have demonstrated that

TABLE VII
SAMPLE VALUES OF CORRELATIONS DETECTED WITH AND WITHOUT CLUSTERING

| **Dataset name (1)** | **Correlation type (2)** | **Correlation coefficient (3)** | **Significance (p-value) (4)** |
|---|---|---|---|
| Whole dataset | | | |
| HEART | BSA & PWDs | 0.35 | 0.05 |
| | BWT & IVSd | -0.35 | 0.05 |
| | BMI & PWDs | 0.46 | 0.01 |
| ECHO | TG & E/A | -0.26 | 0.03 |
| | BMI & DecT | 0.29 | 0.02 |
| | SBP & AEF | 0.29 | 0.02 |
| IUGR | BWT & DBP | -0.36 | 0.01 |
| | HA & BWT | -0.24 | 0.08 |
| Clustering method: k-means | | | |
| HEART | BSA & PWDs | 0.55 | 0.01 |
| | BWT & IVSd | -0.47 | 0.03 |
| | BMI & PWDs | 0.62 | <0.01 |
| ECHO | TG & E/A | -0.45 | 0.03 |
| | BMI & DecT | 0.38 | 0.01 |
| | SBP & AEF | 0.59 | <0.01 |
| IUGR | BWT & DBP | -0.38 | 0.01 |
| | HA & BWT | -0.54 | 0.04 |
| Clustering method: EM | | | |
| HEART | BSA & PWDs | 0.51 | 0.01 |
| | BWT & IVSd | -0.45 | 0.03 |
| | BMI & PWDs | 0.61 | <0.01 |
| ECHO | TG & E/A | -0.44 | 0.01 |
| | BMI & DecT | 0.35 | 0.05 |
| | SBP & AEF | 0.46 | 0.01 |
| IUGR | BWT & DBP | -0.58 | 0.01 |
| | HA & BWT | -0.26 | 0.05 |

the proposed hybrid method allowed to discover relationships which have not been identified previously.

Depending on the dataset, the growth of 10% - 100 % in the number of correlations between diagnosed attributes was obtained. Moreover all the calculated statistically significant dependencies were stronger in clusters rather than in the whole datasets.

The results of the presented investigations can be further implemented in practical diagnostic applications and can constitute the basis for improved medical inference described in [13], [14] and [15].

Future research will consist in developing the proposed methodology by considering some additional problems connected with medical data analysis including data gathering, data quality assurance, feature selection and outliers detection. The last one is especially worth considering as statistical analysis results are deviation sensitive. Therefore, the data need to be checked for outlier instances before proceeding with the analysis. The problem of handling outliers can be considered separately, or can be included as part of clustering process, but in such a case cluster analysis algorithm which deals with outliers should be implemented - random sample consensus (RANSAC) algorithm is regarded as giving satisfactory results [29].

Feature selection plays the crucial role in classification analysis. Although the choice of clustering attributes was carefully examined by medical experts and was consistent with the research presented in [1], we cannot exclude the possibility that considering other attributes may produce new meaningful conclusions. Therefore in future investigations we intend to verify this approach with different automatic feature selection methods, including genetic algorithms [30].

The analysis carried out as part of the current study involved data from laboratory tests, medical observations and their interpretations. However, the data for the analysis can be acquired by imaging studies. For the specified cardiac system the future analysis will concern SPECT images, ECG and EEG signals. Additional further studies will focus on efficient mining in medical imaging data and binding them with any other numerical and text data.

## REFERENCES

[1] S. U. Amin, K. Agarwal and R. Beg: "Data Mining in Clinical Decision Support Systems for Diagnosis", Prediction and Treatment of Heart Disease. Int J Adv Res Comput Eng Technol (IJARCET), 2008 vol. 2(1), pp. 218-223

[2] S.W. Looney and J.L. Hagan: "Statistical Methods for Assessing Biomarkers and Analyzing Biomarkers Data." In: C.R. Rao, J.P. Miller, D.C. Rao (eds): Essential Statistical Methods for Medical Statistics, Elsevier, 2011, pp. 27-65

[3] D.C. Davies, T. Moxham, K. Rees, S. Singh, A.J. Coats, S. Ebrahim, F. Lough and R.S. Taylor: "Exercise based rehabilitation for heart failure", Cochrane Database Syst Rev, 2010 vol. 4(1, pp. 1-57, DOI: 10.1002/14651858.CD001800.pub2

[4] I. Yoo, P. Alafaireet and M. Marinov: "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", J Med Syst, 2012, vol. 36, pp. 2431-2448, DOI: 10.1007/s10916-011-9710-5

[5] P. Haldar, I.D. Pavord, D.E. Shaw, M.A. Berry, M. Thomas, C.E. Brightling, A.J. Wardlaw and R.H. Green: "Cluster Analysis and Clinical Asthma Phenotypes", Am J Resp Crit Care, 2008, vol. 178, pp. 218-224, DOI: 10.1164/rccm.200711-1754OC

[6] X. Zhang, X. Zhou, R. Zhang, B. Liu, and Q. Xie: "Real-world Clinical Data Mining on TCM Clinical Diagnosis and Treatment: A Survey", e-Health Networking. Applications and Services (Healthcom), 2012 IEEE 14th International Conference on, DOI: 10.1109/HealthCom.2012.6380072

[7] A. Poliński, J. Kot A. Meresta: "Analysis of Correlation Between Heart Rate and Blood Pressure", In: IEEE Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS), 2011, pp. 417-420

[8] X.H. Meng, Y.X. Huang, D.P. Rao, Q. Zhang, and Q. Liu: "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", Kaohsiung J Med Sci, 2013, vol. 29, pp. 93-99, DOI: http://dx.doi.org/10.1016/j.kjms.2012.08.016

[9] P.C. Austin, J.V. Tu, J.E. Ho, D. Levy and D.S. Lee: "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes", J Clin Epidemiol, 2013, vol. 66, pp. 398-407, DOI:10.1016/j.jclinepi.2012.11.008

[10] S. Bashir, U. Quamar and M.Y. Javed: "An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis", In: International Conference on Information Society (i-Society), 2014, pp. 259-264, DOI: 10.1109/i-Society.2014.7009056

[11] M. Shouman, T. Turner and R. Stocker: "Using Data Mining Techniques in Heart Disease Diagnosis and Treatment", In: Japan-Egypt Conference on Electronics, Communications and Computers, 2012, pp. 173-177, DOI: 10.1109/JEC-ECC.2012.6186978

[12] B. Falkner, H. Kushner, G. Onesti and E.T. Angelakos: "Cardiovascular characteristics in adolescents who develop essential hypertension" Hypertension, 1981, vol. 3(5), pp. 521-527, DOI: 10.1161/01.HYP.3.5.521

[13] J. Zamojska, K. Niewiadomska-Jarosik, A. Wosiak, and J. Stanczyk: "Evaluation of left ventricular systolic function with the use of tissue Doppler echocardiography in children with primary arterial hypertension" ("Ocena funkcji skurczowej lewej komory z wykorzystaniem metody doplera tkankowego u dzieci z nadciśnieniem tetniczym pierwotnym"). Pol J Cardiol, 2012, vol. 4(2), pp.95-100 (in Polish)

[14] J. Zamojska, K. Niewiadomska-Jarosik, A. Wosiak, P. Lipiec, and J. Stanczyk: "Myocardial dysfunction measured by tissue Doppler echocardiography in children with primary arterial hypertension", Kardiol Pol (Polish Heart Journal), 2015, vol. 73(3), pp. 194-200, DOI: 10.5603/KP.a2014.0189

[15] A. Zamecznik, K. Niewiadomska-Jarosik, A. Wosiak, J. Zamojska, J. Moll and J. Stanczyk: "Intra-uterine growth restriction as a risk factor for hypertension in children six to 10 years old", Cardiovasc J Afr, vol. 25(2), 2014, pp. 73-77, DOI: dx.doi.org/10.5830/CVJA-2014-009

[16] J. Feber and M. Ahmed: "Hypertension in children: new trends and challenges", Clin Sci, 2010, vol. 119, pp. 151âĂ¿161, DOI: 10.1042/CS20090544

[17] G. Chandrashekar and F. Sahin: "A survey on feature selection methods", Computers and Electrical Engineering, vol. 40, 2014, pp. 16-28, DOI: dx.doi.org/10.1016/j.compeleceng.2013.11.024

[18] T.H. Cheng, C.P. Wei and V.S. Tseng: "Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches", Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems 2006, pp. 165 - 170, DOI: 10.1109/CBMS.2006.87

[19] J. Han, M. Kamber and J. Pei: "Data Mining: Concepts and Techniques", Elsevier, USA, 2011

[20] H. Liu, and L. Yu: Toward Integrating Feature Selection Algorithms for Classification and Clustering, IEEE T Knowl Data En, 2005, vol. 17, pp. 491-502, DOI: 10.1109/TKDE.2005.66

[21] I.H. Witten, E. Frank and M.A. Hall: "Data Mining. Practical machine learning tools and techniques", Morgan Kaufmann, San Francisco, USA, 2011

[22] A.P. Dempster, N.M. Laird and Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm". J R Stat Soc, 1977, vol. 39(1), pp. 1-38

[23] A.T. Azar, S.A. El-Said and A.E. Hassanien: "Fuzzy and hard clustering analysis for thyroid disease", Comput Meth Progr Bio, 2013, vol. 111(1), pp. 1-16, DOI: 10.1016/j.cmpb.2013.01.002

[24] Y.F. Wang, M.Y. Chang, R.D. Chiang, L.J. Hwang, C.M. Lee and Y.H. Wang: "Mining Medical Data: A Case Study of Endometriosis", J Med Syst, 2013, vol. 37:9899, DOI: 10.1007/s10916-012-9899-y

[25] N. Esfandiari, M.R. Babavalian, A.M.E. Moghadam and V.K. Tabar: "Knowledge discovery in medicine: Current issue and future trend", Expert Sys Appl, 2014, vol. 41(9), pp. 4434-4463, DOI: 10.1016/j.eswa.2014.01.011

[26] D.G. Altman and J.M. Bland: "Measurement in Medicine: the Analysis of Method Comparison Studies", The Statistician 32, 1983, pp. 307-317

[27] D.E. Hinkle, W. Wiersma and S.G. Jurs: Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin, 2003

[28] F. Figueras and J. Gardosi: "Intrauterine growth restriction: new concepts in antenatal surveillance, diagnosis, and management", Am J Obstet Gynecol, 2011, vol. 204.4, pp. 288-300, DOI: 10.1016/j.ajog.2010.08.055

[29] M.T. El-Melegy: "Model-wise and point-wise random sample consensus for robust regression and outlier detection". Neural Netw, 2014, vol. 59, pp. 23-35, DOI:10.1016/j.neunet.2014.06.010

[30] M.A. Jabbar, P. Chandra and B.L. Deekshatulu: "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection". 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, DOI:10.1109/ISDA.2012.6416610

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 311–320

DOI: 10.15439/2015F327
ACSIS, Vol. 5

# Introduction to Knowledge Discovery in Medical Databases and Use of Reliability Analysis in Data Mining

Elena Zaitseva, Miroslav Kvassay, Vitaly Levashenko, Jozef Kostolny
University of Zilina,
Faculty of Management Science and Informatics
Zilina, Slovakia
Email: {elena.zaitseva, miroslav.kvassay, vitaly.levashenko, jozef.kostolny}@fri.uniza.sk

□

*Abstract*—**Data mining (DM) is a collection of algorithms that are used to find some novel, useful and interesting knowledge in databases. DM algorithms are based on applied fields of mathematics and informatics, such as mathematical statistics, probability theory, information theory, neural networks. Some methods of these fields can be used to find hidden relation between data, what can be used to create models that predict some behavior or describe some common properties of analyzed objects. In this paper, we combine methods of DM with tools of reliability analysis to investigate importance of individual database attributes. Results of such investigation can be used in database optimization because it allows identifying attributes that are not important for purposes for which the database is used. Our approach is based on some coincidence between the key terms of DM and reliability analysis.**

## I. Introduction

ONE of the current problems of modern medicine is processing and analysis of a huge amount of data that are generated by medical systems. This calls for design of automatic or semi-automatic process that could be used to find useful and understandable knowledge from data. This process is known as a Knowledge Discovery in Databases (KDDs) and it involves an (semi-)automatic, exploratory analysis and modeling of large data repositories to identify valid, novel, useful, and understandable patterns from large and complex datasets [1].

The main idea of a KDD is to transform a huge amount of row data into useful information and knowledge that can be very easy interpreted. In general, data represent basic facts and statistics without any context. As an example, let us consider numbers “100” or “22.3” or simple value “no”. When we add a context to the data, then we get information. For example, a patient with plasma glucose concentration at 2 hours in an oral glucose tolerance test of 100 and body mass index (patient’s body mass in kg divided by the square of its height in m) of 22.3 does not suffer diabetes. So, the basic difference between data and information is in its information value, i.e. data have no information value themselves in compare to information.

If we have enough information from some domain, then it can be possible to identify some general facts that characterize the domain. These general facts are known as knowledge. For example, let us consider the example of relation between plasma glucose concentration, body mass index and diabetes of a patient. If we have such information about many patients, then, for example, we can derive that there is a little probability that patients with plasma glucose concentration under 127 and body mass index under 24.6 suffer diabetes. This knowledge can be identified in dataset [2] that contains 768 records about the diabetes incidence in the Pima Indian population living near Phoenix in Arizona.

The relation between data, information and knowledge is very often described by the knowledge pyramid (Fig. 1). It expresses that a huge amount of data can be transformed into information by adding a context and, then, analysis and aggregation of information can lead in the discovery of general patterns in data that represent knowledge about the studied domain. One of the most popular terms for the second phase is Data Mining (DM).

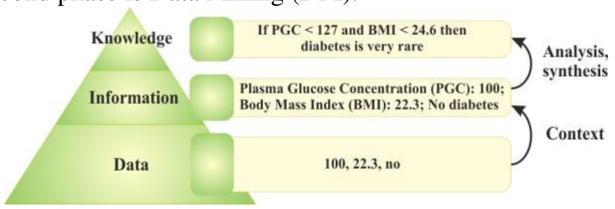

Knowledge
If PGC < 127 and BMI < 24.6 then diabetes is very rare
Analysis, synthesis
Information
Plasma Glucose Concentration (PGC): 100; Body Mass Index (BMI): 22.3; No diabetes
Context
Data
100, 22.3, no

Fig. 1 Knowledge pyramid

## II. Knowledge Discovery Process and Data Mining

In general, DM is a collection of methods focused on building model and finding patterns or trends in data. When DM is used in a process in which their outcome is evaluated, so that we can think about the product as being a new package of information, then we speak about a knowledge discovery process [3].

□ This work was supported by the grant of 7th RTD Framework Program No 610425 (RASimAs).

978-83-60810-66-8/$25.00©2015, IEEE

A knowledge discovery process or KDD is a very complex non-linear process that involves not only data analysis but also its preparation as well as knowledge interpretation and using the discovered knowledge. A KDD involves six important steps that are [3], [4]: understanding the problem domain, understanding the data, preparation of the data, data mining, evaluation of the discovered knowledge and using the knowledge. The term non-linear means that, in any step, there can be identified some problems that cause need to return to some of the previous steps and repeat the whole process from that step.

A KDD is not a one-pass process [3], and its every iteration gives a different view on the data. For example, in the preprocessing phase, we can identify quite a lot of missing data, so their prediction could become the goal of the first iteration of the KDD. In the second iteration, we can focus on the creation of a model that could be used to predict patient's diagnosis from that data (predictive DM) or to identify patients with similar relations between their symptoms and diagnosis (descriptive DM).

### A. *Understanding the Medical Problem Domain*

At the beginning of a KDD, we need to identify what type of knowledge should be found in the data. This requires understanding the problem domain. In case of medical data, the domain comes from medical area and, therefore, the main goals of this phase are [3]:

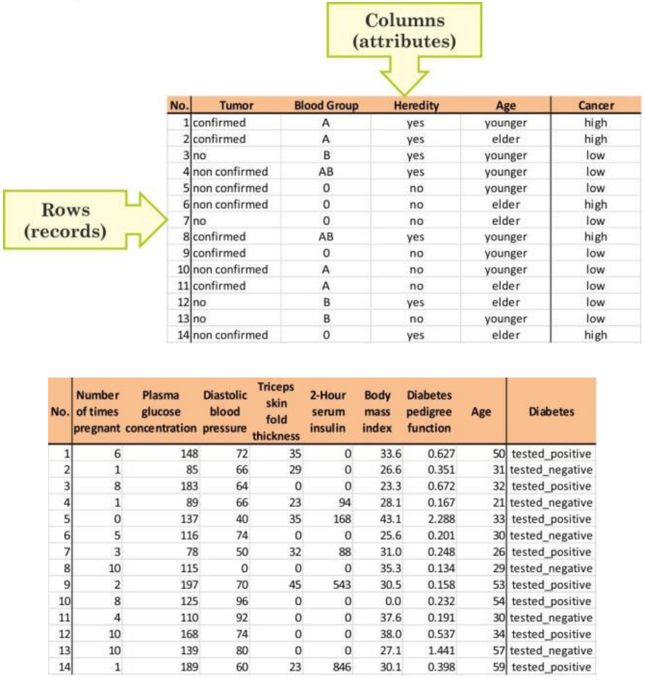
a) translation of medical goals into DM ones, and
b) determination of success criteria from the medical and DM point of view.

### B. *Understanding the Data*

When we understand the medical problem domain, then the data that are available should be analyzed. This analysis identifies which data will be used, and which additional information will be needed. The main goal of this step is to create a dataset for the next steps of KDD.

The result of this step can be very often interpreted as a table with rows and columns. The columns represent individual attributes of analyzed data while rows agree with individual records (instances/patients). Examples of two datasets are in Fig. 2. The first one contains 5 attributes related to cancer and it stores information about 14 patients. The second table is a sample of dataset focused on the diabetes incidence in the Pima Indian population [2]. It has 9 attributes and 14 records (the all dataset has 768 rows).

In Fig. 2, we can see that two different types of attributes exist – categorical and numerical [5] (Fig. 3). Categorical attributes are non-numerical and, usually, they have several possible values. They are also known as qualitative because they describe an object from qualitative point of view. The categorical attributes can be split into two separate groups: nominal and ordinal. The basic difference between them is that nominal attributes contain data that cannot be ordered, i.e. there is define no relation such as "greater/better than" or "lower/worse than", while ordinal ones are defined on data that can be ordered in some way. As an example, let us consider blood groups (A, B, AB, 0) and pain degree (severe, mild, none). There is no reason to assume that blood group A is better than 0 but, in case of pain degree, it is clear that no pain is better than severe one.


Columns (attributes)
Rows (records)

| No. | Tumor | Blood Group | Heredity | Age | Cancer |
|---|---|---|---|---|---|
| 1 | confirmed | A | yes | younger | high |
| 2 | confirmed | A | yes | elder | high |
| 3 | no | B | yes | younger | low |
| 4 | non confirmed | AB | yes | younger | low |
| 5 | non confirmed | 0 | no | younger | low |
| 6 | non confirmed | 0 | no | elder | high |
| 7 | no | 0 | no | elder | low |
| 8 | confirmed | AB | yes | younger | high |
| 9 | confirmed | 0 | no | younger | low |
| 10 | non confirmed | A | no | younger | low |
| 11 | confirmed | A | no | elder | low |
| 12 | no | B | yes | elder | low |
| 13 | no | B | no | younger | low |
| 14 | non confirmed | 0 | yes | elder | high |

| No. | Number of times pregnant | Plasma glucose concentration | Diastolic blood pressure | Triceps skin fold thickness | 2-Hour serum insulin | Body mass index | Diabetes pedigree function | Age | Diabetes |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | tested_positive |
| 2 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | tested_negative |
| 3 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | tested_positive |
| 4 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | tested_negative |
| 5 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | tested_positive |
| 6 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | tested_negative |
| 7 | 3 | 78 | 50 | 32 | 88 | 31.0 | 0.248 | 26 | tested_positive |
| 8 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | tested_negative |
| 9 | 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | tested_positive |
| 10 | 8 | 125 | 96 | 0 | 0 | 0.0 | 0.232 | 54 | tested_positive |
| 11 | 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | tested_negative |
| 12 | 10 | 168 | 74 | 0 | 0 | 38.0 | 0.537 | 34 | tested_positive |
| 13 | 10 | 139 | 80 | 0 | 0 | 27.1 | 1.441 | 57 | tested_negative |
| 14 | 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 | 59 | tested_positive |

Fig. 2 Examples of medical datasets

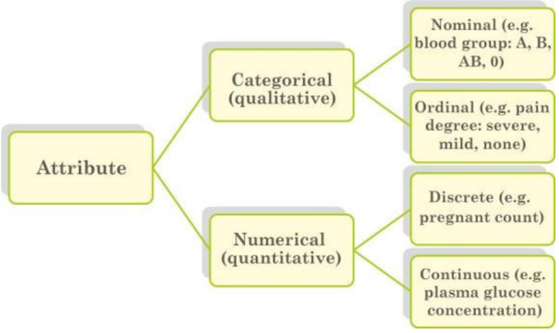

Attribute
Categorical (qualitative)
Numerical (quantitative)
Nominal (e.g. blood group: A, B, AB, 0)
Ordinal (e.g. pain degree: severe, mild, none)
Discrete (e.g. pregnant count)
Continuous (e.g. plasma glucose concentration)

Fig. 3 Types of attributes

Numerical attributes are another class. They are expressed in the form of numbers. These attributes define object properties from quantitative point of view and, therefore, they are also known as quantitative attributes. They can be subdivided into two groups: discrete and continuous. Discrete attributes are often defined on a set of some whole numerical values and very often count numbers of some events. As an example, let us consider the number of times when a woman has been pregnant or the birth year of a patient. When the attribute is continuous, then it means there is no limitation (except the lower and upper limits) on the values that can be taken. Typical examples are patient's height, weight or plasma glucose concentration.

One of the typical problems of KDD that is related to data is how much data is optimal for DM algorithms (Fig. 4).

There exists no definitive answer but, in general, more records are better because we can avoid the problems of underfitting, when the created model is too simple to analyze data that do not come from the original dataset. Similarly, datasets with less attributes are better than ones with many because the discovered knowledge can be interpreted easier.

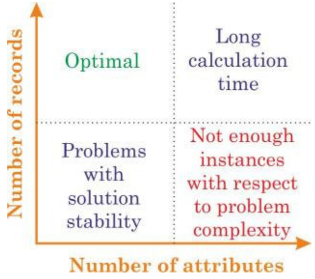

Optimal
Long calculation time
Problems with solution stability
Not enough instances with respect to problem complexity
Number of records
Number of attributes

Fig. 4 Problems of DM related to the number of attributes and records

### C. Preparation of the Data

When we understand the medical problem domain and the data on which KDD will be performed, we can prepare them for DM algorithms. This is one of the most important and the most time consuming steps of KDD [3]. It consists of two phases: data cleansing and data transformation.

The quality of data is the most important factor on which the success of KDD depends. Real databases contain a lot of data. However, these data can include incorrect or missing values. If there are a lot of such values, then the result of a DM algorithm will be a model that is unusable in practice. Therefore, data cleansing is very important step. Its main task is to enhance data reliability. There exist a lot of methods that can be used for this purpose. The simplest ones are based on the assumption that most of the data are correct and, therefore, incorrect data can be handled simply by its removing. More sophisticated methods involve some statistical methods to identify and replace incorrect or missing values and the most complex ones use supervised DM algorithms to predict the correct value of an attribute.

When the data have required quality, then we can prepare them for DM algorithms. This phase involves production of new attributes (when given attributes are not very appropriate for DM) and reduction of attribute count (models created from huge amount of attributes are usually very complicated, which results in problems with interpretation of gained knowledge; also, such models are usually very accurate for data from which they have been produced, but inaccurate for new data and, therefore, their deployment can be very problematic).

New attributes can be produced from one or more existing attributes. Therefore, we distinguish between one-attribute and multi-attribute transformations. Typical examples of one-attribute transformations are: normalization (mapping continuous data to values in interval <0, 1>), percentages (values are related to a specified base value), scores (transformation of ordinal data to discrete ones), etc. Examples of multi-attribute transformations are ratios (quotient of two attributes, e.g. using body mass index instead of weight and height), rates (number of event occurrences divided by time, e.g. number of cigarettes per day), and other linear and nonlinear combinations.

There exist several approaches for reducing the number of attributes. Very often, medical expert knowledge can be sufficient to solve this problem. Another alternative is to use some methods of mathematical statistic such as principal component analysis, kernel principal component analysis, independent component analysis [6], etc.

### D. Data Mining

At the beginning of this phase, the appropriate DM task has to be chosen (Fig. 5). There exist two different goals, for which DM can be used: description and prediction [7].

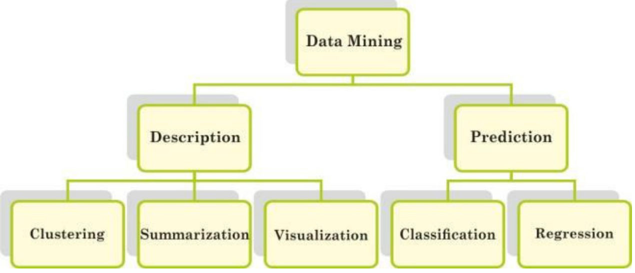


Fig. 5 Basic data mining models

Descriptive methods create models that are used for better understanding of given dataset. Typical examples are clustering, summarization and visualization [3]. The main idea of clustering is to find natural clusters of objects in a dataset. Objects are grouped together if they are similar to one another and dissimilar from objects in other clusters. Summarization is focused on data aggregation that is useful if we want to find some global characteristics of the entire dataset. The global characteristics allow describing data without necessity of knowing exact values of attributes of individual objects in dataset. So, they reduce the dataset size in terms of attributes or records count. Visualization includes techniques that aim is to simplify data understanding.

Predictive methods are used when the attributes can be subdivided into two groups: input and output attributes. In this case, DM can be used to discover the relationship between inputs and output attribute. (For example, the second dataset in Fig. 2 contains 8 input attributes, which more or less relates to diabetes, and 1 output attribute that identifies whether the patient suffer diabetes or not.) Based on the possible values of the output attribute, two types of prediction can be recognized: classification and regression [7]. The former maps the input space into predefined classes or, in general, into a discrete-valued domain, i.e. the output attribute is categorical or discrete numerical (Fig. 3). Typical algorithms of classification include neural networks (existing dataset is used to create and train a neural network that will be used to classify new records), decision trees (existing dataset is used to create a decision tree that will be capable to correct classify new records), instance based learning (every new record is classified according to its similarity with records that have already been classified), etc. Regression models transform the space of input attributes into a real-valued domain, i.e. the output attribute is continuous numerical according to Fig. 3, and the typical examples are linear and non-linear regression.

When we identify the appropriate DM task, then we have to choose and employ DM algorithm that will be used to achieve the goal, e.g. for clustering, we can select from statistical methods, support vector clustering, *k*-means clustering, hierarchical clustering, etc. Every algorithm has some parameters that have to be set correct to get satisfied result, e.g. how many clusters do we want to create, what is the minimal size of a cluster. These parameters are usually obtained by running the algorithm more times with different values of parameters and analyzing the obtained results.

### *E. Evaluation of the Discovered Knowledge*

The result of DM phase is a model that can be used to describe given data or to predict values of some attributes. When we want to interpret the model, then we have to understand the results that are described by it. This means that we have to be able to interpret the results from the medical point of view. This allows us to understand the discovered knowledge and identify whether it is novel and interesting. If the obtained knowledge is novel, then we can check its impact on the medical goal determined at the beginning of the KDD and recognize its usefulness in medical environment.

### *F. Using the Discovered Knowledge*

Finally, when the model is evaluated, then it should be incorporated into another system that will use the knowledge represented by the model. The success of this step determines the effectiveness of the entire KDD because the KDD has been unnecessary without further use of the discovered knowledge. According to [7], there are many challenges in this step, such as losing the “laboratory conditions” under which the model has been created. This means that the model has been produced from a certain static dataset, but the data become dynamic after the model deployment.

The result of KDD is knowledge that can be used for description purposes or for prediction. In case of predictive model that is created from a dataset containing only categorical attributes, the discovered knowledge can be represented as a table that enumerates all combinations of values of input attributes and defines value of the output for each of them. Such kind of table is used also in reliability analysis, and it is known as the structure function. This indicates that some tools of reliability analysis could be used in the analysis of the discovered knowledge.

## III. Reliability Analysis

Reliability is an important characteristic of systems. Every system consists of one or more components (basic parts of the system that are assumed to be indivisible into smaller elements). One of the principal tasks of reliability analysis is investigation of influence of individual system components on system activity and identification of components that are most important for system proper work. This investigation is known as importance analysis.

### *A. Binary- and Multi-State Systems*

Before the importance analysis can be performed, a model of the system has to be created. As a rule two types of models are used in reliability analysis. The first one is known as a Binary-State System (BSS). This model is based on the assumption that the system and all its components can be in one of only two possible states – functioning (labelled by number 1) and failure (represented by number 0). The dependency between states of individual system components and system state is expressed by a special relation that is known as structure function. The structure function of a BSS has the following form [8], [9]:

$$\phi(x_1,...,x_n) = \phi(\boldsymbol{x}) : \{0,1\}^n \to \{0,1\}, \tag{1}$$

where $n$ is a number of system components, $x_i$ is a variable denoting state of the $i$-th component and $\boldsymbol{x} = (x_1,\ldots, x_n)$ is a vector of states of system components (state vector). The structure function of a BSS can be viewed as a Boolean function and, therefore, some approaches related to analysis of Boolean functions can be used in the analysis of BSSs [9].

BSSs have been widely used in reliability analysis, especially in the analysis of systems in which any deviation from perfect functioning results in failure of the system, e.g. nuclear power plants [10], aviation systems [11]. However, these models are not very appropriate for systems that can operate at different performance levels, i.e. systems that can meet their mission also when they are not perfectly functioning, e.g. distribution networks [12] or healthcare systems [13]. Therefore, models that allow defining more than two states in system/components performance are used in the analysis of such systems. These models are known as Multi-State Systems (MSSs).

A general MSS permits defining different number of states for the system and for its components. If we assume that the system has $m$ possible states and its $i$-th component, for $i = 1,\ldots, n$, can be in one of $m_i$ states, then the structure function of the MSS corresponds to the next map [14]–[16]:

$$\begin{gathered}\phi(x_1,...,x_n) = \phi(\boldsymbol{x}) : \\ \{0,...,m_1 - 1\} \times \ldots \times \{0,...,m_n - 1\} \to \{0,...,m-1\},\end{gathered} \tag{2}$$

where 0 corresponds to completely failure of the system/ component while $m$ -1 ($m_i$ -1) means that the system (the $i$-th component) is perfectly functioning.

A special type of MSSs is a homogenous system, in which $m_1 = \ldots = m_n = m$. The structure function of such system can be interpreted as a Multiple-Valued Logic (MVL) function. This fact allows us to use some methods of MVL logic in reliability analysis of MSSs [16].

The mathematical model of structure function used in reliability analysis can be combined with methods of DM to perform analysis of medical databases. This analysis can be used to identify database attributes that carry no important information from the point of view of database purpose.

For example, let us consider the example of database for the analysis of the breast cancer diagnosis from [17]. In this example 4 categorical input attributes are used (Table I): $A_1$ (*Gynecological history*), $A_2$ (*Tumor*), $A_3$ (*Heredity*), and $A_4$ (*Age*). Each combination of values of the input attributes is connected to output attribute B (*Breast Cancer Possibility*). Let the attributes have the next values: $A_1 = \{A_{1,1}, A_{1,2}, A_{1,3}\}$, $A_2 = \{A_{2,1}, A_{2,2}, A_{2,3}\}$, $A_3 = \{A_{3,1}, A_{3,2}\}$, $A_4 = \{A_{4,1}, A_{4,2}\}$, and $B = \{B_1, B_2, B_3\}$ assuming that these values has the meaning presented in Table I.

TABLE I.
ATTRIBUTES VALUES

| Attribute | Attribute Values | Description of Attribute Values |
|---|---|---|
| $A_1$ | $A_{1,1}$ | Gynecological history with high risk |
| | $A_{1,2}$ | Gynecological history with medium risk |
| | $A_{1,3}$ | Gynecological history with low risk |
| $A_2$ | $A_{2,1}$ | Yes and confirmed by medical examination |
| | $A_{2,2}$ | Yes and non-confirmed |
| | $A_{2,3}$ | No |
| $A_3$ | $A_{3,1}$ | Yes |
| | $A_{3,2}$ | No |
| $A_4$ | $A_{4,1}$ | Younger than 40 years |
| | $A_{4,2}$ | 40 years or more |
| B | $B_1$ | High Possibility of Breast Cancer |
| | $B_2$ | Medium Possibility of Breast Cancer |
| | $B_3$ | Low Possibility of Breast Cancer |

In [17], an association rule for breast cancer diagnosis was inducted based on the methods of DM. This rule includes 4 input attributes and one output attribute (Table II). From reliability point of view, the input attributes can be interpreted as system components and the output attribute is considered to be the system state. This implies that the table representing the discovered knowledge can be interpreted as the structure function of a MSS that depends on 4 variables: the 1-st and 2-nd variable has 3 possible values and the 3-rd and 4-th have two values. This structure function is defined based on all possible values of the input attributes and the output attribute is calculated according to the association rule derived in [17].

In section IV, we will use some approaches of reliability analysis to identify input attributes that have the greatest influence on the value of the output attribute. These results can be used to perform some database optimization since we can find attributes that have very little influence on the output attribute and, therefore, it might not be necessary to store them in the database.

### B. Coherent and Noncoherent Systems

Based on the properties of the structure function, two different classes of systems can be recognized – coherent and noncoherent. A system is coherent if a failure/degradation of any system component can result only in system failure/ degradation. This means that the structure function of coherent systems is monotonic (non-decreasing) [8], [14].

TABLE II.
STRUCTURE FUNCTION FOR DATABASE OF BREAST CANCER DIAGNOSIS

| Variables<br>$x_1$ $x_2$ $x_3$ $x_4$ | $\phi(\boldsymbol{x})$ | Variables<br>$x_1$ $x_2$ $x_3$ $x_4$ | $\phi(\boldsymbol{x})$ |
|---|---|---|---|
| 0 0 0 0 | 1 | 1 1 1 1 | 2 |
| 0 0 0 1 | 1 | 1 2 0 0 | 0 |
| 0 0 1 0 | 1 | 1 2 0 0 | 2 |
| 0 0 1 1 | 1 | 1 2 0 1 | 2 |
| 0 1 0 0 | 2 | 1 2 1 0 | 2 |
| 0 1 0 1 | 0 | 1 2 1 1 | 2 |
| 0 1 1 0 | 2 | 2 0 0 0 | 2 |
| 0 1 1 1 | 0 | 2 0 0 1 | 2 |
| 0 2 0 0 | 2 | 2 0 1 0 | 2 |
| 0 2 0 1 | 2 | 2 0 1 1 | 2 |
| 0 2 1 0 | 2 | 2 1 0 0 | 2 |
| 0 2 1 1 | 2 | 2 1 0 1 | 2 |
| 1 0 0 0 | 1 | 2 1 1 1 | 2 |
| 1 0 0 1 | 1 | 2 2 0 0 | 2 |
| 1 0 1 0 | 1 | 2 2 0 0 | 2 |
| 1 0 1 1 | 1 | 2 2 0 1 | 2 |
| 1 1 0 0 | 2 | 2 2 1 0 | 2 |
| 1 1 0 1 | 0 | 2 2 1 1 | 2 |

A system is noncoherent if its structure function is not monotonic. This implies that a noncoherent system admits situations in which component failure/degradation can cause system repair/improvement [18].

The coherency is a typical property of most systems studied in reliability engineering. This indicates that many tools of reliability analysis are based on the assumption that the structure function is monotonic. However, there also exist some systems whose structure function cannot be monotone, e.g. logic networks [19] or *k*-to-*l*-out-of-*n* systems, which are functioning if at least *k* but not more than *l* components are working [20]. The analysis of such systems is more complicated than the analysis of coherent systems and, therefore, it has to be done more carefully.

### C. Availability

The structure function defines system topology, and it carries no information about reliability of individual system components. Therefore, if we want to analyze not only topological properties of the system but also some probabilistic characteristics (e.g. system availability or unavailability, mean time to failure), the probabilities of states of individual system components have to be known:

$$p_{i,s} = \Pr\{x_i = s\}, \quad s \in \{0, \ldots, m_i - 1\}, \quad i \in \{1, \ldots, n\}, \qquad (3)$$

where $m_i = 2$ for BSSs. Please note that in case of BSSs, $p_{i,0}$ is known as unavailability of the *i*-th system component and $p_{i,1}$ as its availability.

Knowledge of system structure function and probabilities (3) can be used to compute three important characteristics of the system – the probability that the system is in state *j* (for *j* = 0,…, *m* -1), the system availability, which is defined with regard to system state *j* as follows [14], [15]:

$$A^{\geq j} = \Pr\{\phi(\boldsymbol{x}) \geq j\}, \quad j \in \{1,\ldots,m-1\}, \tag{4}$$

and the system unavailability, which is defined with regard to system state $j$ as the probability that the system cannot fulfill a requirement that requires at least level $j$ of system performance [14], [15]:

$$U^{\geq j} = \Pr\{\phi(\boldsymbol{x}) < j\} = 1 - A^{\geq j}, \quad j \in \{1,\ldots,m-1\}. \tag{5}$$

Please note, in case of BSSs ($m = 2$ in definitions (4) and (5)), the system availability is defined as the probability that the system is in state 1 while the unavailability agrees with the probability of system 0-state. Therefore, terms “system availability (unavailability)” and “the probability that the system is in state 1 (0)” can be used as synonyms in case of BSSs. However, this is not true for MSSs, in which these two terms represent two different concepts.

System availability and unavailability are very important in reliability analysis. They can be used to estimate mean time to system failure or mean time to system repair. However, they do not allow identifying components that have the greatest influence on system activity. This is very important task because its results can be used to optimize system reliability or to plan system maintenance.

### D. *Importance Analysis*

Importance analysis is a part of reliability engineering. It is used to quantify situations in which a change of component state results in a change of system state. For this purpose, Importance Measures (IMs) are used. There exist a lot of IMs [21]. However, in what follows, we will consider only two of them – the Structural Importance (SI) and the Birnbaum’s Importance (BI).

The SI and BI have originally been developed for the analysis of coherent systems. In [22], the SI of component $i$ has been defined as a relative number of situations in which the component is critical for system failure/functioning, i.e. as a proportion of cases when a failure (repair) of the $i$-th system component results in system failure (repair). In the same paper, the BI has been introduced as the probability that the component failure (repair) causes system failure (repair). These definitions imply that the main difference between the SI and BI is that the former analyzes only the system structure while the latter takes into account not only the structure function but also availabilities (unavailabilities) of system components. Therefore, the SI is primarily used to analyze topological properties of the system.

The considered IMs have been generalized for coherent MSSs in [16], [23]–[25]. These works have introduced several types of the SI and BI depending on whether we are interested in:

a) identification of component states that are the most important for a given system state/ availability [16], [23], [25],
b) identifying component states that have the most influence on the whole system (not only on a specific system state/availability) [24],
c) finding components that are the most important for a given system state/availability,
d) revealing the total importance of individual components for the whole system [25].

Finally, works [18], [26] have introduced definitions of the SI and BI for noncoherent BSSs. These versions of the SI and BI allow quantifying:

a) dependency of system failure (repair) on a failure (repair) of a given component,
b) dependency of system repair (failure) on a failure (repair) of a given component,
c) the total influence of a given component on the system activity.

According to our knowledge, no generalizations of the considered IMs have been proposed for importance analysis of noncoherent MSSs. This can be caused by the fact that these models are used in reliability analysis very rarely.

### E. *Logical Differential Calculus*

Logical differential calculus is a tool used to investigate dynamic properties of Boolean and MVL functions [27]. The central term of this tool is a Boolean/MVL derivative. There exist several types of this derivative. For our purposes, the most important is Direct Partial Logic Derivative (DPLD).

A DPLD of a Boolean function $f(\boldsymbol{x})$ with respect to variable $x_i$ can be defined in the following way [27]:

$$\partial f(j \to \bar{j}) / \partial x_i(s \to \bar{s}) = \\ = \begin{cases} 1, \text{ if } f(s_i, \boldsymbol{x}) = j \text{ and } f(\bar{s}_i, \boldsymbol{x}) = \bar{j} \\ 0, \text{ other} \end{cases}, \tag{6}$$

where $f(a_i, \boldsymbol{x}) = f(x_1,\ldots, x_{i-1}, a, x_{i+1},\ldots, x_n)$ for $a \in \{s, \bar{s}\}$ and $s, j \in \{0, 1\}$. According to this definition, the DPLD of a Boolean function reveals situations in which change of Boolean variable $x_i$ from value $s$ to $\bar{s}$ causes that the Boolean function value changes from $j$ to $\bar{j}$.

In the similar way, a DPLD of a MVL function $f_m(\boldsymbol{x})$ with respect to variable $x_i$ is defined as follows [27]:

$$\partial f_m(j \to h) / \partial x_i(s \to r) = \\ = \begin{cases} 1, \text{ if } f_m(s_i, \boldsymbol{x}) = j \text{ and } f_m(r_i, \boldsymbol{x}) = h \\ 0, \text{ other} \end{cases}, \tag{7}$$

where $f_m(a_i, \boldsymbol{x}) = f_m(x_1,\ldots, x_{i-1}, a, x_{i+1},\ldots, x_n)$ for $a \in \{s, r\}$; $s, r, j, h \in \{0,\ldots, m\text{ -}1\}$, $s \neq r$, and $j \neq h$. Clearly, this derivative models consequence of change of the MVL variable from value $s$ to $r$ on the value of the considered MVL function and, therefore, it can be used to detect situations in which the investigated change of the MVL variable results in the change of the function value from $j$ to $h$.

Since the structure function of a BSS can be interpreted as a Boolean function and the structure function of a homogenous MSS as a MVL function, DPLDs can also be used in reliability analysis of such systems [9], [16]. Moreover, the next little modification of definition (6) also allows applying them to non-homogenous MSSs:

$$\partial\phi(j \to h)/\partial x_i(s \to r) = \\ = \begin{cases} 1, \text{ if } \phi(s_i, \boldsymbol{x}) = j \text{ and } \phi(r_i, \boldsymbol{x}) = h \\ 0, \text{ other} \end{cases}, \tag{8}$$

where $s,\ r \in \{0,\ldots, m_i\text{-}1\}$, $s \neq r$ and $j,\ h \in \{0,\ldots, m\text{-}1\}$, $j \neq h$. Please note that this definition is the most general definition of a DPLD from which definitions (6) and (7) can be obtained simply using the assumption that $s,\ r,\ j,\ h \in \{0,1\}$ or $s, r, j, h \in \{0,\ldots, m\text{-}1\}$ respectively. Therefore, in what follows, we will primarily use this definition.

In terms of reliability analysis, DPLDs are used to identify situations in which a given change of state of the *i*-th system component results in the investigated change of the system state. These derivatives can be split into four groups:

A. $j > h$ and $s > r$ – these derivatives identify situations in which component failure/degradation results in system failure/degradation,

B. $j < h$ and $s < r$ – these DPLDs detect situations in which component repair/improvement causes a repair/improvement of system activity,

C. $j > h$ and $s < r$ – these derivatives can be used to find coincidence between component repair/improvement and system failure/degradation,

D. $j < h$ and $s > r$ – these DPLDs investigate situations in which system repair/improvement is caused by failure/degradation of the considered component.

Based on the definition of a coherent system, only DPLDs from groups A and B are relevant in the analysis of such systems. However, this is not true for noncoherent systems, for which the derivatives from all groups can be nonzero.

### *F. Importance Measures based on Direct Partial Logic Derivatives*

The SI and BI are used to quantify coincidence between component state change and change of system state. Based on the previous paragraphs, this coincidence can be identified based on DPLDs. Therefore, these derivatives can also be used to compute the SI and BI [9], [13], [16].

Firstly, let us consider a coherent BSS. The structure function of this system is monotone, therefore, only DPLDs $\partial\phi(1 \to 0)/\partial x_i(1 \to 0)$ and $\partial\phi(0 \to 1)/\partial x_i(0 \to 1)$ can be nonzero for this type of systems. The former identifies situations in which a failure of component *i* results in system failure, and the latter detects state vectors at which a repair of the component causes that the system begins work. Since the SI of component *i* is defined as a relative number of situations in which a failure (repair) of component *i* results in system failure (repair), this IM can be computed using DPLDs in the following manner [9]:

$$\begin{aligned} \mathrm{SI}_i &= \mathrm{TD}\big(\partial\phi(1 \to 0)/\partial x_i(1 \to 0)\big) \\ &= \mathrm{TD}\big(\partial\phi(0 \to 1)/\partial x_i(0 \to 1)\big), \end{aligned} \tag{9}$$

where TD(.) denotes truth density of the argument interpreted as a function with Boolean-valued output, i.e. a proportion of situations in which the argument takes value 1.

The BI of component *i* can be calculated using DPLDs in the similar way [9]:

$$\begin{aligned} \mathrm{BI}_i &= \Pr\{\partial\phi(1 \to 0)/\partial x_i(1 \to 0) = 1\} \\ &= \Pr\{\partial\phi(0 \to 1)/\partial x_i(0 \to 1) = 1\} \end{aligned} \tag{10}$$

since it is defined as the probability that a failure (repair) of the component causes system failure (repair).

Secondly, let us consider a noncoherent BSS. In this case all four DPLDs that can be defined are relevant because all of them can contain nonzero elements, i.e. DPLDs $\partial\phi(1 \to 0)/\partial x_i(1 \to 0)$, $\partial\phi(1 \to 0)/\partial x_i(0 \to 1)$ can be used to find correlation between system failure and change of state of component *i* while derivatives $\partial\phi(0 \to 1)/\partial x_i(0 \to 1)$, $\partial\phi(0 \to 1)/\partial x_i(1 \to 0)$ identify situations in which a change of component state results in system repair.

It has been proposed in [18] that importance analysis of noncoherent BSSs should be performed in three steps. Firstly, we should quantify influence of component failure on system failure (repair). Secondly, impact of component repair on system failure (repair) should be quantified. Finally, the total influence of the considered component on system failure (repair) can be estimated as the sum of the results obtained in the previous two steps. This implies that several IMs of one type can be defined for one component. For example, in case of the SI, the next four measures can be calculated:

$$\begin{aligned} \mathrm{SI}_{i\downarrow}^{\downarrow} &= \mathrm{TD}\big(\partial\phi(1 \to 0)/\partial x_i(1 \to 0)\big), \\ \mathrm{SI}_{i\uparrow}^{\uparrow} &= \mathrm{TD}\big(\partial\phi(0 \to 1)/\partial x_i(0 \to 1)\big), \\ \mathrm{SI}_{i\downarrow}^{\uparrow} &= \mathrm{TD}\big(\partial\phi(0 \to 1)/\partial x_i(1 \to 0)\big), \\ \mathrm{SI}_{i\uparrow}^{\downarrow} &= \mathrm{TD}\big(\partial\phi(1 \to 0)/\partial x_i(0 \to 1)\big). \end{aligned} \tag{11}$$

The first two SI measures are used to quantify coincidence between component failure (repair) and system failure (repair). The remaining SI measures estimates topological correlation between component failure (repair) and repair (failure) of the system. The total topological influence of component *i* on system failure is computed in the following manner [26]:

$$\begin{aligned} \mathrm{SI}_i^{\downarrow} &= \mathrm{SI}_{i\downarrow}^{\downarrow} + \mathrm{SI}_{i\uparrow}^{\downarrow} \\ &= \mathrm{TD}\big(\partial\phi(1 \to 0)/\partial x_i(1 \to 0)\big) \\ &\quad + \mathrm{TD}\big(\partial\phi(1 \to 0)/\partial x_i(0 \to 1)\big), \end{aligned} \tag{12}$$

and on system repair in the following way:

$$\begin{aligned} \mathrm{SI}_i^{\uparrow} &= \mathrm{SI}_{i\uparrow}^{\uparrow} + \mathrm{SI}_{i\downarrow}^{\uparrow} \\ &= \mathrm{TD}\big(\partial\phi(0 \to 1)/\partial x_i(0 \to 1)\big) \\ &\quad + \mathrm{TD}\big(\partial\phi(0 \to 1)/\partial x_i(1 \to 0)\big). \end{aligned} \tag{13}$$

It is clear that

$$\mathrm{SI}_{i\downarrow}^{\downarrow} = \mathrm{SI}_{i\uparrow}^{\uparrow},\ \mathrm{SI}_{i\uparrow}^{\downarrow} = \mathrm{SI}_{i\downarrow}^{\uparrow},\ \mathrm{SI}_{i}^{\downarrow} = \mathrm{SI}_{i}^{\uparrow} \qquad (14)$$

and, therefore, $\mathrm{SI}_{i\downarrow}^{\downarrow}$, $\mathrm{SI}_{i\uparrow}^{\downarrow}$, and $\mathrm{SI}_{i}^{\downarrow}$ can be used not only in the investigation of system failure but also in the analysis of system repair.

Please note that the same results can also be obtained for the BI of noncoherent BSSs by replacement of the truth densities in equations (11) – (13) with the probabilities that the DPLDs take value 1.

Thirdly, let us focus on coherent MSSs. In this case, we can quantify several dependencies between component state and system state. For simplicity, we will consider only situations in which component degradation coincide with system degradation, i.e. we will not introduce the IMs for component improvement. Furthermore, we will assume that system components can degrade only one state. Using these assumptions, we can calculate the SI of state *s* of component *i* for system state *j* as follows [16]:

$$\mathrm{SI}_{i,s\downarrow}^{j\downarrow} = \sum_{h=0}^{j-1} \mathrm{TD}\big(\partial\phi(j \to h)/\partial x_i(s \to s-1)\big), \qquad (15)$$

for $i \in \{1,\dots,n\}$, $s \in \{1,\dots,m_i\text{-}1\}$, $j \in \{1,\dots,m\text{-}1\}$.

Based on the meaning of DPLDs, this IM corresponds to the relative number of situations in which a minor degradation (i.e. degradation by one state) of state *s* of component *i* results in degradation of system state *j*. Using the ideas presented in [24], [25], this SI can also be used to compute the relative number of situations in which a minor degradation of state *s* of component *i* results in system degradation:

$$\begin{aligned}\mathrm{SI}_{i,s\downarrow}^{\downarrow} &= \sum_{j=1}^{m-1} \mathrm{SI}_{i,s\downarrow}^{j\downarrow} \\ &= \sum_{j=1}^{m-1}\sum_{h=0}^{j-1} \mathrm{TD}\big(\partial\phi(j \to h)/\partial x_i(s \to s-1)\big),\end{aligned} \qquad (16)$$

or the relative number of situations in which a minor degradation of component *i* causes degradation of system state *j*:

$$\begin{aligned}\mathrm{SI}_{i\downarrow}^{j\downarrow} &= \frac{1}{m_i-1}\sum_{s=1}^{m_i-1} \mathrm{SI}_{i,s\downarrow}^{j\downarrow} \\ &= \frac{1}{m_i-1}\sum_{s=1}^{m_i-1}\sum_{h=0}^{j-1} \mathrm{TD}\big(\partial\phi(j \to h)/\partial x_i(s \to s-1)\big),\end{aligned} \qquad (17)$$

or the proportion of state vectors at which a minor degradation of component *i* results in system degradation:

$$\mathrm{SI}_{i\downarrow}^{\downarrow} = \frac{1}{m_i-1}\sum_{s=1}^{m_i-1} \mathrm{SI}_{i,s\downarrow}^{\downarrow}. \qquad (18)$$

Equation (16) can be used to quantify topological influence of state *s* of component *i* on the whole system. Similarly, equations (17) allows us to investigate the total influence of component *i* on system state *j* and formula (18) the total influence on the whole system. In the similar way, the BI for a minor degradation of component state for a coherent MSS can be defined. The only difference is that the truth densities in (15) – (17) have to be replaced with the probabilities that the considered DPLDs are nonzero.

## IV. Use of Importance Analysis in Investigation of Medical Databases

As we mentioned in section III.A, a complete medical database can be obtained from a medical dataset using some tools of DM. The complete database containing only qualitative (categorical) attributes can be viewed as a MSS whose structure function agrees with the relation (discovered knowledge) between the input attributes and the output attribute. However, the main problem is that the database has to be interpreted as the structure function of a noncoherent MSS because it can contain situations in which a decrease in value of input attribute can result in increase of value of the output attribute. For example, in Table II, change of variable $x_2$, which corresponds to value of attribute $A_2$, from value 1 to 0 causes that the value of the structure function of the considered database changes from value 0 to 1 if $x_1 = 0$, $x_3 = 0$, and $x_4 = 1$. This fact requires proposing some generalizations of the SI for noncoherent MSSs if we want to use this measure to find input attributes that have the greatest influence on the output attribute.

In noncoherent systems, not only component degradation but also component improvement can result in system degradation. This implies that we also need to detect situations in which component improvement results in system degradation. DPLDs $\partial\phi(j \to h)/\partial x_i(s \to s+1)$ in which $j > h$ can be used for this purpose. Based on these DPLDs, topological influence of a minor improvement of state *s* of component *i* on degradation of system state *j* can be estimated using the next version of SI:

$$\mathrm{SI}_{i,s\uparrow}^{j\downarrow} = \sum_{h=0}^{j-1} \mathrm{TD}\big(\partial\phi(j \to h)/\partial x_i(s \to s+1)\big), \qquad (19)$$

for $i \in \{1,\dots,n\}$, $s \in \{0,\dots,m_i\text{-}2\}$, $j \in \{1,\dots,m\text{-}1\}$.

In case of noncoherent BSSs, the total influence of a given component on system failure is computed as the sum of SI measures analyzing consequences of the component failure and repair. Therefore, in case of MSSs, the total importance of state *s* of component *i* for degradation of system state *j* can be computed simply using the next SI:

$$\mathrm{SI}_{i,s}^{j\downarrow} = \begin{cases} \mathrm{SI}_{i,s\uparrow}^{j\downarrow} & \text{if } s = 0, \\ \mathrm{SI}_{i,s\downarrow}^{j\downarrow} & \text{if } s = m_i - 1, \\ \mathrm{SI}_{i,s\downarrow}^{j\downarrow} + \mathrm{SI}_{i,s\uparrow}^{j\downarrow} & \text{else,} \end{cases} \qquad (20)$$

where $\mathrm{SI}_{i,s\downarrow}^{j\downarrow}$ is computed based on formula (15).

SI measures (15), (19), and (20) are useful for evaluation of influence of a given component state on degradation of a given system state. However, they do not allow identifying importance of the whole component on degradation of a given system state or importance of a given component state on the entire system (regardless of a concrete system state). For these purposes, other versions of the SI have to be defined. This can be done in the similar way as in the case of coherent MSSs, i.e. the total importance of a given component state on system activity can be computed as follows:

$$\mathrm{SI}_{i,s}^{\downarrow}=\sum_{j=1}^{m-1}\mathrm{SI}_{i,s}^{j\downarrow}=\begin{cases}\mathrm{SI}_{i,s\uparrow}^{\downarrow} & \text{if } s=0,\\ \mathrm{SI}_{i,s\downarrow}^{\downarrow} & \text{if } s=m_i-1,\\ \mathrm{SI}_{i,s\downarrow}^{\downarrow}+\mathrm{SI}_{i,s\uparrow}^{\downarrow} & \text{else,}\end{cases}\qquad(21)$$

where $\mathrm{SI}_{i,s\downarrow}^{\downarrow}$ (definition (16)) quantifies consequences of deterioration of state $s$ of the $i$-th system component on system activity, and $\mathrm{SI}_{i,s\uparrow}^{\downarrow}$ calculates results of improvement of state $s$ of component $i$ on system degradation. Please note that $\mathrm{SI}_{i,s\uparrow}^{\downarrow}$ is computed similarly as $\mathrm{SI}_{i,s\downarrow}^{\downarrow}$, i.e.:

$$\begin{aligned}\mathrm{SI}_{i,s\uparrow}^{\downarrow}&=\sum_{j=1}^{m-1}\mathrm{SI}_{i,s\uparrow}^{j\downarrow}\\&=\sum_{j=1}^{m-1}\sum_{h=0}^{j-1}\mathrm{TD}\big(\partial\phi(j\to h)/\partial x_i(s\to s+1)\big).\end{aligned}\qquad(22)$$

The total topological influence of component $i$ on a given system state can be calculated based on the next formula:

$$\mathrm{SI}_i^{j\downarrow}=\frac{1}{m_i-1}\sum_{s=0}^{m_i-1}\mathrm{SI}_{i,s}^{j\downarrow}=\mathrm{SI}_{i\downarrow}^{j\downarrow}+\mathrm{SI}_{i\uparrow}^{j\downarrow},\qquad(23)$$

where $\mathrm{SI}_{i\downarrow}^{j\downarrow}$ (definition (17)) quantifies results of degradation of the $i$-th system component on degradation of system state $j$, and $\mathrm{SI}_{i\uparrow}^{j\downarrow}$ evaluating consequences of improvement of the $i$-th component on degradation of system state $j$ is computed using the following formula:

$$\begin{aligned}\mathrm{SI}_{i\uparrow}^{j\downarrow}&=\frac{1}{m_i-1}\sum_{s=0}^{m_i-2}\mathrm{SI}_{i,s\uparrow}^{j\downarrow}\\&=\frac{1}{m_i-1}\sum_{s=0}^{m_i-2}\sum_{h=0}^{j-1}\mathrm{TD}\big(\partial\phi(j\to h)/\partial x_i(s\to s+1)\big).\end{aligned}\qquad(24)$$

Based on the meaning of DPLDs, it can be shown simply that SI (21) agrees with the relative number of state vectors at which a minor change (i.e. a change by one state) of state $s$ of component $i$ results in system degradation, SI (22) identifies the relative count of state vectors at which a minor improvement of state $s$ of component $i$ results in system degradation, SI (23) corresponds to the relative number of situations in which a degradation or improvement of component $i$ causes decrease in state $j$ of the system, and SI (24) agrees with the proportion of state vectors at which an improvement of component $i$ causes deterioration of state $j$ of the system

Finally, the total topological importance of a given component on system activity can be defined as the relative number of state vectors at which a change of component state results in system deterioration and, therefore, it can be computed as follows:

$$\mathrm{SI}_i^{\downarrow}=\mathrm{SI}_{i\downarrow}^{\downarrow}+\mathrm{SI}_{i\uparrow}^{\downarrow},\qquad(25)$$

where $\mathrm{SI}_{i\downarrow}^{\downarrow}$ (definition (18)) quantifies results of degradation of the $i$-th system component on system degradation, and $\mathrm{SI}_{i\uparrow}^{\downarrow}$ computed based on the next formula:

$$\mathrm{SI}_{i\uparrow}^{\downarrow}=\frac{1}{m_i-1}\sum_{s=0}^{m_i-2}\mathrm{SI}_{i,s\uparrow}^{\downarrow}\qquad(26)$$

evaluates consequences of improvement of the considered component on system degradation.

In this section, we have proposed a lot of SI measures that can be used in the investigation of topological properties of noncoherent MSSs. For clarity, we summarize them in Table III. The similar formulae could also be proposed for BI measures. The only difference is that the probabilities that the DPLDs are nonzero have to be used in formulae (15) – (26) instead of the truth densities.

TABLE III.
STRUCTURAL IMPORTANCE MEASURES FOR NONCOHERENT MULTI-STATE SYSTEMS

| SI | Coherent Part (Influence of Component Degradation) | Noncoherent Part (Influence of Component Improvement) |
|---|---|---|
| $\mathrm{SI}_{i,s}^{j\downarrow}$ | $\mathrm{SI}_{i,s\downarrow}^{j\downarrow}=\sum_{h=0}^{j-1}\mathrm{TD}\left(\frac{\partial\phi(j\to h)}{\partial x_i(s\to s-1)}\right)$ | $\mathrm{SI}_{i,s\uparrow}^{j\downarrow}=\sum_{h=0}^{j-1}\mathrm{TD}\left(\frac{\partial\phi(j\to h)}{\partial x_i(s\to s+1)}\right)$ |
| $\mathrm{SI}_{i,s}^{\downarrow}$ | $\mathrm{SI}_{i,s\downarrow}^{\downarrow}=\sum_{j=1}^{m-1}\mathrm{SI}_{i,s\downarrow}^{j\downarrow}$ for $s>0$ | $\mathrm{SI}_{i,s\uparrow}^{\downarrow}=\sum_{j=1}^{m-1}\mathrm{SI}_{i,s\uparrow}^{j\downarrow}$ for $s<m_i-1$ |
| $\mathrm{SI}_i^{j\downarrow}$ | $\mathrm{SI}_{i\downarrow}^{j\downarrow}=\frac{1}{m_i-1}\sum_{s=1}^{m_i-1}\mathrm{SI}_{i,s\downarrow}^{j\downarrow}$ | $\mathrm{SI}_{i\uparrow}^{j\downarrow}=\frac{1}{m_i-1}\sum_{s=0}^{m_i-2}\mathrm{SI}_{i,s\uparrow}^{j\downarrow}$ |
| $\mathrm{SI}_i^{\downarrow}$ | $\mathrm{SI}_{i\downarrow}^{\downarrow}=\frac{1}{m_i-1}\sum_{s=1}^{m_i-1}\mathrm{SI}_{i,s\downarrow}^{\downarrow}$ | $\mathrm{SI}_{i\uparrow}^{\downarrow}=\frac{1}{m_i-1}\sum_{s=0}^{m_i-2}\mathrm{SI}_{i,s\uparrow}^{\downarrow}$ |

Based on the relation between medical database and the structure function of a noncoherent MSS, the proposed SI measures can be used to analyze importance of individual input attributes on the value of the output attribute. For illustration, let us consider the medical database defined by Table II. The total topological importance of individual input attributes is computed based on formula (25) in Table IV. Based on the data presented in this table, the input attribute that has the greatest influence on the output attribute is $\mathrm{A}_2$. On the other hand, value of attribute $\mathrm{A}_3$ has no influence on attribute B. This implies that attribute $\mathrm{A}_3$ is not important for tasks for which the table is used (i.e. decision whether the breast cancer has high possibility or not) and, therefore, it is not necessary to store its values in the database.

## V. CONCLUSION

This paper focuses on correlation between some key terms of KDD (or DM) and reliability analysis. We illustrated that KDD is a very complex process whose main part is DM. DM is used to discover some new information (knowledge) in a database (predictive DM) or for better understanding of data stored in a database (descriptive DM). In case of predictive DM used on databases containing only categorical attributes, the discovered knowledge can be interpreted as a table that can be viewed as the structure function of a MSS. This allows us to use some methods of reliability analysis in investigation of database properties. One of them is importance analysis, which identifies influence of system components on the system activity.

TABLE IV.
IMPORTANCE OF INDIVIDUAL INPUT ATTRIBUTES FOR THE OUTPUT ATTRIBUTE

| Input Attribute | $\mathrm{SI}_i^{\downarrow}$ |
|---|---|
| A1 | 0.25 |
| A2 | 0.50 |
| A3 | 0 |
| A4 | 0.22 |

In this paper, we considered use of importance analysis in investigation of coincidence between change of input attributes and change of the output attribute of a table representing the discovered knowledge. This required extending some measures used in importance analysis on noncoherent MSSs. The extension was done using logical differential calculus. The presented approach can be used to optimize number of attributes occurring in the table. Furthermore, it can also be used to decide which attributes have to be measured with the most accuracy to ensure that the prediction based on the table representing the discovered knowledge will be correct.

## REFERENCES

[1] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, "Knowledge discovery in databases: An overview," in *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, Eds. Cambridge, MA: AAAI/MIT Press, 1–27, 1991.

[2] V. Sigillito, "Pima Indians Diabetes Database," UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Pima Indians Diabetes]. Phoenix, AZ: National Institute of Diabetes and Digestive and Kidney Diseases, 1990.

[3] K. J. Cios and G. W. Moore, "Medical data mining and knowledge discovery: Overview of key issues," in *Medical Data Mining and Knowledge Discovery*, K. J. Cios, Ed. New York, NY: Physica Verlag Heidelberg, 2001, pp. 1–20.

[4] K. J. Cios, A. Teresinska, S. Konieczna, J. Potocka and S. Sharma, "A knowledge discovery approach to diagnosing myocardial perfusion," *IEEE Engineering in Medicine and Biology Magazine*, vol. 19, no. 4, pp. 17–25, Jul.–Aug. 2000.

[5] A. Petrie and C. Sabin, *Medical Statistics at a Glance*, 2nd ed. Oxford, UK: Blackwell Publishing Ltd, 2005.

[6] L. J. Cao, K. S. Chua, W. K. Chong, H. P. Lee and Q. M. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, no. 1–2, pp. 321–336, Sep. 2003.

[7] O. Maimon and L. Rokach, "Introduction to knowledge discovery in databases," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. New York, NY: Springer Science+Business Media, Inc., 2005, pp. 1–17.

[8] M. Rausand and A. Høyland, *System Reliability Theory: Models, Statistical Methods, and Applications*. Haboken, NJ: John Wiley & Sons, Inc., 2004, 664 p.

[9] E. N. Zaitseva and V. G. Levashenko, "Importance analysis by logical differential calculus," *Automation and Remote Control*, vol. 74, no. 2, pp. 171–182, Feb. 2013, http://dx.doi.org/10.1134/S000511791302001X.

[10] Y. Watanabe, T. Oikawa, and K. Muramatsu, "Development of the DQFM method to consider the effect of correlation of component failures in seismic PSA of nuclear power plant," *Reliability Engineering & System Safety*, vol. 79, no. 3, pp. 265–279, Mar. 2003, http://dx.doi.org/10.1016/S0951-8320(02)00053-4.

[11] B. Nystrom, L. Austrin, N. Ankarback, and E. Nilsson, "Fault tree analysis of an aircraft electric power supply system to electrical actuators," in *Probabilistic Methods Applied to Power Systems, 2006. PMAPS 2006. International Conference on*, 2006, pp. 1–7, http://dx.doi.org/10.1109/PMAPS.2006.360325.

[12] W.-C. Yeh, "A simple approach to search for all d-MCs of a limited-flow network," *Reliability Engineering & System Safety*, vol. 71, no. 1, pp. 15–19, Jan. 2001, http://dx.doi.org/10.1016/S0951-8320(00)00070-3.

[13] E. Zaitseva, V. Levashenko, and M. Rusin, "Reliability analysis of healthcare system," in *2011 Federated Conference on Computer Science and Information Systems, FedCSIS 2011*, 2011, pp. 169–175.

[14] B. Natvig, *Multistate Systems Reliability Theory with Applications*. New York, NY: Wiley, 2011, 262 p., http://dx.doi.org/10.1002/9780470977088.

[15] A. Lisnianski, I. Frenkel and Y. Ding, *Multi-state System Reliability Analysis and Optimization for Engineers and Industrial Managers*. London, UK: Springer-Verlag London Ltd., 2010, 393 p., http://dx.doi.org/10.1007/978-1-84996-320-6.

[16] E. Zaitseva and V. Levashenko, "Multiple-valued logic mathematical approaches for multi-state system reliability analysis," *Journal of Applied Logic*, vol. 11, no. 3, pp. 350–362, Sep. 2013, http://dx.doi.org/10.1016/j.jal.2013.05.005.

[17] V. Levashenko and E. Zaitseva, "Fuzzy decision trees in medical decision making support system" in *2012 Federated Conference on Computer Science and Information Systems, FedCSIS 2012*, 2012, pp. 213–219.

[18] J. D. Andrews and S. Beeson, "Birnbaum's measure of component importance for noncoherent systems," *IEEE Transactions on Reliability*, vol. 52, no. 2, pp. 213–219, Jun. 2003, http://dx.doi.org/10.1109/TR.2003.809656.

[19] E. Zaitseva, M. Kvassay, V. Levashenko, and J. Kostolny, "Reliability analysis of logic network by logical differential calculus," in *2014 ELEKTRO*, 2014, pp. 245–250, http://dx.doi.org/10.1109/ELEKTRO.2014.6848895.

[20] S. J. Upadhyaya and H. Pham, "Analysis of noncoherent systems and an architecture for the computation of the system reliability," *IEEE Transactions on Computers*, vol. 42, no. 4, pp. 484–493, Apr. 1993, http://dx.doi.org/10.1109/12.214699.

[21] W. Kuo and X. Zhu, *Importance Measures in Reliability, Risk, and Optimization*. Chichester, UK: John Wiley & Sons, Ltd, 2012, 472 p., http://dx.doi.org/10.1002/9781118314593.

[22] Z. W. Birnbaum, "On the importance of different components in a multicomponent system," in *Multivariate Analysis*, vol. 2, P. R. Krishnaiah, Ed. New York, NY: Academic Press, 1969, pp. 581–592.

[23] D. A. Butler, "A complete importance ranking for components of binary coherent systems, with extensions to multi-state systems," *Naval Research Logistics Quarterly*, vol. 26, no. 4, pp. 565–578, Dec. 1979, http://dx.doi.org/10.1002/nav.3800260402.

[24] W. S. Griffith, "Multistate reliability models," *Journal of Applied Probability*, vol. 17, no. 3, pp. 735–744, Sep. 1980, http://dx.doi.org/10.2307/3212967.

[25] S. Wu, "Joint importance of multistate systems," *Computers & Industrial Engineering*, vol. 49, no. 1, pp. 63–75, Aug. 2005, http://dx.doi.org/10.1016/j.cie.2005.02.001.

[26] J. Kostolny, M. Kvassay, and S. Kovalik, "Reliability analysis of noncoherent systems by logical differential calculus and binary decision diagrams," *Komunikacie*, vol. 16, no. 1, pp. 114–120, 2014.

[27] S. N. Yanushkevich, D. M. Miller, V. P. Shmerko and R. S. Stankovic, *Decision Diagram Techniques for Micro- and Nanoelectronic Design. Handbook*. Boca Raton, FL: CRC Press, 2006, 952 p.

# 5th International Workshop on Advances in Semantic Information Retrieval

RECENT advances in semantic technologies form a solid basis for a variety of methods and instruments that support multimedia information retrieval, knowledge representation, discovery and analysis. They influence the way and form of representing documents in the memory of computers, approaches to analyze documents, techniques to mine and retrieve knowledge. The abundance of video, voice and speech data also raises new challenging problems to multimedia information retrieval systems.

We believe that our workshop will facilitate discussions of new research results in this area, and will serve as a meeting place for researchers from all over the world. Our aim is to create an atmosphere of friendship and cooperation for everyone, interested in computational linguistics and semantic information retrieval. The ASIR'15 workshop will continue to maintain high standards of quality and organization, set in the previous years. We welcome all the researchers, interested in semantic information retrieval, to join our event.

## Topics

The workshop addresses semantic information retrieval theory and important matters, related to practical Web tools. The topics and areas include but not limited to:

- Domain-specific semantic applications.
- Evaluation methodologies for semantic search and retrieval.
- Models for document representation.
- Natural language semantic processing.
- Ontology for semantic information retrieval.
- Ontology alignment, mapping and merging.
- Query interfaces.
- Searching and ranking.
- Semantic multimedia retrieval.
- Visualization of retrieved results.

## Event Chairs

**Klyuev, Vitaly,** University of Aizu, Japan

**Mozgovoy, Maxim,** University of Aizu, Japan

## Program Committee

**Carrara, Massimiliano,** Universita di Padova, Italy

**Dobrynin, Vladimir,** Saint Petersburg State University, Russia

**Goczyła, Krzysztof,** Gdansk University of Technology, Poland

**Haralambous, Yannis,** Institut Telecom - Telecom Bretagne, France

**Homenda, Wladyslaw,** Warsaw University of Technology, Poland

**Jin, Qun,** Waseda University, Japan

**Lai, Cristian,** CRS4, Italy

**Leonelli, Sabina,** University of Exeter, United Kingdom

**Nalepa, Grzegorz J.,** AGH University of Science and Technology, Poland

**Pyshkin, Evgeny,** Peter the Great St. Petersburg Polytechnic University, Russia

**Shtykh, Roman,** CyberAgent Inc., Japan

**Slezak, Dominik,** University of Warsaw & Infobright Inc., Poland

**Soldatova, Larisa,** Brunel University, United Kingdom

**Suárez-Figueroa, Mari Carmen,** Ontology Engineering Group, Scool of Computer Science at Universidad Politécnica de Madrid, Spain

**Tadeusiewicz, Ryszard,** AGH University of Science and Technology, Poland

**Vacura, Miroslav,** University of Economics, Czech Republic

**Zadrozny, Slawomir,** Systems Research Institute, Poland

**Ławrynowicz, Agnieszka,** Poznan University of Technology, Poland

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 323–328
DOI: 10.15439/2015F130
ACSIS, Vol. 5

# The data retrieval optimization from the perspective of evidence-based medicine

Vladimir Dobrynin*, Julia Balykina*, Michael Kamalov*,
Alexey Kolbin†, Elena Verbitskaya† and Munira Kasimova‡
*Saint-Petersburg State University, Universitetskaya nab., 7-9, Saint-Petersburg, Russia
Email: v.dobrynin@bk.ru, Email: {julia.balykina, mkamalovv} @gmail.com
†Pavlov First Saint-Petersburg State Medical University, L'va Tolstogo str., 6/8, Saint-Petersburg, Russia
Email: alex.kolbin@mail.ru, Email: elena.verbitskaya@gmail.com
‡Tashkent Institute of Postgraduate Medical Education, Parkent str., 51, Tashkent, Uzbekistan
Email: drkasimovamunira@mail.ru

***Abstract*—The paper is devoted to classification of MEDLINE abstracts into categories that correspond to types of medical interventions - types of patient treatments. This set of categories was extracted from Clinicaltrials.gov web site. Few classification algorithms were tested including Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM implementations from sklearn machine learning library. Document marking was based on the consideration of abstracts containing links to the Clinicaltrials.gov Web site. As the result of an automatical marking 3534 abstracts were marked for training and testing the set of algorithms metioned above. Best result of multinomial classification was achieved by Linear SVM with macro evaluation precision 70.06%, recall 55.62% and F-measure 62.01%, and micro evaluation precision 64.91%, recall 79.13% and F-measure 71.32%.**

## I. Introduction

AT THE moment an evidence-based medicine approach is actively developing in medical practice. This approach requires an expert to choose a method of patient treatment based on available evidences of safety and efficiency of the method. Complexity of evidence-based medicine application in practice involves not only control over saving new research results, but also assessment of the quality and reliability of existing ones. To solve this problem in evidence-based medicine a grading scale for ranking studies by level of evidence is used. For example, in the USA the National Guideline Clearinghouse[1] recommends to follow levels of evidence and grades (table I, table II).

TABLE I
Levels of Evidence

| | |
|---|---|
| I A | Evidence from meta-analysis of randomized controlled trials (RCTs) |
| I B | Evidence from at least one randomized controlled trial |
| II A | Evidence from at least one controlled study without randomization |
| II B | Evidence from at least one other type of quasi-experimental study |
| III | Evidence from non-experimental descriptive studies, such as comparative studies, correlation studies, and case-control studies |
| IV | Evidence from expert committee reports or opinions or clinical experience of respected authorities, or both |

TABLE II
Grades of Recommendations

| | |
|---|---|
| A | Directly based on Level I evidence |
| B | Directly based on Level II evidence or extrapolated recommendations from Level I evidence |
| C | Directly based on Level III evidence or extrapolated recommendations from Level I or II evidence |
| D | Directly based on Level IV evidence or extrapolated recommendations from Level I, II, or III evidence |

However, in some cases studies corresponding to the first level of evidence may contain errors in the correctness of randomized controlled trials (RCTs). More detailed description of some examples with errors in studies with the first level of evidence is reviewed in [1]. Solution of the problem is the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system. This system evaluates level of evidence for different studies and ranks them by recommendation significance with due consideration of additional criteria for evaluation. Additional criteria for GRADE are presented in table III and described in more detail in [2]. GRADE considers only two classes of recommendations: strong or low-level recommendations. The quality level of evidence is presented in 4 levels. Thus, using additional GRADE factors it becomes possible to rise or lower the value of research.

Another actively developing research trend is information retrieval application in the field of medicine based directly on the use of MEDLINE[2] database. For example, in [3] MEDIE search engine developed for MEDLINE database, that executes semantic search, keyword search and generalized concordance lists (GCL) search is described. In [4] the Hierarchical Hidden

[1] http://www.guideline.gov/

[2] http://www.nlm.nih.gov/

978-83-60810-66-8/$25.00©2015, IEEE

TABLE III
QUALITY ASSESSMENT CRITERIA

| Study design | Quality of evidence | Lower if | Higher if |
|---|---|---|---|
| Randomized trial | High | Risk of bias<br>-1 Serious<br>-2 Very serious<br>Inconsistency<br>-1 Serious<br>-2 Very serious<br>Indirectness<br>-1 Serious<br>-2 Very serious<br>Imprecision<br>-1 Serious<br>-2 Very serious<br>Publication bias<br>-1 Serious<br>-2 Very serious | Large effect<br>+1 Large<br>+2 Very large<br>Dose response<br>+1 Evidence of gradient<br>All plausible can founding<br>+1 Would reduce a demonstrated effect or<br>+1 Would suggest a spurious effect when results show no effect |
| | Moderate | | |
| Observational study | Low | | |
| | Very low | | |

Markov Models algorithm for retrieving information about protein and its location from the MEDLINE abstract database is considered. Study [5] considers a method of automatic term extraction developed specifically for indexing documents from large medical collections. Computational experiments are conducted on a set of documents from MEDLINE database. In [6] an unsupervised clustering technique called SOPHIA is presented, that is evaluated on the MEDLINE testing set collection. Study [7] describes an experiment that changes the ranking strategy using the term-graph data structure for assessing the importance of a document to a user's query to the MEDLINE database. In [8] existent question-answering system based on principles of evidence based medicine is presented. Study [11] describes a fuzzy VIKOR framework for ranking internet health information providers.

Based on the relevance and demand for joint studies in the field of medicine and information retrieval, it was decided to start a development of a search engine for the MEDLINE database on the basis of the Saint-Petersburg State University with the support of Pavlov First Saint-Petersburg State Medical University and Tashkent Institute of Postgraduate Medical Education. The main goal of the project is to develop a new ranking method for search results, which takes account for a level of evidence and GRADE criteria.

## II. Problem Statement

The object of analysis of this work are documents containing abstracts of articles from the MEDLINE international database of medical research. The goal is to group abstracts according to subtypes of medical interventions. Subtypes of medical interventions correspond to various methods for patient treatment and prophylaxis. Examples of medical intervention subtypes were taken from the Clinicaltrials.gov[3] Internet resource:

1). Drug;
2). Biological;
3). Device;
4). Dietary Supplement;
5). Procedure;
6). Radiation;
7). Behavioral;
8). Genetic;
9). Other.

This website is a public register approved by the US International Committee of Medical Journal Editors. It provides relevant structured information about conducting clinical studies for a wide range of diseases.

The assigned task falls within the domain of machine learning and requires an implementation of the following auxiliary problems:

1). Development of a method for automatic markup of abstracts from a training and test set by medical intervention subtypes, based on the existence of a link between documents that represent paper abstracts from MEDLINE database and contents of registered clinical trials on Clinicaltrials.gov. The link is presented by reference to Clinicaltrials.gov Web resource.

2). Training methods for multinomial classification by means of selected set of classical algorithms such as Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM from the sklearn[4] library for further evaluation and selection of a more effective algorithm.

First it was decided to evaluate linear multinomial classification algorithms for an obtained marked sample of MEDLINE abstracts. Therefore, the following linear algorithms were chosen: Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM. In the future it is planned to choose a set of nonlinear multinomial classification algorithms and conduct experiments with the same marked sample of MEDLINE abstracts.

## III. Methods for abstracts markup

At the first stage of handling the problem 90 paper abstracts of the year 2011 taken from MEDLINE database and

[3] https://clinicaltrials.gov/

[4] http://scikit-learn.org/

which contained links to the Clinicaltrials.gov Web site were examined. To simplify abstract processing, an abstracts.xml document was created which has the following structure (per entry). Document structure (1):

```
<document>
   <doc_id></doc_id>
   <date></date>
   <title></title>
   <body></body>
   <topics></topics>
   <place></place>
   <author></author>
   <type></type>
</document>
```

where: <document> is a document container; <doc_id> is a paper identifier; <date> – publication date of the article; <title> – title of the article; <body> – body of the abstract; <topics> – article keywords; <place> – journal where the article was published; <author> – paper authors; <type> – a subtype of medical intervention. All abstracts were transfered into this structure. Every record in abstracts.xml satisfies the structure listed above.

These 90 documents were marked manually based on the search for links between abstracts and Clinicaltrials.gov Web resource. Linkage was performed by finding in abstract a reference (eg., NCT00893711). Such a reference meant that the study was indexed at Clinicaltrials.gov. The sequence of manual markup was as follows:

1) A link to Clinicaltrials.gov site was retrieved from an abstract.

2) With the help of Clinicaltrials.gov internal search engine a search was performed in order to find studies corresponding to the reference given in the abstract. An example of using internal search engine is presented in Fig. 1.

3) After the search, a found subtype of medical intervention was manually added to the document structure in the <type> </type> field. It was proposed to impose a restriction on the Clinicaltrials.gov web-resource search results: if the study was represented by two subtypes of medical intervention as it is shown in Fig. 1, it was suggested to use the first subtype because it contains the main information about the study.

As a result of manual marking it was possible to group 60 out of 90 abstracts into the following subtypes: Behavioral, Biological, Device, Dietary_Supplement, Drug, Other, and Procedure. The remaining 30 abstracts were divided into 4 groups:

1) no link to clinicaltrials.gov;

2) contains a link, but no information about the subtype of medical intervention;

3) contains a link but studies are of observational type;

4) contains a link with an error (eg., ISRCTN51481987).

It was further decided to consider corpus containing 2000000 abstracts for years 2006 to 2013 and automate the process of markup. An automation has been implemented with the help of an application developed in python that performs a search for links in abstracts (eg. NCT00893711) as a regular expression and a web-crawler that searches through links http://clinicaltrials.gov/show/NCT00776256?resultsxml=true replacing the part (NCT00893711) with the one found in the abstract and extracting data from the xml page contained in the <intervention_type> field. Results extracted from <intervention_type> field were then added to the <type> </ type> field of document structure (1) using the developed software application. The result of parsing xml pages also imposes restrictions: in case of two fields <intervention_type> is marked by the first field.

As a result of an automatic marking of 2000000 abstracts, 3534 abstracts were marked. Remaining abstracts were divided into groups:

1) no link to clinicaltrials.gov;

2) contained a link, but when referring to web-crawler on corresponding page an error appeared "404 - page not found";

3) contained a link but corresponding studies were of observational type;

4) contained a link with an error (eg., CTNO1481987).

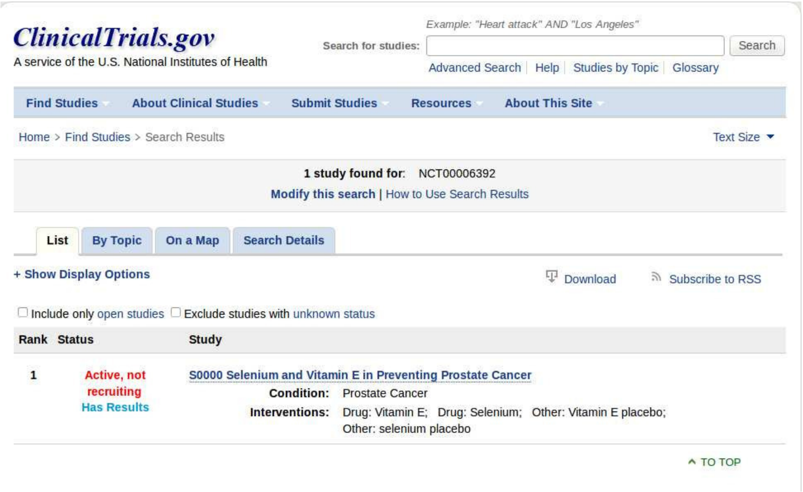
As a result of marking, every subtype of medical intervention aggregated the following number of the abstracts: Behavioral – 585, Biological – 242, Device – 238, Dietary Supplement – 191, Drug – 1619, Other – 333, Procedure – 300, Radiation – 18, and Genetic – 8. After that, 3534 abstracts were divided into training set and testing set for performing training and testing of the following classifiers: Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM using the sklearn library. Also, experiments with changing parameters of classifiers were performed in order to determine the most efficient algorithm for classification.

## IV. Experimental part

This section presents experiment results for multinomial classification of automatically marked 3534 MEDLINE abstracts by subtypes of medical interventions. Such algorithms as Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM were used (names of algorithms in sklearn library: MultinomialNB, LinearSVC, LogisticRegression). Marked abstracts were divided into training set and testing set in ascending order by date of publication as follows:

- For training:
  - 2651 abstracts from 2006 till 2012 year containing the following number of classes with abstracts: Behavioral - 450, Biological - 174, Device- 189, Dietary Supplement - 145, Drug - 1198, Other - 266, Procedure - 209, Radiation - 12, Genetic - 8.
- For testing:
  - 883 abstracts from 2012 till 2013 year containing the following number of classes with abstracts: Behavioral - 135, Biological - 68, Device - 49, Dietary Supplement - 46, Drug - 421 Other - 67, Procedure - 91, Radiation - 6.

As a partition result, testing set contained the following sub-types of medical interventions: Behavioral, Biological,


ClinicalTrials.gov
A service of the U.S. National Institutes of Health
Example: "Heart attack" AND "Los Angeles"
Search for studies:
Search
Advanced Search | Help | Studies by Topic | Glossary
Find Studies
About Clinical Studies
Submit Studies
Resources
About This Site
Home > Find Studies > Search Results
Text Size
Modify this search | How to Use Search Results
List
By Topic
On a Map
Search Details
+ Show Display Options
Download
Subscribe to RSS
TO TOP

1 study found for: NCT00006392

□ Include only open studies □ Exclude studies with unknown status

| Rank | Status | Study |
|---|---|---|
| 1 | Active, not recruiting<br>Has Results | S0000 Selenium and Vitamin E in Preventing Prostate Cancer<br>Condition: Prostate Cancer<br>Interventions: Drug: Vitamin E; Drug: Selenium; Other: Vitamin E placebo; Other: selenium placebo |

Fig. 1. Search results by link at Cliniclatrials.gov

Dietary Supplement, Drug, Other, Procedure, and Radiation. To evaluate classification result for the testing set for each subtype of medical intervention such measures as precision, recall and F-measure were used:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F\text{-}measure = 2 * \frac{Precision * Recall}{Precision + Recall},$$

where $TP$ – true positive classification value, i.e. the classifier identified an element of the testing set correctly. $FP$ – false positive value, i.e. the classifier referred an element to the class falsely. $FN$ – false negative elements, i.e. the classifier falsely did not refer an element to the class. To assess different multinomial classification algorithms, macro and micro precision and recall values, as well as F-measure were calculated using the following formulas:
Macro:

$$precision = \frac{\sum_{n=1}^{m} Precision_n}{m},$$

$$recall = \frac{\sum_{n=1}^{m} Recall_n}{m},$$

$$F\text{-}measure = 2 * \frac{precision * recall}{precision + recall}.$$

Micro:

$$precision = \frac{\sum_{n=1}^{m} TP_n}{\sum_{n=1}^{m} TP_n + \sum_{n=1}^{m} FP_n},$$

$$recall = \frac{\sum_{n=1}^{m} TP_n}{\sum_{n=1}^{m} TP_n + \sum_{n=1}^{m} FN_n},$$

$$F\text{-}measure = 2 * \frac{(precision) * (recall)}{(precision) + (recall)},$$

where $m$ is the number of classes. Calculation of micro precision and recall was performed by summing up all true positive, false positive and false negative results of classification for each class.

In this experiment we use the «bag of words» model [9]. For vector of features we use vector of terms from the dictionary, composed of all annotations from corpora. In process of forming dictionary no stemming was used. With the help of $tf\text{-}idf$ metric, weight for every term was assessed.

$$tf\text{-}idf(t, d) = tf(t, d) * idf(t),$$

where $tf$ – term frequency:

$$tf(t, d) = \frac{n_{t,d}}{|d|},$$

$t$ – term, $d$ – document, $n_{t,d}$ – entry of $t$ term occurence in $d$ document, $|d|$ – total number of terms in $d$ document; $idf$ – inverse document frequency:

$$idf(t) = \log \frac{N}{df(t)},$$

where $N$ – number of documents in corpora; $df$ – number of documents, in which term $t$ occures.

During the experiments, the optimal parameters of classifiers were selected for the case with the removal of stop words, as well as for the case with no stop words removing from the dictionary. More detailed options of the algorithms and their values are given in the documentation for the library

TABLE IV
RESULT OF CLASSIFICATION WITH THE REMOVAL OF STOP WORDS

| | **Precision** | | | **Recall** | | | **F-measure** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Multinomial Naive Bayes** | **Linear SVM** | **Maximum Entropy** | **Multinomial Naive Bayes** | **Linear SVM** | **Maximum Entropy** | **Multinomial Naive Bayes** | **Linear SVM** | **Maximum Entropy** |
| **Behavioral** | 77% | 73% | 70% | 49% | 78% | 76% | 60% | 75% | 73% |
| **Biological** | 100% | 89% | 96% | 24% | 72% | 65% | 38% | 80% | 77% |
| **Device** | 0% | 62% | 71% | 0% | 41% | 10% | 0% | 49% | 18% |
| **Dietary Supplement** | 0% | 58% | 58% | 0% | 63% | 24% | 0% | 60% | 34% |
| **Drug** | 53% | 79% | 65% | 99% | 94% | 97% | 69% | 86% | 78% |
| **Other** | 100% | 34% | 50% | 1% | 22% | 9% | 3% | 27% | 15% |
| **Procedure** | 0% | 64% | 65% | 0% | 42% | 14% | 0% | 51% | 23% |
| **Radiation** | 0% | 100% | 0% | 0% | 33% | 0% | 0% | 50% | 0% |

TABLE V
RESULT OF CLASSIFICATION WITH THE REMOVAL OF STOP WORDS (MICRO AND MACRO EVALUATION)

| | **Precision** | | **Recall** | | **F-measure** | |
|---|---|---|---|---|---|---|
| | **Micro** | **Macro** | **Micro** | **Macro** | **Micro** | **Macro** |
| **Multinomial Naive Bayes** | 55.77% | 41.27% | 86.87% | 21.62% | 67.93% | 28.37% |
| **Linear SVM** | 64.91% | 70.06% | 79.13% | 55.62% | 71.32% | 62.01% |
| **Maximum Entropy** | 65.66% | 59.34% | 76.63% | 36.93% | 70.72% | 45.53% |

TABLE VI
RESULT OF CLASSIFICATION WITHOUT STOP WORDS REMOVAL

| | **Precision** | | | **Recall** | | | **F-measure** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Multinomial Naive Bayes** | **Linear SVM** | **Maximum Entropy** | **Multinomial Naive Bayes** | **Linear SVM** | **Maximum Entropy** | **Multinomial Naive Bayes** | **Linear SVM** | **Maximum Entropy** |
| **Behavioral** | 87% | 73% | 74% | 1% | 84% | 76% | 17% | 78% | 75% |
| **Biological** | 0% | 89% | 96% | 0% | 75% | 65% | 6% | 82% | 77% |
| **Device** | 0% | 56% | 67% | 0% | 39% | 8% | 0% | 46% | 15% |
| **Dietary Supplement** | 0% | 59% | 57% | 0% | 70% | 26% | 0% | 64% | 36% |
| **Drug** | 48% | 80% | 65% | 100% | 93% | 97% | 65% | 86% | 78% |
| **Other** | 0% | 59% | 63% | 0% | 38% | 13% | 0% | 47% | 22% |
| **Procedure** | 0% | 100% | 0% | 0% | 33% | 0% | 0% | 50% | 0% |
| **Radiation** | 0% | 100% | 0% | 0% | 33% | 0% | 0% | 50% | 0% |

TABLE VII
RESULT OF CLASSIFICATION WITHOUT STOP WORDS REMOVAL (MICRO AND MACRO EVALUATION)

| | **Precision** | | **Recall** | | **F-measure** | |
|---|---|---|---|---|---|---|
| | **Micro** | **Macro** | **Micro** | **Macro** | **Micro** | **Macro** |
| **Multinomial Naive Bayes** | 49.03% | 16.88% | 77.87% | 13.67% | 60.18% | 15.10% |
| **Linear SVM** | 61.72% | 69.72% | 75.74% | 56.73% | 68.01% | 62.56% |
| **Maximum Entropy** | 63.49% | 57.87% | 72.62% | 37.19% | 37.19% | 45.28% |

sklearn. Below the optimum parameters of the algorithms are presented.

The following algorithm parameters were used for the classification of documents from the test set with no account for stop words:

MultinomialNB: alpha = 1.0, fit_prior = True;

LinearSVC: penalty = 'l2', loss = 'squared_hingle', multi_class = 'ovr', C = 1.0;

LogisticRegression: penalty = 'l2', multi_class = 'ovr', C = 1.0, solver = 'liblinear';

Results are presented in tables IV, V.

The following algorithm parameters were used for the classification of the testing set with stop words in the dictionary:

MultinomialNB: alpha = 2.0, fit_prior = True;

LinearSVC: penalty = 'l1', loss = 'hingle', multi_class = 'crammer_singer', C = 0.8;

LogisticRegression: penalty = 'l1', multi_class = 'multinomial', C = 0.8, solver = 'newton-cg'.

Corresponding results are presented in tables VI, VII.

Based on the results of computational experiments, the best results were obtained without accounting for stop words in the dictionary and when using LinearSVC with the following parameters: penalty = 'l2', loss = 'squared_hingle', multi_class = 'ovr', C = 1.0.

Corresponding results for macro evaluation: precision= 70.06%; recall = 55.62%; F-measure = 62.01%. Results of micro evaluation: precision = 64.91%; recall= 79.13%; F-measure = 71.32%.

## V. Discussion

Relatively low classification quality rates are associated with the fact, that documents for classification describe medical studies, which were performed during patients treatment. Some differences in documents from various classes are related only to subtypes of medical treatments, that were considered in the studies, and can describe patients suffering from the same disease. The result that was recieved can be compared with results from the research [10],where maximum F-measure value of 80% has been achieved by using linear SVM during the classification of abstracts on RCTs and on non RCTs. In our case maximum value of macro F-measure = 62.01% and micro F-measure = 71.32% has been also retrieved when using linear SVM, with multiclassification of abstracts by subtypes of medical interventions.

## VI. Conclusion

This article describes methods that allow to automate grouping MEDLINE abstracts by subtypes of medical interventions. Computational experiments were carried out using the following algorithms: Multinomial Naive Bayes, Multinomial Logistic Regression, and Linear SVM from the sklearn library. Linear SVM algorithm showed the best result of multinomial classification.

For further research it is planned to perform the following tasks:

- chose the set of nonlinear multinomial classifier algorithms and examine 3534 MEDLINE abstracts using these algorithms;
- classify the remaining 1996466 unmarked abstracts using Linear SVM algorithm;
- extract facts from the marked abstracts about a specific subtype of a medical intervention described in the study;
- group abstracts by subtypes of medical intervention using a catalog of natural science subjects MESH[5] and contents of the <topics> field from the document structure (1).

## References

[1] G. Guyatt,G. Vist, Y. Falck-Ytter, R. Kunz, N. Magrini ,H. Schunemann for the GRADE* working group. "An emerging consensus on grading recommendations?," (Editorial). *ACP J Club,* 2006, Jan-Feb;144(1):A08, PMID: 17216711.

[2] G. Guyatt,G. Vist, Y. Falck-Ytter, R. Kunz, N. Magrini ,H. Schunemann for the GRADE* working group."GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables,"*Journal of Clinical Epidemiology,* 2011, vol. 64, pp. 383-394, doi: 10.1016/j.jclinepi.2010.04.026.

[3] T. Ohta, Y. Tsuruoka, J. Takeuchi, J. Kim, Y. Miyao, A. Yakushiji et al. "An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing," *Proceedings of the COLING/ACL on Interactive presentation sessions,* Stroudsburg, PA, USA, 2006, vol. 4, pp. 17-20, doi: 10.3115/1225403.1225408.

[4] S. Kaneko, A. Hayashi, N. Suematsu, K. Iwata, "Hierarchical hidden conditional random fields for information extraction," *Proceedings of the 5th international conference on Learning and Intelligent Optimization,* Springer-Verlag Berlin, Heidelberg, 2011, vol. 12, pp. 191-202, doi: 10.1007/978-3-642-25566-3_14.

[5] A. Hliaoutakis, K. Zervanou, E. G.M. Petrakis, E. E. Milios, "Automatic document indexing in large medical collections," *Proceedings of the international workshop on Healthcare information and knowledge management*, New York, USA, 2006, vol. 8, pp. 1-8, doi: 10.1145/1183568.1183570.

[6] V. Dobrynin, D. Patterson, M. Galushka, N. Rooney, "SOPHIA: An Interactive Cluster Based Retrieval System for the OHSUMED collection," *in IEEE Trans. on Information Technology for Biomedicine*, 2005 , vol. 9, pp. 256-265, PMID: 16138542 .

[7] K. Veningston, R. Shanmugalakshmi, "Information Retrieval by Document Re-ranking using Term Association Graph," *Proceedings of the 2014 International Conference on Interdisciplinary Advances in Applied Computing*, New York, USA, 2014 , vol. 8, Article No. 21., doi:10.1145/2660859.2660927

[8] D. Demner-Fushman, J. Lin, "Answering Clinical Questions with Knowledge-Based and Statistical Techniques," *Journal of Computational Linguistics*, 2007 , vol. 33, pp. 63-103, doi: 10.1162/coli.2007.33.1.63

[9] Ch. D. Manning, P. Raghavan, H. Schutze, "Introduction to Information Retrieval," *Cambridge University Press*, Cambridge, England, 2008 , pp. 482, isbn: 9780521865715

[10] A. M. Cohen , N. R. Smalheiser , M. S. McDonagh , C. Yu , C. E. Adams , J. M. Davis et al. "Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine," *Journal of Am Med Inform Assoc* , 2015, pp. 707-717, doi: http://dx.doi.org/10.1093/jamia/ocu025

[11] E. Afful-Dadzie , S. Nabareseh , S. Kominkova Oplatkova, "Fuzzy VIKOR approach: evaluating quality of internet health information," *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems* , Warsaw, Poland, 2014, vol. 2, pp. 183-190, doi: 10.15439/2014F203

[5] http://www.nlm.nih.gov/mesh/

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 329–335

DOI: 10.15439/2015F50
ACSIS, Vol. 5

# Mapping Evaluation for Semantic Browsing

Veslava Osinska
Institute of Information Science and Book Studies, Nicolaus Copernicus University,
ul. Bojarskiego 1, 87-100 Toruń, Poland
Email: wieo@umk.pl

Adam Jozwik
Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Science, Warsaw;
Institute of Computer Science, College of Social and Media Culture, Toruń, Poland
Email: adamj346@wp.pl

Grzegorz Osinski
Institute of Computer Science, College of Social and Media Culture, ul. Starotoruńska 3
87-100 Toruń, Poland
Email: gos@fizyka.umk.pl

***Abstract*—The paper contributes to the problem solving in semantic browsing and analysis of scientific articles. With reference to presented visual interface, four – the most popular methods of mapping including own approach - MDS with spherical topology, have been compared. For a comparison quantitative measures were applied which allowed to select the most appropriate mapping way with an accurate reflection of the dynamics of data. For the quantitative analysis the authors used machine learning and pattern recognition algorithms and described: clusterization degree, fractal dimension and lacunarity. Local density differences, clusterization, homogeneity, and gappiness were measured to show the most acceptable layout for an analysis, perception and exploration processes. Visual interface for analysis how computer science evolved through the two last decades is presented on website. Results of both quantitative and qualitative analysis have revealed good convergence.***

## I. Introduction

Nonlinear growth of scientific writing imposes a new forms of academic databases management. The latter includes the both retrieval and analytical exploration. Analysts, science of science professionals, science policy makers need various computing, statistical and visualization tools to monitor how science or scientific domains evolved.

Authors designed and described in a series of papers [1], [2] the visual interface for analysis of dynamics of computer science through the two last decades[1]. Screenshot of spherical application is shown on Figure 1. Users can interact, manipulate and browse the data and see how graphical pattern change in time. The nodes represent scientific articles from digital library and the colour – appropriate thematic category. Similarity metrics was based on semantic relations between documents [2]. In order to generate 3D layout, multidimensional scaling (MDS) technique was applied and enriched by Morse potential [1].

Overlapping spots show where the categories mutually integrate that means an articles at that location are semantically similar. Visualization of classified documents reveals both organization of digital library content as well as allows users to track how it changes over time. This paper presents further study on visualization maps. Authors decided to test this prototype regarding to mapping algorithms. Four mapping methods were compared in terms of dynamics and analytical possibility of output visualization. The next chapter shows the outline of VxOrd, MDS, VOS, SOM as the most popular methods.

## II. Mapping preview

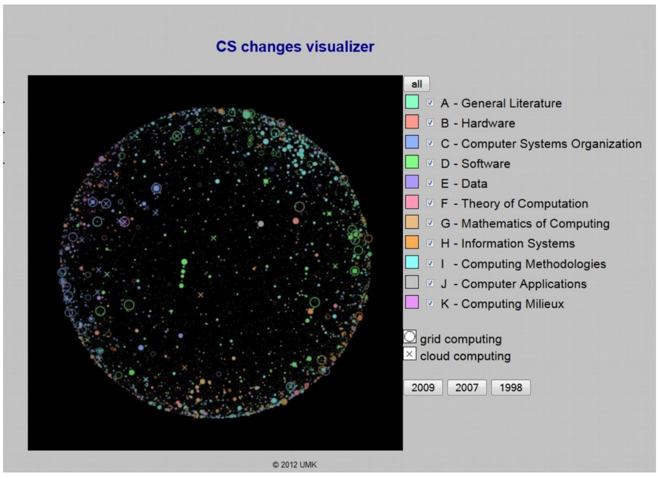
How we, as analysts, perceive and understand the connections between data, depends on graphical layout. Thus, the final structure of visualized knowledge can be drawn either by spatial arrangement (2D or 3D) of analysis units or by the relationship between nodes in graph or combination of these two.

One of the basic ordination algorithm - VxOrd extends a traditional force-directed approach [3]. VxOrd determines the both number and size of clusters automatically based on the data. Popular software for data mapping and visualization - Gephi[2] uses this technique. Due to Gephi users can analyse large networks consisting of even millions of nodes. The most popular technique for dimension reduction is MDS, which involves minimizing the difference between Euclidean and graph-theoretic distances. MDS has been widely applied for constructing knowledge maps of authors, articles, journals, and keywords [3]-[5]. The same satisfactory representation of knowledge can be produced by use of new mapping technique VOS introduced in series of works by Van Eck and Waltman [6],[7]. The idea of VOS is to minimize a weighted sum of the squared distances between all pairs of items.

Dimension reduction can be also achieved in self organized maps – the kind of unsupervised neural network which aimed to project high-dimensional data into a lower-dimensional space [4]. The nodes (input vectors) form two dimensional regular grid; node's neighbourhood is defined to be all connected nodes. During training process similar input vectors stimulate adjacent neurons and therefore output SOM map shows semantic relationships between data, where similar items are mapped close together. Comparing MDS and VOS,

[1] http://www-users.mat.umk. pl/~garfi/vis2009v3/

[2] http://gephi.github.io/



☑ A - General Literature
☑ B - Hardware
☑ C - Computer Systems Organization
☑ D - Software
☑ E - Data
☑ F - Theory of Computation
☑ G - Mathematics of Computing
☑ H - Information Systems
☑ I - Computing Methodologies
☑ J - Computer Applications
☑ K - Computing Milieux

○ grid computing
☒ cloud computing

Fig. 1 Interface screenshot - the prototype of application for semantic browsing

researchers concluded that maps constructed using VOS approach provide a more satisfactory representation of the underlying dataset [8]. MDS-VOS tests revealed the first one is distance preserving while the second - topology preserving technique [9],[10].

## III. Methodology

The complex structure of large graphs commonly is measured by modularity. According to Fortunato [11], modularity can be defined as the function which evaluates the goodness of partitions of graph and is defined by ties between vertices, vertices and hubs. Structure described by linked objects does not match with character of present data. Authors analyse the visualization of classified documents. Primary categories are taken from library classification. All relationships between collected classes and documents were predefined by specialists permanently working on computer science taxonomy. Visualized articles are assumed to reveal semantics while keeping their thematic similarity. These new correlations (down-top) between data allow to rich independence of the original organization of items (top-down). As every rigid scheme, it is characterized by adequacy and disjointedness of subclasses. Therefore, evaluation metrics can be based on spatial configuration of visual layout instead of links outline. To analyze the final nodes distribution, image processing methods were applied. Then evaluation of graphical pattern should be carried out taking into account the accuracy, topology (space filling and capacity) and perception abilities of users.

The authors present alternative approach: a sphere surface has been selected as a target mapping space. There are some arguments for a sphere surface: it “has no edges and therefore it is possible to represent not only local similarities but also large-scale ones regarding the whole space. The benefit of a curved surface in comparison to a plane one is a more capacious exploration space” [1,10]. 3D visualization is a popular but also challenging method in large dataset mapping and modelling.

### A. Assumptions

By implication, evaluation process will touch how to fit interface to the requirements of analysts and domain experts. The study is based on the following assumptions.

1. A given visualization layout might serve as a graphical interface for the exploration and semantic retrieval of scientific articles. From this point of view the most important is configuration on the bottom level (documents) – then can be evaluated the spatial distribution of nodes.
2. Current modifications by editors of the original classification are aiming at its improvement. The classification reflects the most of current changes in computer science. Quickly developing categories will form dense clusters and overlap each other. These tendencies must be visible on visualization maps generated for different time periods.
3. In the construction of the ergonomic user interface, such features as capacity, homogeneous distribution and edgelessness must be taken into consideration.

Short movie[3] shows how three dimensional configuration allows the user to analyze semantic distribution of articles and its behavior in time.

### B. Evaluation steps

On Figure 2 we can see the elements of evaluation process. Continuity characteristics is crucial for present study. For this purposes clusterization potential was validated by machine learning and image recognition algorithms. Structural complexity can be evaluated by fractals analysis. Quantitative measures are different for several maps and the changes tendency are essential too.

All dimension reduction methods determine the arrangement of classes and subclasses nodes. Documents distribution was calculated by using geometrical rules in 2D or 3D space [10]. Obtained pattern became the basic material for comparison and further study.

If we plan to involve users to scientific domain analysis, visualization interface must be user-friendly and carry good navigational features. Another usefulness of such application is retrieval of semantically similar documents. Precision in this case will be an appropriate measure of this visual searching system.

### C. Research material

Visualization maps were obtained by using the same data but distinct in terms of data configuration (like matrixes versus data pairs), mapping algorithms and space topology. The series of every ten-year layouts show the changes of pattern and thus the evolution of the ACM classification and computer science knowledge (see Appendix). An insight into the differences in graphical patterns could reveal the most and the least complex structures due to human perception. The system of human perception is able to recognize a natural tex

[3] www.wizulizacjainformacji.pl/unas/interface.avi

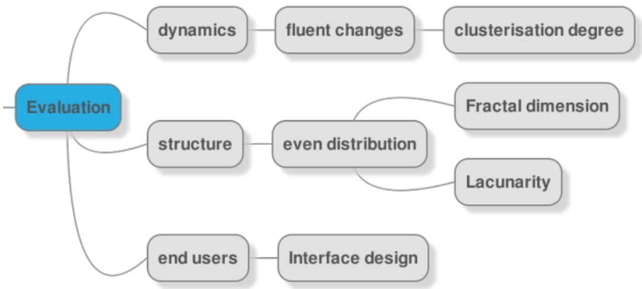


Fig. 2 Decision map of visualization evaluation steps

ture appearing in nature as a result of evolutionary adaptation. The human vision allows determining approximately whether the perceived structure differs one from another in terms of complexity[12].

For example, the first series of maps is characterized by a relative even distribution while VxOrd by data grouping on edges and bends. Furthermore, VxOrd and VOS distributions are highly limited to the output geometry [10]. The result may be a non-effective space for navigation in those cases. But human perception cannot be one of the main criteria for comparison and estimation of visual layouts, although useful in the final conclusions. Quantitative approach requires that the authors analyze local density differences and quantify clusterization, rarefaction, homogeneity and porosity.

Output maps can be described by both density and colour of nodes. If information about the main thematic category assignment is excluded (i.e. the colour), the clusters can be identified by density only. Consequently, clusterization and its changes can deliver information on how knowledge advances and how knowledge organization changes throughout two decades, independently of the primary (original) classification.

## IV. Quantitative Analysis of Maps

### A. Clusterization and its dynamics

To identify clusters on given maps we used the most popular partition technique, based on distances between points and/or points to centroids – k-means clustering [11]. Algorithm aims to minimize the within-cluster sum of squares:

$$\sum_{i=1}^{k}\sum_{x_j\in S_i}\|x_j-c_i\|^2 \qquad (1)$$

where $S_i$ indicates the subset of points in the $i$-cluster, $c_i$ – the centroid of cluster.

The disadvantage of this algorithm in our assignment is the requirement to know the number of clusters. To find optimal number of classes we modeled the data by a set of Gaussian distributions [13].

The number of components can be estimated by the Bayesian Information Criterion [14] (BIC), which is based on a penalization of the observed log-likelihood – the function of $x$, $\theta$. The preffered model is the one with the lowest value of BIC which decides about the number of clusters. Thus the 6 clusters are recognized for the MDS –sphere (authors method) and SOM and 8 for both VxOrd and VOS. These new clusters reorqanize initial data assignment to the 11 main initial categories, coded by colour. By k-means clustering the centroids of clusters are found, demonstrated on Figure 3 according to the the MDS –sphere map.

Dynamical characteristics of clustering are crucial for a final evaluation of the presented approaches in terms of structural analysis. In any sequence of maps it is possible to find the one with a highly developed clusterization just intuitively.

To evaluate clustering and its dynamics, a misclassification rate offered by the standard k nearest neighbour (k-NN) rule was used as a criterion. That error rate was estimated by the leave-one-out method [15],[16]. The k-NN rule assigns the classified object to the class most heavily represented from among its nearest objects in the training set (i.e. nearest neighbours). The reference set, also called a training set, is a set of objects with a known class membership and in a certain sense it defines the considered classes. The leave-one-out method consists in the classification of each object from the reference set by the decision rule obtained from the training set decreased by the currently classified object. The ratio of the number of misclassified objects to a numerical force of the reference set estimates the above mentioned error rate used as the clustering quality criterion.

The low error rate value denotes that the considered classes (or clusters) differentiate easily, but high values of the error rate mean that the classes overlap. The leave one out method is very convenient in the case of classifiers based on the k-NN rule since no training is required. This property of the k-NN classifiers was intesively used for creating a fuzzy k-NN rule proposed by one of the authors of the present work [17] and for introducing the more sophisticated pair-wise k-NN classifier [18].

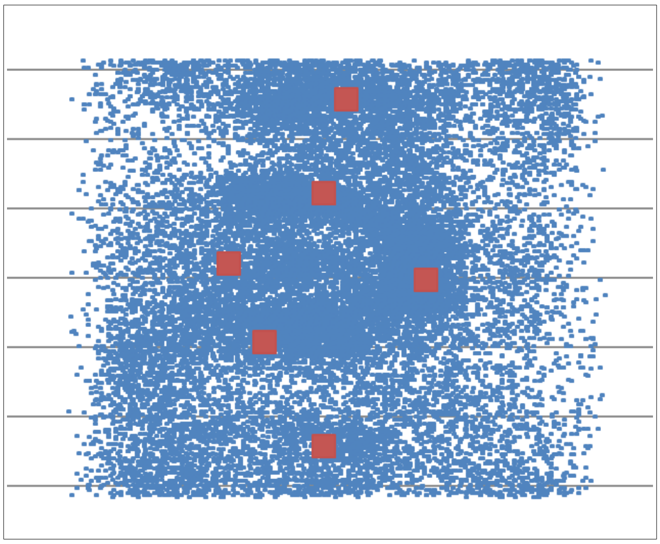


Fig. 3 Data distribution with six clusters centroids (along “horseshoe” shape).

TABLE I.EVALUATION OF CLUSTERIZATION BY K-NEAREST NEIGBOURS METHOD FOR EACH VISUALIZATION MAP BASED ON ERROR RATES.

| **Phase** | **Training** | | | **Testing** | | |
|---|---|---|---|---|---|---|
| | **1988** | **1998** | **2009** | **1988** | **1998** | **2009** |
| Authors' method | 0.0140 | 0.0115 | **0.0074** | 0.8890 | 0.8300 | **-** |
| VOS | 0.0119 | **0.0078** | 0.0110 | 0.9011 | **-** | 0.8710 |
| VxOrd | 0.0084 | 0.0107 | **0.0045** | 0.8707 | 0.7898 | **-** |
| SOM | 0.0143 | **0.0121** | 0.0128 | 0.8480 | | 0.8090 |

In every series of visualizations it is possible to point at the map with the clearest clustering structure – model map. The other, as it can be assumed, develops towards clustering pattern. If the algorithm for pattern (points) clustering of model map is trained, other ones can be tested by using the nearest neighbour method. Which one serves as a training set and which one as a testing set can be found first by evaluation of the standard k-NN classifier. The lower error rate means a better clustering structure (bold numbers in *Training* part of Table 1).

The results of training allow selecting an apropriate dataset (bold) for testing. During the testing phase (*Testing* part) the clusterization quality can be tracked on the basis of error rate changes. It is worth noting that comparison must be made along rows, not columns, because of the use of different methods to generate patterns.

The data in Table 1 show that the clusterization increases in one case – the first row, which characterizes the authors' approach. It proves continuous changes of CCS towards overlapping categories and reorganization needs.

### B. Even distribution, FD and Lacunarity

The authors' study of visualization maps relates to visual interface: which method of mapping can deliver the best way to explore a complex dataset of scientific articles?

Restrained homogenous distribution on a sphere surface can be estimated by the volume of empty places. The more holes in the pattern, the more heterogeneous it is. The appropriate parameter is *Lacunarity* - the degree of holes distribution having the lowest value for indeterminate structure. Lacunarity is often used in medical imaging for detection of structural changes in bone texture on radiographs [19].

Visualization maps can be considered as flat textures associated with the patterns of documents nodes distribution. Lacunarity $\lambda$ is defined as:

$$\lambda_{\epsilon,g} = CV^2_{\epsilon,g} = \left(\frac{\sigma}{\mu}\right)^2_{\epsilon,g} \qquad (2)$$

where $\sigma$ is the standard deviation and $\mu$ is the mean for pixel per box at this size $\varepsilon$, in a box at this orientation $g$. Lacunarity pertains to both gaps and heterogeneity. To simplify, the more *gappiness* in the image (i.e. sparsely occupied maps), the higher lacunarity. Some recent research has shown that there is a correlation between lacunarity and *fractal dimension*, FD [20],[21].

The FD is a complexity indicator with a non-integer value. The fractal dimension could be characterized as a scale of transition to homogeneity and is therefore very practical in the dynamics study case. Because the maps were generated by three different mapping algorithms they present distinct homogeneity, what is according to our assumption, one of the criteria of the ergonomic visual interface.

The values of lacunarity and the FD for every map are shown in Table 2. The highest value in each row (bold numbers in the first part of Table 2) indicates a map with the large porosity (gappiness). Bold FD values in the second part of the same table means the best formed structure (implicitly clear clusterization). The first row data presents a continuous growth of complexity degree with simultaneously dense occupation (high FD and lowest lacunarity values). Other (VOS, SOM and VxOrd) demonstrate oscillations are difficult to interpret. Consequently, the dynamics of each index across time for every method was evaluated. Lacunarities of every method should not be compared because of different spanning geometry.

If should be taken into consideration that the fractal dimension for random (or pseudo-random) distribution equals 2.77, the more does the FD tend to this value, the pattern is more homogeneous [22]. And inversely, the low FD (bold numbers in Table 2) means the distribution resembles linear. A stable structural change (in contrary of step change) in time is proved by the authors' method.

## V. DISCUSSION

According to authors' conception, in order to measure dynamics of graphical patterns we need to focus on how complexity evolves. Therefore clustering resolution has been tested by use of machine learning and pattern recognition algorithms.

TABLE II. LACUNARITY AND FD FOR EACH VISUALIZATION MAP.

| **Method** | **Lacunarity** | | | **Fractal Dimension** | | |
|---|---|---|---|---|---|---|
| | **1988** | **1998** | **2009** | **1988** | **1998** | **2009** |
| Authors' method | 0.0185 | 0.0155 | **0.0147** | 2.34 | 2.39 | **2.50** |
| VOS | **0.0035** | 0.0144 | 0.0067 | 2.23 | 2.15 | **2.40** |
| VxOrd | **0.0247** | 0.054 | 0.0421 | 2.18 | 2.15 | **2.23** |
| SOM | 0.3340 | **0.305** | 0.319 | 1.82 | 1.84 | **1.97** |

The authors proposed the qualitative measures for evaluation of structural changes of pattern: FD and lacunarity.

In general, we can conclude: the larger complexity degree, the lower randomness. On the other hand, the complex structure is also can be determined by the clustering level. The current findings confirmed by presented measures (Table 1, Table 2) show that clusters have become more explicit with time at the maps generated by authors' approach and at the same time tend to uniform distribution (lowest lacunarity and FD resembles the value of random distribution). The high lacunarity informs about dense network of holes among others, due to overlapping pattern. Next acceptable technique in the terms of changes continuity is SOM. Qualitative approach to compare visualization maps [10] shows the similar results: both MDS-sphere and SOM reveal consequence in dynamics changes, moreover VOS and VxOrd – inappropriate topology for data exploration [10, Appendix]

How easy users can play with data and analyze their change – show the ergonomic properties of visualization interface. Homogeneous occupation of visual layout, edgeless, continuity in changes should feature good visualization [23]. Several papers described particularly this application from the end-users-analysts point of view [10], [24]. Another practical aspect pertain relevant documents retrieval due to visual representation. This still remains the main direction of current study. After receiving good precision for a small sample, authors intend to repeat experiment with bigger dataset and all presented methods.

Recent research [20],[21] show a strong correlation between the FD and lacunarity. To check this, it is required to have a more representative dataset i.e. be multi-various. To find essential changes including paradigms, the period of analysis must be extended to three or four decades, i.e. until 2017. There basic technological problem has appeared: the ACM has changed the classification and applied it to the collection of 2013. The success to supplement the dataset depends on whether the ACM will standardize the old classification schemes according the new version and adapt it to the whole dataset.

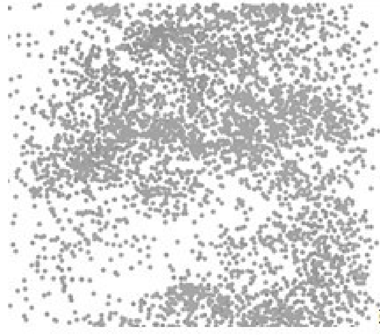
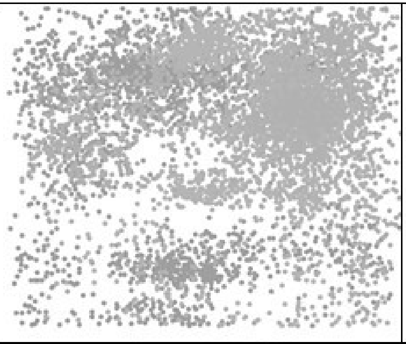
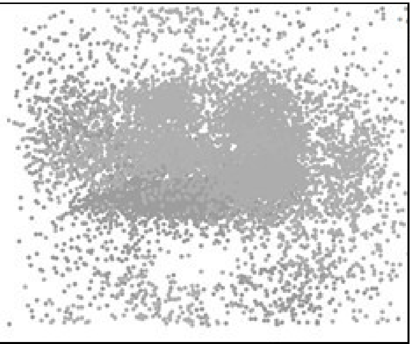
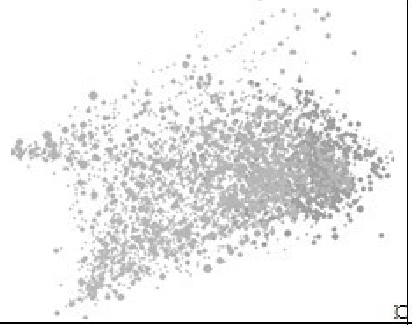
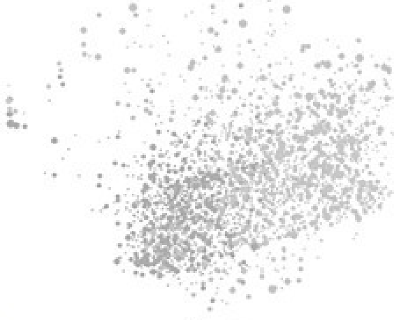
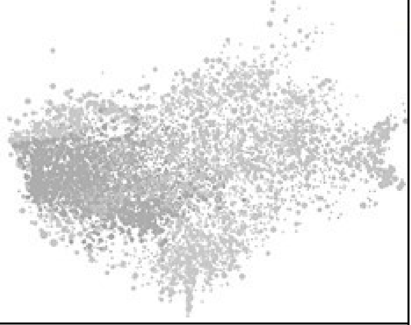
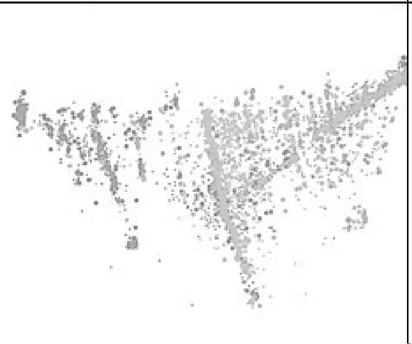
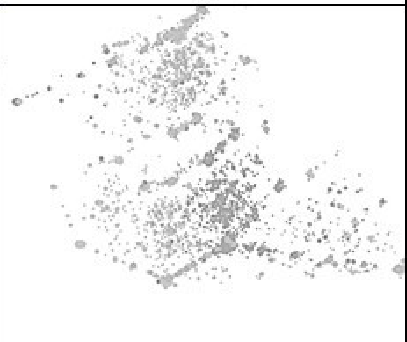
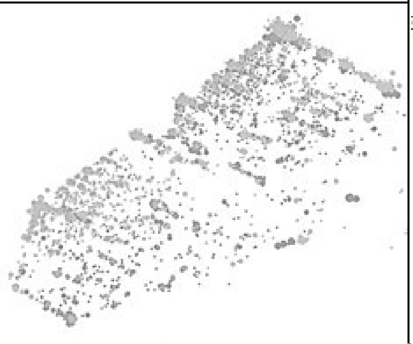
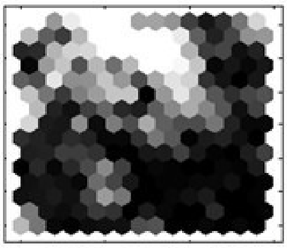
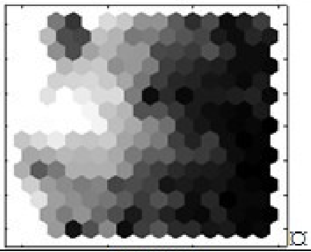
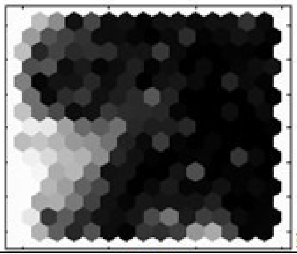
Undoubtedly, a sequential series of three maps is not enough to estimate knowledge evolution dynamics. A more multi-variety dataset to track all changes in fast growing knowledge is needed, but truly objective circumstances concerning data gathering were appeared. However, proposed measures can be considered if we need to select the best data distribution in the terms of interface functionality.

## VI. SUMMARY

Visual interface for analyzing how computer science evolved through the two last decades is briefly presented in current paper. This application includes an interactive 3D map of scientific articles organized by their semantic relationships. The authors proposed the conception how to quantitatively evaluate different visualization maps in respect of possibilities of dynamics analysis. They also characterize topological arrangement in the terms of navigation functionality.

Four methods of mapping including own approach of mapping (MDS with spherical topology) have been compared. Quantitative measures allowed selecting the most appropriate mapping way with an accurate reflection of the current changes of computer science. In the quantitative analysis authors tracked the changes of pattern clusterization over time. Clusterization degree they evaluate using machine learning and pattern recognition algorithms (Table 1). They adopted both lacunarity and the fractal dimension of visualization patterns to find the scale of randomness in dynamics (Table 2). Moreover the local density differences, clusterization, rarefaction, homogeneity, and gappiness were measured to show the most acceptable layout for analysis, perception and exploration processes. 3D MDS maps (authors' approach) and SOM have shown the better properties than VOS and VxOrd. These results have proved the findings and interpretations obtained from qualitative analysis [9]. Given maps have revealed essential changes in computer science literature during the time of the development of the CCS classification compared.

## APPENDIX

| | | | |
|---|---|---|---|
| 3D·MDS¤ |  |  |  |
| VOS¤ |  |  |  |
| VxOrd¤ |  |  |  |
| SOM¤ |  |  |  |
| ¤ | 1988¤ | 1998¤ | 2009¤ |

## References

[1] V. Osinska and P. Bala, “Classification Visualization across Mapping on a Sphere”, in: *New trends of multimedia and Network Information Systems.* Amsterdam: IOS Press, pp. 95-107, 2008. ISBN 978-1-58603-904-2.

[2] V. Osinska, P. Bala and M. Gawarkiewicz, “Information Retrieval across Information Visualization”. IEEE Xplore Digital Library: *Proceeedings of 2012 Federated Conference on Computer Science and Information (FedCSIS),* Wroclaw, 2012, pp. 233 – 239.

[3] K. W. Boyack, B. N. Wylie and G.S.Davidson. “Domain visualization using VxInsight for science and technology management”. *Journal of the American Society for Information Science and Technology,* 53(9): 764-774, 2002. doi: 10.1002/asi.10066.

[4] Ch. Chen, *Information Visualization. Beyond the Horizon.* 2nd ed. London: Springer, 2006, pp.143-170. ISBN: 978-1-84628-579-0.

[5] K. W. Boyack, R. Klavans and K. Börner, “Mapping the backbone of science”, *Scientometrics,* vol. 64(3): 351-374, 2005. doi: 10.1007/s11192-005-0255-6.

[6] N. J. Van Eck and L. Waltman, “VOS: a new method for visualizing similarities between objects”, in *Advances in Data Analysis: Proceedings of the 30th Annual Conference of the German Classification Society* (eds HJ Lenz, R Decker), London: Springer, pp. 299-306, 2007

[7] N. J. Van Eck and L. Waltman, “How to normalize cooccurrence data? An analysis of some well-known similarity measures”, *Journal of the American Society for Information Science and Technology,* 60(8): 1635-1651,2009. doi: 10.1002/asi.21075.

[8] N. J. Van Eck, L. Waltman, R. Dekker and J. Van den Berg, “A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS”. *Journal of the American Society for Information Science and Technology,* 61(12): 2405-2416, 2010.

[9] F. Moya-Anegón, V. Herrero-Solana and E. Jiménez-Contreras. “A connectionist and multivariate approach to science maps: the SOM, clustering and MDS applied to library and information science research”. *Journal of Information Science,* 32(1): 63-77, 2006. doi:10.1177/0165551506059226.

[10] V. Osinska and P. Bala, “Study of dynamics of structured knowledge: Qualitative analysis of different mapping approaches”, *Journal of Information Science,* 1-12, 2014. doi: 10.1177/0165551514559897.

[11] S. Fortunato, “Community detection in Graphs”, *Physics Reports,* 486: 75-174, 2010. doi: 10.1016/j.physrep.2009.11.002.

[12] C. Ware, *Information Visualization: Perception for Design.* CA: Morgan Kaufmann, pp. 11, 188, 273, 2004. ISBN 0123814642.

[13] J. D. Banfield and A.E. Raftery AE, “Model-based gaussian and non-gaussian clustering”, *Biometrics,* 49: 803-821, 1993. doi: 10.1093/biomet/63.3.413.

[14] C. Biernacki, G. Celeux and G. Govaert, “Assessing a mixture model for clustering with the integrated completed likelihood”, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22: 719-725, 2000.

[15] P. A. Devijver and J. Kittler, *Pattern recognition. A statistical approach,* London: Prentice Hall, 1982.

[16] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification,* New York: John Wiley & Sons, 2001.

[17] A. Jozwik, “A learning scheme for a fuzzy k-NN rule”, *Pattern Recognition Letters,* 1: 287-289, 1983. doi: 10.1016/0167-8655(83)90064-8.

[18] A. Jozwik, S. Serpico and F. Roli, “A parallel network of modified 1-NN and k-NN classifiers -application to remote-sensing image classification”, *Pattern Recognition Letters,* 19: 57-62, 1998.

[19] R. E. Plotnick, R. H. Gardner and R. W. O'Neill “Lacunarity indices as measures of landscape texture”, *Landscape Ecology,* 8(3): 201-211, 1993.

[20] A. Forsythe et al., “Predicting Beaty: Fractal dimension and Visual complexity in art”, *British Journal of Psychology,* 102, 49-70,2011. T. G. Smith, G. D. Lange and W.B.Marks, “Fractal Methods and Results in Cellular Morphology”, *Journal of Neuroscience Methods,* 69: 1123-126, 1996. doi: 10.1016/S0165-0270(96)00080-5.

[21] V. Osinska, “Fractal analysis of Knowledge Organization in Digital Library”, in Katsirikou A, Skiadas CH (eds) New Trends in *Qualitative and Quantitative Methods in Libraries,* Singapore: World Scientific Publishing, pp. 17-23, 2011.

[22] W. A. Pike et al., “The Science of Interaction”, *Information Visualization,* vol. 8, 4: pp. 263-274, 2009.

[23] V. Osinska, J. Dreszer-Drogorob, G. Osinski and M. Gawarkiewicz “Cognitive Approach in Classification Visualization. End-Users Study”, in *Classification & Visualization: interfaces to knowledge* (ed A. Slavic et al), Hague, Holland, 23 -25 October 2013, Würzburg: Ergon Verlag, pp. 273-283. ISBN 978-3-95650-007-7.

DOI: 10.15439/2015F186
ACSIS, Vol. 5

# Automatic Summarization of Polish News Articles by Sentence Selection

Krzysztof Jassem, Łukasz Pawluczuk
Adam Mickiewicz University
in Poznań
ul.Wieniawskiego 1, 61-712 Poznań, Poland
Email: jassem@amu.edu.pl, lp44246@st.amu.edu.pl

***Abstract*—This paper describes the automatic summarization system developed for the Polish language. The system implements sentence-based extractive summarization technique, which consists in determining most important sentences in document due to their computed salience. A structure of the system is presented, as well as the evaluation method and achieved results. The presented attempt is intended to serve as the baseline for future solutions, as it is the first summarization project evaluated against the Polish Summaries Corpus, the standardized corpus of summaries for the Polish language.**

## I. Introduction

AUTOMATIC text summarization is a very active research field in recent years. Its purpose is to reduce a text document, by extracting its most important parts in order to create more condensed, but still human-readable form, known as summary. The task consists in the creation of an appropriate computer application and a framework for testing and evaluation.

In this paper we focus on sentence-based extractive summarization using machine learning. We implement well-known techniques, improved and merged into a single summarizing system. The system uses a list of features applied in previous projects, supplemented by new ones, introduced by the paper's authors. Polish Summaries Corpus, a resource created by Ogrodniczuk and Kopeć in [1] has been used as the dataset for training machine learning algorithms. No one has ever used this corpus to create a summarizing system before.

Moreover, an evaluation method has been developed. It is based on the ROUGE summarization evaluation package introduced at Document Understanding Conference (DUC) in 2004, by Chin-Yew Lin [2], who proved it to be a correct measure for the task. We propose to use this evaluation method in future automatic summarization solutions for the Polish language for the sake of objective comparison. The present solution could then serve as the baseline for new systems.

The paper is organized as follows: the rest of the current section describes briefly the aim of the summarization task and main methods in the field. Section 2 provides a review of already existing summarization systems for the Polish language. In section 3 Polish Summaries Corpus is described in detail. Section 4 outlines our solution, it's overall framework, as well as the employed set of features. Section 5 introduces the evaluation methodology and presents our experiments and their results. Eventually, section 6 contains some conclusions and the outline for future work.

### A. Aim of summarization

Modern digital technologies, including World Wide Web, result in information excess. Everyday brings vast amount of new on-line information of various type. Processing this continuously growing information databases is not possible by a single human. Automatic summarization is an attempt to confront information processing needs. It is based on the assumption that a computer system can read all data quickly and present its condensed from. Summarization is useful in medicine, law or scientific areas, as well as in everyday life.

Formally, in the area of text summarization, "summary can be defined as a text that is produced from one or more texts, that contains a significant portion of information in the original text(s), and that is no longer than half of the original text(s)" [3].

### B. Methods of summarization

Automatic text summarization may be classified according to program's input or output. As regards input, summarization may concern one document or multiple documents (multi-document summarization). Further, in case of multi-document summarization, input data may be *monolingual* or *multilingual*. As regards output, one may distinguish *extracts* and *abstracts*. Mani (2001) claims that "an extract is a summary consisting entirely of material copied from the input" (which in fact can be paragraphs, sentences, phrases, terms or even single nouns) and "abstract is a summary at least some of whose material is not present in the input". Extractive summaries are obviously easier to obtain. Moreover, summaries may be *indicative* or *informative*, which means they can indicate source text's topics and give a brief idea of what the original text is about, or cover the topics in the source text, respectively [3]. Finally, *generic* and *user-focused* (a.k.a. *query-driven*) summaries may be distinguished. *Generic* summaries try to cover all relevant information from the source text, while *user-focused* ones respond to user's information needs expressed as topic or query [3].

Literature often considers automatic summarization a three-stage process. Lloret (2006) names the following steps of the process:

978-83-60810-66-8/$25.00©2015, IEEE

- *interpretation* of the source text in order to obtain a text representation,
- *transformation* of the text representation into a summary representation,
- *generation* of the summary text from the summary representation.

Methods of text summarization may differ as far as the level of processing is concerned: *surface*, *entity*, or *discourse* levels [4]. It is worth noting that there exist systems, which adopt hybrid-approaches.

*Surface-level* approaches make use of shallow features to analyze information included in a text document. Usually, these features are combined together into a salience function used to extract information. Examples of such features are:

- Thematic features — based on term frequency analysis and statistically salient terms,
- Location features — based on position in text, paragraph or section depth,
- Background features — based on presence of title or headings terms, or a user's query,
- Cue words and phrases — based on presence of special 'bonus' or 'stigma' terms.

*Entity-level* approaches are based on the internal representation of text. They model text entities and their relationships across a document. Examples of such relationships between entities are:

- Similarity — e.g. vocabulary overlap,
- Proximity — distance between text units,
- Co-occurrence — words occurring in common contexts,
- Thesaural relationship among words — e.g. synonymy, hypernymy,
- Coreference — e.g. anaphora, cataphora, noun phrases,
- Logical relations — e.g. agreement, contradiction, entailment, consistency,
- Syntactic relations — e.g. relations based on parse trees,
- Meaning representation-based relations — e.g. predicate-argument relations.

*Discourse-level* approaches model the global structure of text, and its relation to communicative goals. Examples of such structures are:

- Format of the document,
- Threads or topics as they are revealed in the text,
- Rhetorical structure of the text.

## II. Review of experiments on summarization of Polish Texts

This section covers experiments on automatic summarization for the Polish language, resulting in theoretical works, as well as working implementations. All of them apply extractive methods of summarization.

### A. PolSum2 (S. Kulikow)

The first attempt on automatic text summarization for the Polish language was made by Ciura et al. [5] and resulted in the *PolSum* system, which then evolved into *PolSum2*. The system is still available at http://las.aei.polsl.pl/PolSum/. *PolSum2* is an extractive system. It performs various kinds of text analysis (morphological, syntactic, semantic) in order to extract most important sentences from an input document. The system also recognizes anaphora, which results in better coherence between selected sentences.

*PolSum2* performs in three stages of summarizing[5]. The first stage, called 'Calling remote analyzer' is intended to call the remote server, which performs text analysis. The *Linguistic Analysis Server* (LAS) is used for this purpose. This tool, created by the same authors, performs linguistic analysis on the levels of: morphological, syntactic and semantic analysis. The syntactic analysis builds a parse tree on the basis of Syntactic Group Grammar for Polish (SGGP) [6]. The system also performs the analysis of anaphoric relations. The seconds stage of summarization process is 'Selecting the essential sentences'. There is no concrete information on the criteria for sentence weighting. The last stage is called 'linearization'. It is designed to create coherent output. Proper forms of words are generated and placed in proper places in sentence. The system also performs homonyms reduction and anaphora substitution for better result reading.

The papers that describe the system do not provide any information about evaluation results.

### B. Lakon (A. Dudczak)

Adam Dudczak's *Lakon* is another automatic text summarization system created for the Polish language [7]. It is available on-line at http://www.cs.put.poznan.pl/dweiss/research/lakon/. The system was developed as a result of author's Master Thesis, whose one of main goals was to compare effectiveness of some popular extractive methods for the Polish language. Three methods were developed. They were based on the following heuristics:

- $tf \times idf$ and *Bm25 Okapi* — assumes that words occurrence frequency determines sentence's salience
- sentence's position in text — assumes that most important sentences are often at the beginning of paragraphs,
- lexical chain — assumes that relations across sentences determine their salience.

The system was evaluated on the corpus created from 10 manually summarized newspaper articles. 60 volunteers manually created totally 285 summaries of these articles. Evaluation results indicated that the most effective features were words occurrence frequency and sentence's position. The lexical chains method was proved to be worse than the others.

### C. Summarizer (J. Świetlicka)

*Świetlicka's Summarizer* [8] is the latest tool created for Polish. It is available on-line at http://clip.ipipan.waw.pl/Summarizer. This solution is the most similar to the one proposed here. It uses various machine learning methods for training an extractive summarizer based on a set of sentence's features. These features include:

- LLR — *Log Likelihood Ratio*,
- $tf \times idf$,

- Sentence's centrality,
- Occurrence of characteristics phrases — bonus and stigma words, popularity of one or two first words of a sentence,
- Similarity to the title — indicating occurrence of words from the title in a sentence,
- Number of words starting with uppercase — indicating Named Entities,
- Number of tokens that are not proper words — i.e. punctuation or numbers,
- Localization — position of the sentence in paragraphs and position of the sentence in the whole text,
- Length of sentence,
- Length of paragraph,
- Length of text,
- Type of sentence — based on the last token: declarative, interrogative, imperative.

A number of tests were performed on different subsets of these features. The author used about 13 different machine learning algorithms in order to compare their effectiveness. The corpus was created by the author on his own and contains 102 newspaper articles for training and 67 articles for evaluation.

*Świetlicka's Summarizer* also performs simple summary linearization. It consists of three steps. At first, sentences are sorted in the order of their appearance in the document. Secondly, fragments in parentheses are removed in order to make sentences shorter. Lastly, some special words are removed from the beginnings of sentences, such as: therefore, moreover or however.

The discussed work contained the following conclusions:

- localization-based features, particularly sentence position in the paragraph and the whole document, tend to be the most important ones,
- sentence centrality feature is also very effective,
- cue words feature are not so effective,
- machine learning algorithms tend to be an effective solution for automatic summarization. Using a set of features result in better quality than using each separate feature.

## III. Polish Summaries Corpus

Polish Summaries Corpus is a resource created by Ogrodniczuk and Kopeć in 2014 [1]. Its aim is to provide a high quality corpus containing manual summarization examples. The corpus forms a significant facilitation for further researchers, who can build their own summarization tools based on this corpus, as well as evaluate them. Ogrodniczuk and Kopeć notice that previous works on automatic summarization in the Polish language lacked a common corpus and a common evaluation method, therefore their results are not comparable. *Rzeczpospolita corpus* — a collection of articles from the Web archive of a Polish newspaper [9] was used as the base corpus for Polish Summaries Corpus.

Polish Summaries Corpus contains 569 text documents divided into 7 categories: Society and Politics, Sport, Economy, Culture news, Law, National news and Science and Technology. All these texts have been manually summarized by independent annotators. All 569 documents have the extractive summaries and 154 have also the abstractive summaries. For each document in the corpus 5 independent propositions of summarization have been created. Each proposition of summarization contains 3 summaries of a given text of the approximate length of 5%, 10%, 15% of the original, respectively. The summaries are included in one another: 10% summary contains only fragments from previously selected 20% summary and so on. Therefore, the corpus size is 8355 summaries.

## IV. The proposed solution

The solution presented here implements sentence-based extractive summarization. It consists of two main components: linguistic analysis and summarization application. The latter component selects essential sentences and generates the result summary. The summarization component appplies neural networks as a machine learning algorithm. The Open Source implementation — PyBrain[10] was used.

### A. Methodology

The linguistic analysis component performs various kinds of text analysis. The input document is divided into paragraphs, sentences and tokens. Subsequently, lemmatization is performed, parts of speech, named entities and headers are determined. Finally, the internal document model is created and transferred to the summarization component.

The summarization component works as a three-stage process. The first stage computes feature values for each sentence in the document. The second stage is sentence weighting based on the previously trained machine learning model and computed features. In the third stage, the summary is prepared according to the obtained sentences weights. This stage includes the sorting of result sentences, according to their order in the original document.

### B. Description of features

This section describes each feature used in the system. Selection of the features was based on literature [3], [4], [11], [7], [8] as well as a few new ideas. The complete list of used features includes:

- *TfIdf* — sum of *term frequency – inverse document frequency* value for every word in sentence,
- *Centrality* — arithmetic average of sentences similarity to every other sentence in the document. Cosine similarity is used as a similarity measure between two sentences,
- *SentLocPara* — position of a sentence in the paragraph: in the first, second or third of equal parts,
- *ParaLocSection* — position of the paragraph in the document: in the first, second or third of equal parts,
- *SentSpecialSection* — occurrence in a special section like the beginning (introduction) or ending (conclusion) of document,

$$\text{ROUGE-N} = \frac{\sum\limits_{S\in\{ReferenceSummaries\}} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S\in\{ReferenceSummaries\}} \sum\limits_{gram_n \in S} Count(gram_n)} \tag{1}$$

- *SentInHighestTitle* — number of words from heading or title in the sentence,
- *ParaLength* — paragraph length: short (up to 1 sentence), average (2–5 sentences) or long (more than 5 sentences),
- *SentLength* — sentence length: short (up to 7 words), average (7–14 words) or long (more than 14 words),
- *SentType* — type of the sentences, based on its last punctuation mark: declarative, interrogative or imperative.
- *MetaInfo* — sentences not referring to the document content, i.e.: an information about document's author or photo signatures,
- *AvWordLength* — the average of words lengths in sentences,
- *Verb* — existence of the final verb,
- *Nouns* — number of nouns in sentence,
- *Pronouns* — number of pronouns in sentence,
- *SentInHighestPname* — number of Named Entities in the sentence as found by a naive method, recognizing Named Entity as a word starting with capital letter,
- *NER* — number of Named Entities in the sentence as found by NERf Named Entities Recognition tool [12],
- *NERTf* — sum of every Named Entity frequency in the whole document, occurring in given sentence,
- *PersNameNE* — number of recognized NE of the "person" type,
- *OrgNameNE* — number of recognized NE of the "organization" type,
- *PlaceNameNE* — number of recognized NE of the "place" type,
- *DateNE* — number of recognized NE of the "date" type,
- *GeogNameNE* — number of recognized NE of the "geography" type,
- *TimeNE* — number of recognized NE of the "time" type.

The features applied by authors of this paper, which were not mentioned in the referred works, are: *MetaInfo*, *AvWordLength*, *Verb*, *Nouns*, *Pronouns*, *NER*, *NERTf*, *PersNameNE*, *OrgNameNE*, *PlaceNameNE*, *DateNE*, *GeogNameNE* and *TimeNE*.

## V. Evaluation

### A. Evaluation method ROUGE (DUC conference)

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation [2]. It was introduced by Chin-Yew Lin at Document Understanding Conference (DUC) in 2004 and since then it has became the standard method for the evaluation of automatic summarization systems. It provides a set of measures to automatically determine the quality of summary in comparison to ideal summaries created by humans. The measures are based on overlapping units such as n-grams, word sequences and word pairs. ROUGE has been proved to be highly correlated with human judgements. This section describes ROUGE-N methods, which were proved to work well in single document summarization tasks.

ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries [2].

It is computed using the (1) formula, where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. It is worth noting that the denominator of (1) increases if more than one reference documents are used. Moreover, larger weight is assigned to matching n-grams occurring in multiple references, so if words are shared by more references, ROUGE-N favors them.

### B. Experiments and Results

A number of experiments were performed. Different subsets of features were used in order to achieve the best results in summarization. Every learned model for each features susbset was evaluated with ROUGE-1, ROUGE-2 and ROUGE-3 methods. Random summarization was used as a baseline. Features were divided into subsets, as follows:

- *Sub1* $\in \{TfIdf, ParaLength, Centrality, SentType, SentSpecialSection, SentInHighestTitle, SentLength, SentLocPara, ParaLocSection\}$
- *Sub2* $\in$ *Sub1* $\cup \{Pronouns, MetaInfo, Verb, Nouns, AvWordLength\}$
- *Ner1* $\in \{SentInHighestPname\}$
- *Ner2* $\in \{NER, NERTf\}$
- *Ner3* $\in \{OrgNameNe, GeogNameNe, DateNe, PlaceNameNe, PersNameNe, TimeNe\}$

Evaluation results are placed in Table I. It is clear that almost every subset of features used in experiments gave nearly the same results, which were about 15% better than the baseline, according to the F-1 score. No feature subset performed clearly better than the others. New features, included in *Sub2* raised the score slightly, just as dividing the Named Entities information into categories did (*Ner3*). In fact, using the NER tool, instead of naive methods tends to give slightly better results in summarization. Summing up, the best results in ROUGE-1, ROUGE-2 and ROUGE-3 were achieved using the largest subset of features. The experiments have shown that developing new features may be quite useful, but there is no single feature that separately raises the score significantly.

## VI. Conclusion and Future Work

In this article, we have presented the document summarizing approach for the Polish language. It is based on sentence extraction and applies neural networks as a machine learning

TABLE I
EXPERIMENTS' RESULTS.

| | ROUGE-1 | | | ROUGE-2 | | | ROUGE-3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Recall | Precision | F-1 | Recall | Precision | F-1 | Recall | Precision | F-1 |
| RANDOM | 0.29 | 0.34 | 0.30 | 0.14 | 0.17 | 0.15 | 0.13 | 0.15 | 0.13 |
| $Sub1$ | 0.45 | **0.45** | 0.44 | 0.34 | **0.33** | 0.33 | 0.33 | **0.32** | 0.32 |
| $Sub2$ | 0.47 | **0.45** | 0.45 | 0.35 | **0.33** | 0.33 | 0.35 | **0.32** | 0.33 |
| $Sub1 \cup Ner1$ | 0.44 | 0.42 | 0.43 | 0.32 | 0.29 | 0.30 | 0.31 | 0.28 | 0.29 |
| $Sub1 \cup Ner2$ | 0.49 | 0.43 | 0.45 | 0.37 | 0.31 | 0.34 | 0.37 | 0.31 | 0.33 |
| $Sub1 \cup Ner3$ | 0.49 | 0.42 | 0.45 | 0.37 | 0.30 | 0.33 | 0.36 | 0.29 | 0.32 |
| $Sub2 \cup Ner1$ | 0.48 | 0.42 | **0.46** | 0.37 | 0.32 | 0.34 | 0.36 | 0.31 | 0.33 |
| $Sub2 \cup Ner2$ | 0.47 | 0.42 | 0.44 | 0.35 | 0.3 | 0.32 | 0.34 | 0.29 | 0.31 |
| $Sub2 \cup Ner3$ | **0.51** | 0.43 | **0.46** | **0.39** | 0.32 | **0.35** | **0.39** | 0.31 | **0.34** |

algorithm. This approach seems to be promising in achieving acceptable summarizing method for the Polish language, however there are some difficulties in choosing the proper features set and tuning machine learning algorithm. We conclude that there is still much work to do in the field. We hope that our approach will serve as a inspiration, as well as a baseline for the future research at the task of automatic summarization for Polish.

## REFERENCES

[1] M. Ogrodniczuk and M. Kopeć, "The polish summaries corpus," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014. ISBN 978-2-9517408-8-4

[2] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL workshop on Text Summarization Branches Out*, 2004, p. 10.

[3] E. Lloret, "Text summarization: An overview," [on-line] http://www.dlsi.ua.es/~elloret/publications/TextSummarization.pdf.

[4] I. Mani and M. Maybury, *Advances in Automatic Text Summarization*. MIT Press, 1999. ISBN 9780262133593

[5] M. Ciura, D. Grund, S. Kulików, and N. Suszczanska, "A system to adapt techniques of text summarizing to polish," in *International Conference on Computational Intelligence, ICCI 2004, December 17-19, 2004, Istanbul, Turkey, Proceedings*, A. Okatan, Ed. International Computational Intelligence Society, 2004. ISBN 975-98458-1-4 pp. 117–120.

[6] N. Suszczańska and M. Lubiński, "Polmorph, polish language morphological analysis tool," in *19th IASTED International Conference APPLIED INFORMATICS - AI'2001, Innsbruck (Austria)*, 2001, pp. 84–89.

[7] A. Dudczak, J. Stefanowski, and D. Weiss, "Automatyczna selekcja zdań dla tekstów prasowych w języku polskim," Institute of Computing Science, Poznan University of Technology, Poland, Technical Report RA-03/08, 2008.

[8] J. Świetlicka, "Metody maszynowego uczenia w automatycznym streszczaniu tekstów," Master's thesis, University of Warsaw, 2010.

[9] D. Weiss, "Korpus rzeczpospolitej," [on-line] http://www.cs.put.poznan.pl/dweiss/rzeczpospolita.

[10] T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. Rückstieß, and J. Schmidhuber, "PyBrain," *Journal of Machine Learning Research*, 2010.

[11] I. Mani, *Automatic Summarization*, ser. Natural language processing. J. Benjamins Publishing Company, 2001. ISBN 9789027249869

[12] J. Waszczuk, K. Głowińska, A. Savary, and A. Przepiórkowski, "Tools and methodologies for annotating syntax and named entities in the National Corpus of Polish," in *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT 2010): Computational Linguistics – Applications (CLA'10)*. Wisła, Poland: PTI, 2010, pp. 531–539.

DOI: 10.15439/2015F142
ACSIS, Vol. 5

# Approach to Building a Web-based Expert System Interface and Its Application for Software Provisioning in Clouds

Evgeny Pyshkin
Institute of Computing and Control
Peter the Great St. Petersburg Polytechnic University
St. Petersburg, Russia, 195251
Email: pyshkin@icc.spbstu.ru

Andrey Kuznetsov
St. Petersburg Software Center
Motorola Solutions Inc.
St. Petersburg, Russia, 192019
Email: andrei.kuznetsov@motorolasolutions.com

***Abstract*—This paper focuses on a generalized approach to providing user interface to a web-based expert system (WBES). We examine MVC and MVP design patterns used traditionally to construct a web application user interface. In order to leverage the strength of the MVC/MVP design patterns we propose a special ontology representing a user communication domain. We describe a self-service networked infrastructure for automatic deployment of command line interface (CLI) applications. We demonstrate how to apply the proposed ontology for the design of a WBES aimed at supporting client software re-execution in clouds. In particular, we address the problems existing in the area of software development for music information retrieval algorithms implementation.**

## I. Introduction

Pervasive nature of modern software is a popular subject of the present-day technology discourse, whether the question concerns computer-assisted education, interface design, usability or elementary forms of programming, which are one of necessary elements of modern information literacy. In particular, service-oriented software and cloud technology significantly transformed the way we use computing and storage capabilities.

There is a constant interest to organizing processes of research software distribution in order to make computational and data resources available for other users. For example, in the domain of information retrieval (IR) many developed approaches are semantic relatedness centered. The focus of such works is on developing algorithms for better semantic relatedness evaluation, semantic classification or clusterization. An obvious way to evaluate an IR algorithm is to use various test collections, while an algorithm itself might be implemented in the form of a computer program. However, those programs often remain unpublished. In some IR domains, particularly, in music information retrieval (MIR), even if a software implementation is reusable, test collections might not be available for a third party researcher either for the reason of their big size or due to the copyright restrictions. That's one of the reasons explaining difficulties of comparing or reproducing results achieved by other researchers. From the study [1] we know some statistics of what MIR researchers are as software developers. The figures are rather discouraging: 82% of researchers did develop software, but only 39% of those took steps to achieve better reproducibility. It is no wonder that only 35% of those developers published any code, whereas 51% said their code had never left their own computer. Often the only way to follow is to believe in the results reported in papers without any possibility to be sure that the reported results came from a research method, and not from bugs in the software. Thus, researchers often re-implement algorithms based on the published descriptions; such re-implementations are not often executed in the same context as it was for the original software.

In [2] the best practices are examined, which are aimed to improve research reproducibility. Among them there are such ones as using collaborative platforms (like *Github*) and resource sharing mechanisms in order to lower the barrier to reproduce the third party work. In [3] a platform for re-executing software in the same context is described. That solution is based on capturing files and environments required for an experiment, and building an archive with subsequent software re-execution on a third party machine.

In [4] the authors took the significant step toward research reusability and reproducibility in the domain of machine learning. They developed and maintain a networked system (OpenML) for sharing and organizing data sets and algorithms solving the typical machine learning tasks. In our work we pay attention to two broader aspects of the reproducibility problem. Firstly, it might not be permitted to distribute the source code, the binary files or the datasets due to the legal issues. Secondly, due to manifold existing supporting tools (such as version control systems, build systems or runtime environments) it is not easy to configure a local environment in a way to be capable to run a third party software not limited by the specific data and task types.

A possible solution addressing both mentioned issues is to distribute both software and datasets as services accessible via a standard client (e.g. a web browser). In such a case neither data copying, nor environment configuration is required. However, distributing algorithms and datasets as services is far

978-83-60810-66-8/$25.00©2015, IEEE

from being a trivial problem. There is a certain similarity with Milne and Witten's consideration on data mining. Researchers who want to use Wikipedia as a knowledge source have two major options: either to base their work on secondary structures, or to build their own algorithms from scratch [5]. There are difficulties to share the algorithms due to lack of supporting platforms. We think that the similar situation exists in the domain of MIR.

Clouds can serve as a platform to share algorithms as services. One significant problem is a relatively high barrier to entry for non-experts. In our earlier work we described sources, advantages and problems of deploying research software in clouds and proposed an architecture of a provisioning service for automatic CLI applications deployment in computing clouds [6]. The proposed solution targets problems of build and run error discovery and handling, with special emphasis on errors conditioned by possible misconfiguration of a virtual platform where client software modules have to be executed.

The remaining text is organized as follows. In section II-A we discuss the process of CLI software deployment and provide a brief survey of existing approaches in order to explain their limitations and constraints (conditioning difficulties of provisioning applications as services). For the reason that we consider using expert systems as one of possible ways to overcome such difficulties, in section II-B we pay attention to a web based expert system (WBES) user interfaces. In section II-C we analyze a state-of-the art example of a web based expert system and evaluate its architecture with reuse and change scenarios. We discover that the modification of a model requires changes in all the model-view-presenter (MVP) architecture components, which, in a sense, goes against the motivation to use MVC/MVP pattern in design. In section III we examine two approaches to define a model interface: a domain *aware* approach and a domain *agnostic* approach. In order to better separate a presenter from a model, we propose to complement a subject domain ontology with a user communication ontology. For the problem of provisioning software in clouds, we define an ontology of software provisioning partially described in section III-A as well as a user communication ontology (section III-B). In section IV we describe a software provisioning self-service networked infrastructure, its architecture and its major components. In section V we demonstrate how the proposed approach helps in developing a web-based expert system for CLI applications deployment in computing clouds. We list some experiments we arranged in order to evaluate the knowledge based approach of using introduced networked infrastructure for provisioning a series of projects developed in the domain of MIR. We compare the knowledge based approach with the other existing implementations (section V-A) and describe a scenario based architecture evaluation process (section V-B).

## II. Related Work

In order to position our work within the framework of the service distribution domain we have to examine three major issues. First, we attempt to have a look at existing systems for deploying CLI applications in clouds with respect to the user expertise necessary to use such systems properly. Second, we analyze existing works on expert system user interface development with special attention paid to WBESs. Third, we analyze an emerging problem of providing a web-based user interface of an expert system to end users: specifically, how to apply MVC or MVP design patterns (widely used in web application architectures) for a WBES.

### A. Deploying a CLI Application in a Cloud

Research software (especially in the MIR domain) is often developed as desktop applications which primarily were not intended to be executed in networked or distributed environments. This aspect causes difficulties of their deployment in clouds without significant changes in software code. For example, an *OpenShift PaaS*[1] provides two ways to deploy an application in a cloud. The first way is to develop a custom cartridge, while the second one is to develop a module for an existing cartridge (e.g. *JavaEE* cartridge, *Python* cartridge, *Ruby-on-Rails* cartridge, etc.). Unfortunately, both approaches seem to be unsuitable for deploying CLI applications having no any networking capabilities. In order to support networking features without modification of an existing application, a proxy component is required [7]. The reality is that a deployer must be provided with the exact configuration describing a runtime environment. If the configuration is not valid (for example, an incorrect *Python* version is selected) users get unrecoverable deployment errors. In order to recover deployment errors automatically we could use such approaches as *AutoBash* [8].

In our earlier work we described how to extend the *AutoBash* approach by applying knowledge engineering formalisms [9]. We designed a proxy architecture which, in turn, is an enhanced *MEDEA*[2] proxy where a knowledge base is leveraged to control deployment and execution processes. In fact, the system described in the following sections of this paper can be considered as a web based expert system helping users to deploy a CLI application both in a cloud and within a desktop environment. Thus, the idea is to make deployment and invocation process as easy as uploading applications and datasets via web forms. Specifically, in the MIR domain the main algorithm experimentation scenarios are the following:

- A researcher wishes to test his/her algorithm by using one of the standard test collections, which might not be publicly available;
- A researcher wishes to compare the algorithm to other algorithms by using the same corpus;
- A researcher wishes to test a third party algorithm by using his/her own test corpus;
- A researcher wishes to test the corpus by running third party algorithms.

In the above mentioned work [9] we analyzed two popular software platforms facilitating the above mentioned scenarios:

[1] http://openshift.redhat.com

[2] MEDEA – Message, Enqueue, Dequeue, Execute, Access

one used in MIR, while the second used as a software deployment infrastructure.

*NEMA*[3] [10] provides access to the MIR software and data sets through the Internet. The environment was developed in 2008–2010, just about the period when the very first cloud service commercial implementations appeared. In fact, the *NEMA* provides a platform as a service (PaaS) and allows provisioning client software as a service (SaaS). The NEMA uses a set of preconfigured virtual machine images; each image provides a platform (e.g. *Python*, *Java*, etc.), which is completely configured to be used by the *NEMA* flow service. Storing a large set of custom images is expensive, moreover, any image modification has to be done manually; that's why using *NEMA* in public clouds is not easy.

The *MEDEA*[4] [7] infrastructure enables the deployment of an arbitrary CLI application on an arbitrary cloud platform. The *MEDEA* uses standard virtual machine images provided by a cloud and uploads a special wrapper (a task worker, in *MEDEA* terms) to the running virtual machine. The wrapper initializes the respective execution environment (*Python* or *Java*, for example) and then executes a client application. If we consider *MEDEA* as a self-service platform (within the context of MIR), there are two issues to be observed. Firstly, MIR applications often have dependencies on third party components and libraries, therefore, a wrapper might not initialize the environment properly. Sometimes researchers are unable to upload these libraries due to certain license restrictions; sometimes they don't know how to create an application package containing all the required dependencies. Secondly, there are MIR test collections which are not publicly available, so the code executed against those collections shall be considered to be "unsafe", and shall be executed in managed way in order to avoid dataset leaking. Let us mention that within the framework of the proposed architecture we address both issues by introducing a deployment manager that includes a wrapper component as it is described in the following sections.

The special case is *MIREX*[5], which is not a platform but an organization providing a service for testing algorithms delivered by its creators. In addition to the published test collections, some "secret" test collections are also used, and the evaluation process is arranged in the form of an annual contest. Thus, researchers have to wait for the results till the next competition, hence, this is not a way for everyday use.

### B. Expert System User Interfaces

As mentioned in [11], "Expert systems is a branch of Artificial Intelligence that makes extensive use of specialized knowledge to solve problems at the level of a human expert." As a result of Internet evolution and telecommunication tools development a new type of expert systems appeared: Web Based Expert Systems (WBES)[12]. In our work we pay attention to one important aspect of WBES design, i.e. user communication with WBES.

The focus of the majority of current works on expert system user interfaces is on the functional requirements (e.g. on capabilities the UI does provide and the reasons for them). Examples of such requirements are listed in [11]. We also found research works focused on a particular expert system development containing some screenshots of an expert system user interface and a discussion on the system architecture (see [13] for instance). Fewer works are focused on WBES development process [14]. We only found few works discussing a level of workload separation between a server and a client of an expert system [15]. The same is true for architectural issues of providing a user interface for a generic expert system [16], [17].

In [18] the authors realized that existing approaches to evaluate an expert system are connected mostly with the rules evaluation, paying less attention to user interaction issues. Some researchers complain about "a lack of a general methodology for developing web-based expert systems" [12] and notice that "web sites that enclose an expert system have been developing ad hoc and their developers do not follow any systematic method or process" [19].

### C. A Case Study: Web Based Expert Systems and Design Patterns

In order to better understand WBES user communication issues, we studied one of the rare state-of-the-art examples of the WBES where there is a discussion on WBES architecture and WBES-related design patterns.

For an end user, a WBES is presented as a web application. As we can see from many works (see [20], [21] for example) a common way to implement web applications is to use a model-view-controller (MVC) pattern [22] or its generalization known as a model-view-presenter (MVP) pattern [23]. However, using MVP might present a problem if we evaluate an architecture by using any scenario-based method (for example, *SAAM* [24]).

Let us think, for instance, of an MVC-based solution for an automatic price negotiation proposed in [25]. The authors suggest we use a production knowledge base with an inference engine as a *Model*, while the generated HTML pages are *Views* and a mediator component is a *Controller* (see Figure 1). As it is well known, if we follow an MVC pattern we expect to have such an advantage that each of the three structural components (e.g. a *Model*, a *View* and a *Controller*) can be modified independently. It allows to improve such software quality properties as reusability, modifiability and reduce a ripple effect appearing if one of the components changes significantly [23]. Let's evaluate this statement with the following scenarios:

1) *Reusability*: we can change an expert system domain from price negotiation to CLI application execution. If we consider this rather substantial change, we have to modify the *Model* (despite keeping the inference engine,

[3] NEMA – Networked Environment for Music Analysis: www.music-ir.org/?q-nema/overview

[4] MEDEA – Message, Enqueue, Dequeue, Execute, Access

[5] MIREX – Music Information Retrieval Evaluation eXchange: http://www.music-ir.org/mirex/

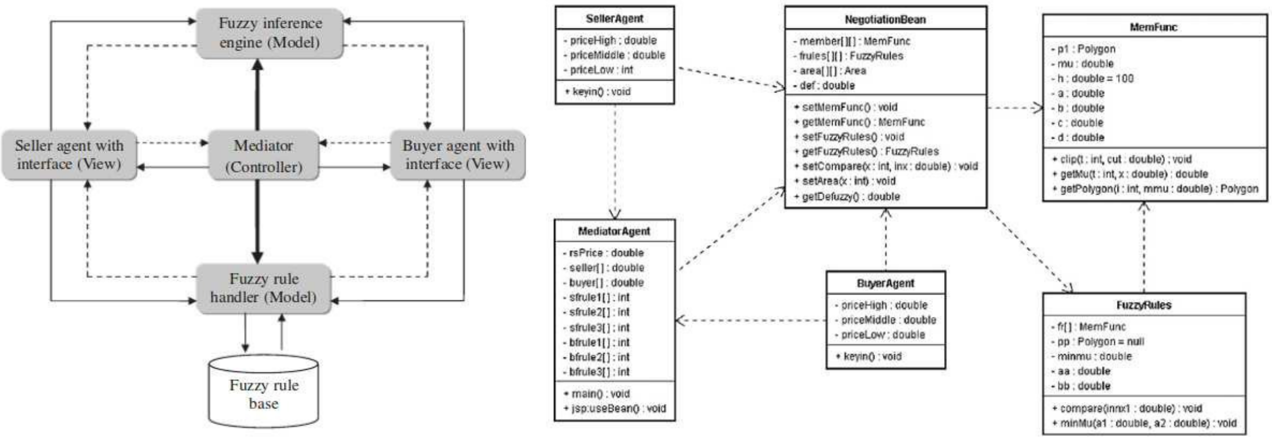


Fig. 1. The MVC design pattern and the price negotiation system class diagram [25]

we still have to change the knowledge base), the *View* (in order to be compatible with another ontology serving us as a dictionary of concepts used for production rules) as well as the *Controller* (since the model interface has to be updated in order to be compatible with that new ontology).

2) *Changeability*: during the evolution the price negotiation domain description changes significantly: temporal aspects are added to the price negotiation rules similar to "if a seller decreased the price on a small value followed by decreasing the price on a medium value then ...". If we consider this change, we have to modify the *View* (in order to make it compatible with another set of concepts) as well as the *Model* and the *Controller* (for the similar reasons as those in the reusability scenario description).

Thus, the above scenarios require changes in all the three MVC major components.

Applying methods of formal architecture evaluation allows to evaluate component relations and discover whether a system follows a loosely coupled design strategy [26]. Let us note that using component interfaces doesn't guarantee system loose coupling since a scenario might affect changes in an interface, which, in turn, leads to the changes in all the components depending on this interface.

Both of the above mentioned scenarios require an ontology to be changed (e.g. the dictionaries used in production rules). This is a key factor since other two components (e.g. the *View* and the *Presenter/Controller*) directly depend on the domain ontology.

We think that the problem is that an interface is strongly connected to the subject domain. If we succeed to remove the subject domain related information from the interface, we are able to construct a user friendly GUI without having to redesign an expert system architecture in order to fit every change in the expert system subject domain.

Following [23], there are two major problems to be resolved while developing a GUI application:

- *UI*: How does the user interact with my data?
- *Data Management*: How do I manage my data?

Each problem falls into three more concrete questions (see Figure 2). In our work we only address the following questions: *What is my data?* (see Section III-A), *How do I change my data?* (see Section IV-C1), and partially the question *How do I display my data?* (see Section IV-C2). In MVC terms the questions are: *What is the interface of my model for the view?* and *What is the interface of my model for the controller?*.

## III. Introducing the Model

After the analysis of a series of existing expert system implementations, we realized that there are two basic approaches to define a *Model* interface:

1) **Expert system domain aware model interface:** There are interface methods directly connected to the subject domain as it is implemented in work [25] (e.g. *setPriceHigh* in the example of automated price negotiation on the web). Figure 5 (1) illustrates this issue.
2) **Domain agnostic model interface:** A user interface communicates directly with the inference engine interface as it is implemented in works [27], [28]. It means that there are methods like *assertFact(fact: Fact)* (see Figure 5 (2)).

If we follow the first approach, we have to change the *Model* interface in order to respond to subject domain changes. If we rely on the second one, we are able to keep the *Model*–presenter interaction interface, but changes in the subject domain still require changes both in the *Model* (the knowledge base rules) and in the *Presenter* (since the latter should be able to assert new facts to the knowledge base). Thus, both approaches are not aimed at using the MVP pattern in the best way.

In order to separate the *Presenter* and the *Model* we propose to complement an expert system subject domain ontology (e.g. automatic price negotiation or CLI application deployment) with a user communication subject domain ontology (*UserComm*, see Figure 5 (3)). As far as a problem of user

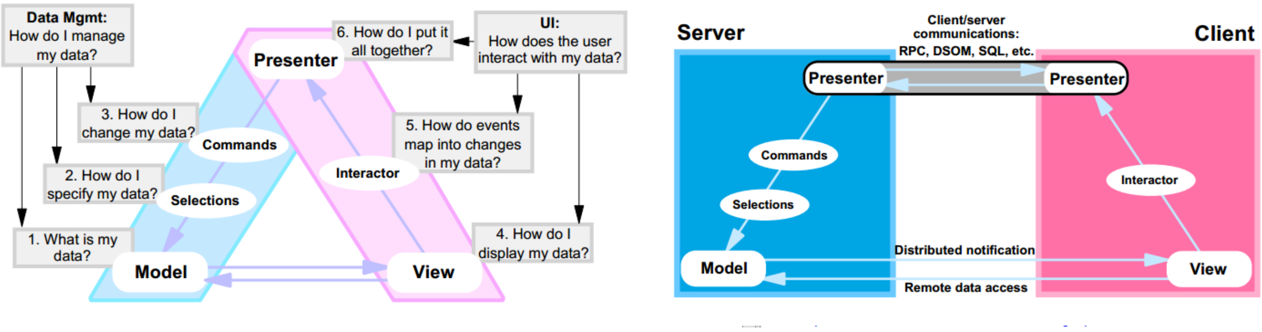


Fig. 2. The MVP design pattern [23]

communication can be described with no connection to the subject domain problems, a user communication model can be developed once and then reused oftentimes as long as its interface is well designed and doesn't change. Therefore, as a *Model*, we propose to use a knowledge base (with rules description and its working memory), but an interface of the model for other components includes only the concepts from the *UserComm* domain.

### A. Introducing a Request-Activity and Related Facts Core Ontology

In order to formalize both software provisioning and execution error description, as well as the relationships between an error and an error resolution procedure, knowledge engineering formalisms are required. In [29] and [9] we proposed and argued for a *Software Provisioning Ontology* that describes processes of software code building and execution with much attention paid to represent build and execution errors as well as the actions required to fix the recognized errors. In the earlier mentioned works we also demonstrated how ontologies of specific tasks can be defined by extending the core ontology base entities.

Hereinafter we only introduce basic entities of the above mentioned software provisioning ontology aimed at describing and resolving the problems of CLI software provisioning to a virtual platform. Let us mention again that we focus on deployment problems in relation to the special software class – *research software implementing the MIR algorithms*. The authors of such programs are usually able to implement an algorithm in the form of CLI-based console application, which transforms the input data to the output according to the data formats required by a certain algorithm evaluation system. However, it is common that a developer might not be experienced enough to resolve runtime environment failures or to guarantee that virtual platform requirements are satisfied.

The major concepts of this ontology are *Activities* and activity *Requests* (see Figure 3). An *Activity* is a sequence of *Actions* aimed at achieving an activity goal, while a *Request* can be considered as a new goal setting. An activity might aggregate requests (being subrequests, in a sense), while a request might consist of activities: if an activity fails, an error has to be identified and fixed, then the activity for the same request has to be restarted. If the activity failure can not be fixed, the request is considered failed.

In order to describe activity results, we introduce a concept of an *Activity status*, which is twofold: there may be an *Activity runtime status* and an *Activity completion status*. The *Activity runtime status* instances are an *Activity being executed* and an *Activity suspended*. The *Activity completion status* instances are an *Activity succeeded* and an *Activity failed*. We assume that an activity is completed successfully if the activity goal is reached (for example, for the activity *Unpacking* the artifact has been successfully unpacked). Otherwise the activity is failed (for example, some file artifact has not been unpacked for the reason that the required archiving utility has not been found in the system).

The *Request* features a necessary and appropriate condition that there starts an activity of a particular type. Similar to an activity concept, a *Request* might also have its status, which is also twofold: there are a *Request runtime status* and a *Request completion status*. The *Request runtime status* instances are a *Request being executed* and a *Request suspended*, while the *Request completion status* instances are a *Request succeeded* and a *Request failed*. If at least one activity for the request is completed successfully, the request is considered to be completed successfully too. By contrast, if all the activities associated with the given request failed, the request is considered to be failed.

Subject domain ontologies are rarely used in expert systems directly: they are usually too common to describe the subject domain-related specific tasks. However, we are able to define an ontology of specific tasks by extending the base entities of the core ontology, and in so doing to follow an *extendibility* principle of the ontology design: "an ontology should be designed so as to allow to use shared vocabularies and to support monotonic ontology extension or/and specialization (i.e. new terms might be introduced without revising existing definitions)" [30].

Let us note that in the earlier mentioned work [9] we also demonstrated how to construct the knowledge base production rules in order to manage processes of client application building and execution with detecting respective errors while using some building tool (e.g. *maven*) as a kind of specific building system.

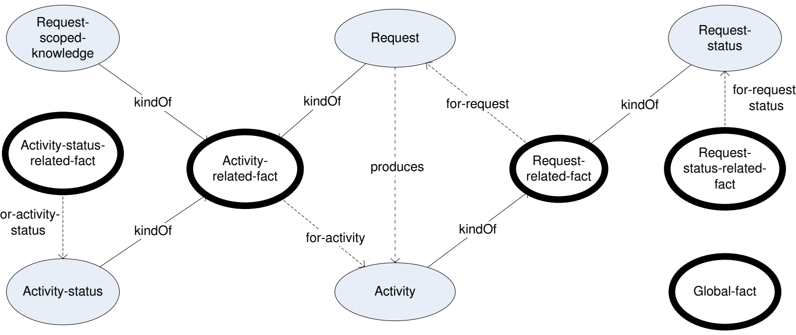


Fig. 3. Activities and requests are main ontology concepts

### B. User Communication Ontology

The user communication ontology (that we refer to as *UserComm*) is based on *Request-RequestStatus* and *Related-Facts* concepts of the core ontology mentioned in the previous section.

The *UserComm* ontology provides the following concepts (see Figure 4) which represent users' requests:

- A **user request** extends a *Request* concept and represents a request generated by a user. Associated *request related facts* provide more details on the user request
- An **ordered argument** extends a *request related fact* and represents a part of the user request in the form of an *integer-value*, where a value is an arbitrary string.
- A **key-value argument** extends a *request related fact* concept and represents a part of the user request in the form of *key-value* where *key* and *value* are strings.
- A **key-(multivalue argument)** extends a *request related fact* concept and represents a part of the user request in the form of *key-(array of values)*, where *key* and each *value* are strings.
- A **named artifact** extends a *request related fact* concept and represents a part of the user request in the form of a *name-(binary file)*, where *name* is a string
- An **unnamed artifact** extends a *request related fact* concept and represents a part of the user request in the form of *binary file*. At most one unnamed artifact may be associated with a request.

The *UserComm* domain description serves as an abstraction layer used by both the *Presenter* and the *Model* allowing to hide real expert system's domain from the *Presenter* as shown in Figure 5 (3). UML diagram presenting the *Model* interface to be used by a *Presenter*, a *Controller* or a *View* is shown in Figure 6. In the following sections we describe how the *Presenter* and the *View* use the *Model* interface, and how the expert system domain rules communicate with a user via the *UserComm* ontology abstraction layer.

## IV. Software Provisioning Self-Service Networked Infrastructure: An Architecture and Major Components

An architecture of a system for automated experiments with algorithms developed in MIR is shown in Figure 7 (see also [6]). It includes the following components:

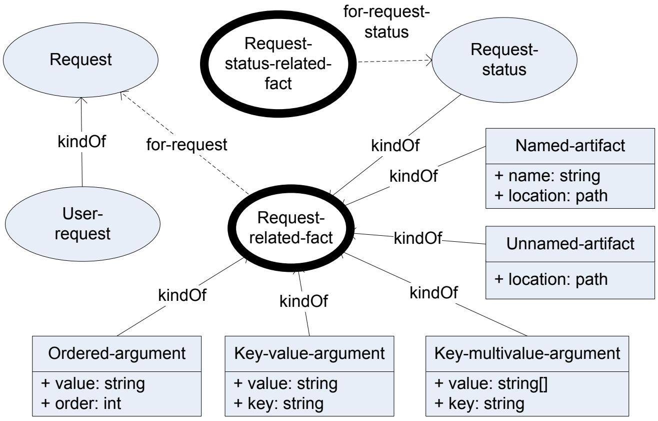


Fig. 4. *UserComm* domain ontology

- User interface
- Submitted applications repository
- Virtual machine images repository
- Deployment knowledge base
- Deployment manager
- Input provider service
- Result collecting service
- Statistics service
- Authentication and authorization service
- Client virtual machines
- Cloud manager
- Cloud administrator console

Some of the listed components (e.g. a cloud administrator console, authentication and authorization services, etc.) are provided by a cloud infrastructure.

Provisioning client applications to a cloud is supported by two major components of a virtual platform: a *cloud broker* and a *deployment manager* (see Figure 8). The latter is a composition of a *deployment manager agent* and a *configuration manager*. A *knowledge base* (KB), an *inference engine* and its *working memory* are components of an expert system controlling the provisioning process.

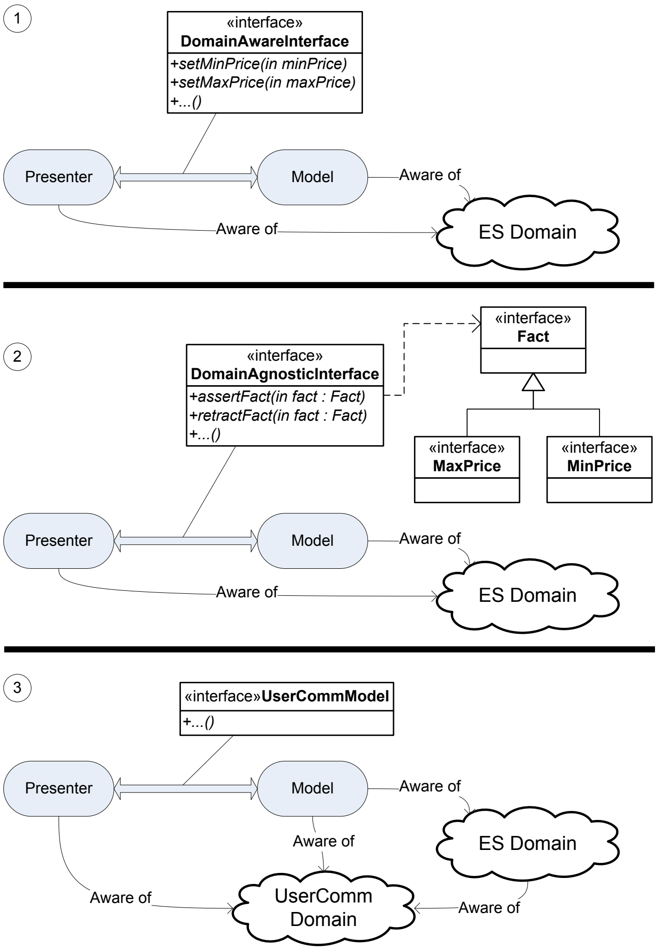


Fig. 5. Model interface for a presenter: (1) expert system domain aware; (2) expert system domain agnostic; (3) user communication domain aware

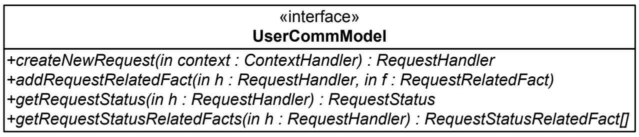

«interface»
UserCommModel
+createNewRequest(in context : ContextHandler) : RequestHandler
+addRequestRelatedFact(in h : RequestHandler, in f : RequestRelatedFact)
+getRequestStatus(in h : RequestHandler) : RequestStatus
+getRequestStatusRelatedFacts(in h : RequestHandler) : RequestStatusRelatedFact[]

Fig. 6. Model interface

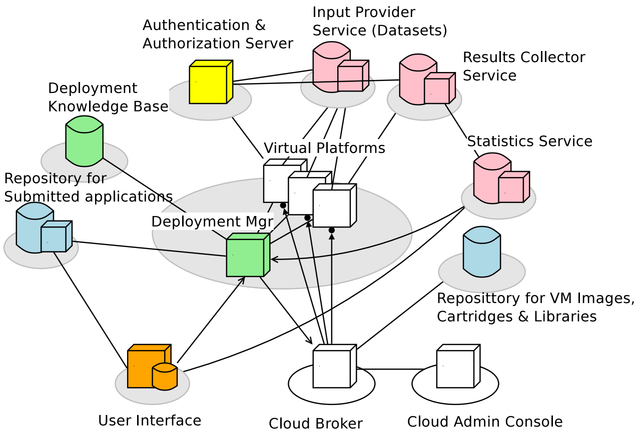


Fig. 7. CLI software provisioning service architecture

### A. Deployment Manager

Similar to the *MEDEA* approach, a wrapper is uploaded to a virtual machine deployed in a cloud. The wrapper provides an HTTP interface and executes a client CLI application in response to user inputs provided via an HTTP proxy interface. Normally the wrapper consists of two components: a *proxy* and an *executor*, where the proxy invokes the executor directly according to HTTP commands received via an HTTP interface. In our implementation, in contrast to a traditional approach, the proxy and the executor never interact directly but via indirect communications using a knowledge base. In terms of the MVP design pattern the *Proxy* is a *Presenter*, the *Knowledge Base* is a *Model*. while the *View* could be either a client side web browser, or a server side HTML code generator component.

Let us note that an expert system often communicates not only with a user but with other components of the system. For instance, for the purpose of CLI application deployment an expert system might need executing a command (a communication with the executor) or changing a platform configuration (a communication with the configuration manager). The problem is how to define an interface that doesn't need to be changed if an expert system domain changes. This problem is similar to the problem of interface definition between a *View* and a *Model*, as well as between a *Presenter* and a *Model* in MVP pattern. Hence, we can use the same approach. We can define an ontology (*ActionExecution* ontology or *ConfigManagement* ontology) used for communication between an expert system and any external component.

### B. Deployment Manager Agent

The deployment manager agent gathers runtime information about the client application and about the environment state and uses the knowledge base in order to resolve deployment and execution errors such as absence of required components or libraries, improper runtime environment version, etc. The agent interacts with the configuration manager by using the *ConfigManagement* ontology (including high-level commands like "need Python3") in order to reconfigure the platform properly. In turn, the configuration manager interacts with the cloud broker (which is a component provided by a cloud itself) by using low-level commands (e.g. "change VM image"). In so doing, the configuration manager controls the installation of the external components (such as language runtimes, necessary middleware or databases) to the platform. It also controls virtual machines recreation if required.

### C. Proxy

The proxy component is responsible for providing a capability to access the expert system by supporting two routines:

1) Asserting user requests to the knowledge base;
2) Retrieving the execution status.

In a sense, the proxy acts as an adapter transforming the data representation from one form (HTTP) to another (*UserComm* facts) and vice versa. The *UserCommModel* interface (see Figure 6) is used for interaction with the knowledge base. Request related facts for a request generated by the proxy

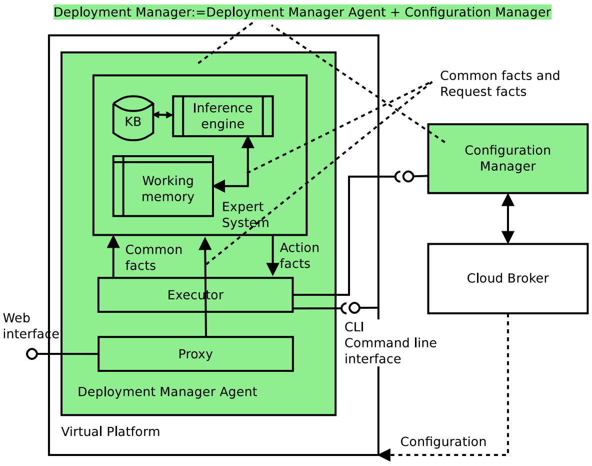


Fig. 8. Managing application deployment in a cloud

should be only concepts from the set of concepts defined by the *UserComm* ontology. In order to interact with a user the REST-like HTTP interface is provided (as defined in Table I).

Two major procedures for HTTP requests processing are described in the following subsections.

*1) Proxy: Processing POST Requests:* In this section we describe a method for an arbitrary HTTP request transformation to a knowledge base facts representation with use of a fixed set of *UserComm* ontology concepts. The method consists of five major steps:

1) Parse an HTTP request (according to RFC 2616).
2) Map parts of the HTTP request to ontology facts according to the rules defined in Table II.
3) Obtain a $ContextHandler$[6] object from the HTTP request URL.
4) Obtain a $RequestHandler$ object by invoking $createNewRequest$ method (see Figure 6) with a $ContextHandler$ parameter obtained in the previous step. In fact, this invocation asserts new $UserRequest$ object to the knowledge base of the model.
5) Assert all the $RequestRelatedFacts$ by invoking $addRequestRelatedFact$ method with a $RequestHandler$ parameter obtained in the previous step.

*2) Proxy: Processing GET Requests:* In this section we describe a method for querying a $RequestResult$ with an HTTP request. The method consists of six major steps:

1) Parse an HTTP request according to RFC 2616.
2) Extract $RequestHandler$ object from the request URL.
3) Get a $RequestStatus$ for a $RequestHandler$ by invoking $getRequestStatus$ method (see Figure 6) with a $RequestHandler$ parameter obtained in the previous step.
4) Get a $RequestStatusRelatedFacts$ for a $RequestHandler$ by invoking $getRequestStatusRelatedFacts$ method (see Figure 6).
5) Decide on view layout on the base of the $RequestStatus$ properties (including runtime type information) and a variety of $RequestStatusRelatedFacts$ (or use some default layout).
6) Draw each object with its own widget being a part of the layout.

### D. Execution Process

As we described in our earlier work [6], we extended the *MEDEA* and *NEMA* execution model by adding the second phase of the execution process. During the first phase (command execution) some debugging information can be written to *stdout*/*stderr*, to the environment logs and so on. These execution results are represented as facts and asserted into the working memory. During the second phase the results are analyzed. As soon as the expert system determines the execution failure and the error cause is determined, the expert system issues appropriate reconfiguration command. This command is handled by the executor, which either performs the necessary operations by itself or delegates them to the configuration manager.

An *executor* is a platform-specific component. It observes a working memory for the presence of action facts. We use a deferred action execution model: as soon as the inference engine does not have any active rules, the executor performs the required actions described in the form of action facts stored in the working memory. Such an approach is tolerant to action facts addition, deletion or modification up until the moment when the agent performs the action.

### E. Configuration Manager

The configuration manager interacts with the cloud broker in order to provide the required configuration, i.e. to install/uninstall necessary/unnecessary system components, (frameworks, applications). As soon as the reconfiguration process is completed, the configuration manager asserts new environment configuration facts. As a result, this assertion might activate the rules asserting actions in order to execute the failed command again or to notify the user about an unrecoverable failure detected.

## V. Evaluation

First, let us demonstrate how to apply the proposed system architecture to develop a web-based expert system for CLI applications deployment in a cloud. We consider only user communication aspect of the system in this example.

The *Proxy* component provides an API as described in Table I. We extended the proxy interface with a GET method for all URLs supporting POST requests. The response to the GET request to such a URL returns a simple HTML page that a user can use to upload an artifact (zip archive, for example)

[6] According to Section III-A a context for a *Request* is an *Activity*. By convention a top-level activity may represent a user and a context for all top-level requests. For a subrequest its context is represented by the request's parent *Activity*

TABLE I
REST INTERFACE OF THE PROXY COMPONENT

| URL | HTTP method | Description |
|---|---|---|
| /action/<contextHandler>/<relative path>?<query string> | POST | Submit a user request within the context $contextHandler$. The request is executed according to the procedure described in section IV-C1. |
| /status/<requestHandler> | GET | Get execution result for a request identified by $requestHandler$. The request is executed according to the procedure described in section IV-C2. |

TABLE II
HTTP REQUEST TO *UserComm* CONCEPTS TRANSFORMATION RULES

| HTTP message part | Ontology concept | Description |
|---|---|---|
| Segment(*) of <relative path> | OrderedArgument(segNo, segValue) | The segment is represented by its value *segValue* (path segment name as it appears in the URL) and its order *segNo*(i.e. number of '/' signs in the URL before the segment) |
| Parameter(**) from <query string><br>('key=value') | KeyValueArgument(key, value) | The query string parameter is represented by its *key* and *value*. |
| Parameter from <query string><br>('key' or 'key=values[]') | KeyMultivalueArgument(key, value[]) | The query string parameter is represented by its *key* and *value[]* (zero or several, but not exactly one value). |
| HTML form input field<br>(input field type is not 'file') | KeyValueArgument(fieldId, value) | The form input field is represented by its id *fieldId* as it appears in HTML code or in multipart HTTP message and *value* (the contents of the field or the entity value in multipart HTTP message) |
| HTML form input field<br>(input field type is 'file') | NamedArtifact(fieldId, fileContentsPath) | The file input field is represented by its id *fieldId* as it appears in HTML code or in multipart HTTP message and a path *fileContentsPath* to a local copy of the binary file received from a user |
| HTTP named entity<br>(not HTML form) | NamedArtifact(entityName, fileContentsPath) | Named entity is represented by its id *entityName* as it appears in multipart HTTP message and a path *fileContentsPath* to a local copy of the binary file received from a user |
| HTTP default entity | UnamedArtifact(fileContentsPath) | Default entity is represented by a path *fileContentsPath* to a local copy of the binary file received from a user |

(*) According to RFC 3986 path of a URL consists of zero or more segments separated by slash ('/') character.
(**) RFC 3986 doesn't set any restrictions on a query string format. In practice Web developers use ampersand ('&') separated 'key=value', 'key=values[]' or 'key' format as defined in RFC 1866.

to the server. Selecting a file to upload followed by clicking on "Submit" button causes the form to be uploaded to the same URL that is used to download the form. In our system the GET URL path is "/action/user/deploy/" so the POST URL path is also "/action/user/deploy/".

According to the algorithm described in Section IV-C1, the proxy parses an HTTP message in the following way: a *contextHandler* is string "user", a relative path consists of one segment "deploy". Hence an input field with an artifact is transformed to $NamedArtifact(\text{“}artifact'', \text{“}/local/file/path'')$ and URL path is transformed to $OrderedArgument(1, \text{“}deploy'')$. The proxy invokes $createNewRequest$ method (see Figure 6) of the model and receives a $RequestHandler$ that is used to assert all other facts via $addRequestRelatedFact$ method and returns the $RequestHandler$ to the user's browser. The latter redirects the response to URL "/status/RequestHandler".

In the knowledge base we can construct a rule translating a *UserComm* domain to an expert system domain as follows:

```
RULE 'Deploy an artifact'
IF
  ctx : Context('user')
  req : UserRequest(ctx)
  exists OrderedArgument( order==1
                    AND value=='deploy'
                    AND request==req)
  artifact : NamedArtifact(name=='artifact'
                     AND request==req)
THEN
  assert(Expert system domain facts)
END RULE
```

Similarly, we are able to define the rules to translate the expert system domain back to the *UserComm* domain.

Redirection to "/status/RequestHandler" URL enables users to monitor execution process. In the simplest case the request result might be described as a status string, e.g. "in progress", "completed" and so on.

### A. Expert System Evaluation: Preview

Currently we tested the approach by implementing a prototype system that was successfully used for automatic deployment of two *Java* projects (built with *maven*) selected among the projects submitted to the *MIREX 2013* contest[7]. During the deployment there were several configuration errors. For each error we provided the knowledge base rule in order to detect and fix configuration errors. Table III lists the examples of build and run errors discovered during our experiments.

### B. Scenario Based System Architecture Evaluation

*1) Reuse Scenario: Domain That Changes:* Let's revisit the case study investigated in Section II-C: the change of the *Model* provoked changes in both the *View* and the *Controller*. As we discovered in Section II-C, these changes are conditioned by the fact that the *Model* provides a domain-specific interface for data modification. If we use the described architecture (which includes two domains: the communication domain (*UserComm*) and the expert system domain), changing the expert system domain doesn't lead us to the communication domain changes. Thus, the *Model* interface can be preserved, and the *Presenter* remains unchanged. The *View* might need to be updated in order to support new *RequestStatusRelatedFacts* introduced by the changed expert system domain. The communication interface between a *View* and a *Model* (as well as between a *View* and a *Presenter*) remains unchanged. To sum up, updating the *Model* component leads to changes in the *View* only because this new domain has new concepts to be visualized.

*2) Change Scenario: Model That Changes:* The *Model* change can be handled the same way as the domain change (see section V-B1). Updating the *Model* might lead to changes in the *View* if this new model has new concepts to be visualized.

*3) Change Scenario: Managed and Unmanaged Execution Modes:* As we described in [6], the architecture with isolated proxy and executor components allows implementing different execution modes (e.g. system behavior) without modifications of neither the proxy, nor the executor. In practical cases, at least two execution modes are useful: *managed mode* and *unmanaged mode*. Regardless of the current mode, the knowledge base is able to detect and fix recoverable errors in order to support client software automatic deployment.

In the *managed mode* we have a predefined client code invocation command, as well as we use strict validation of the input and output data. Validated output data are automatically published as request status related facts. Especially in case of MIR, in the managed mode it is possible to keep private music collections safe while providing access to these collections for processing by third party algorithms.

In the *unmanaged mode* the invocation command, as well as its input and output data are defined in an HTTP request. The framework doesn't check input and output data, but it still detects and fixes recoverable execution errors.

While these modes differ significantly in behavior, they are implemented entirely at the knowledge base level. It means that in order to support managed or unmanaged modes, only changes in the knowledge base component (i.e. in the *Model*) are required.

*4) Change Scenario: Deploying a CLI application in IaaS or PaaS clouds:* The basic idea of this evaluation scenario is to check whether the expert system is able to deploy a CLI application to *PaaS* and *IaaS* clouds. The platform configuration can be handled in three different ways:

1) In *PaaS* clouds the configuration manager can delegate all operations to the cloud broker (as in *OpenShift*). Necessary components can be installed as cartridges developed by the communities (e.g. *Python* or *Ruby* cartridges). This approach is the easiest to implement, but it requires support from the cloud as Figure 9 (left) illustrates.
2) In *PaaS* and *IaaS* clouds the configuration manager is able to install all the required software itself. Software installation can be implemented using the same interfaces that end users have. Indeed, within the context of deployment there is no big difference whether the deployment manager deploys a particular version of a build system (e.g. *maven*), or a MIR research application. This approach is illustrated in Figure 9 (middle).
3) In *IaaS* clouds it is possible to have a set of preconfigured virtual machines, hence configuring means switching of virtual machine images as shown in Figure 9 (right). This way might be resource consuming but in some cases there is no other choice. For instance, it is useful to have two images: one with **nix* OS and the other one with *Windows* OS, because there is no way to install *Windows* applications to **nix* or vice versa with no virtual machines usage.

In its pure forms neither *PaaS* nor *IaaS* fits the task of MIR research software automatic deployment and execution. For example, using the only *PaaS* it is often impossible to implement access to big local data. If we consider an example of *MSD* (Million Songs Dataset [31]) with its size of about 240 GB, we immediately face two problems: 1) a virtual platform is usually limited by only several GBs of disk space, 2) for a virtual platform it is usually not allowed to mount new partitions, volumes or remote file systems. Our subject domain restrictions make it almost impossible to force client software developers to use some network file system similar to *webdav*.

In turn, an *IaaS* does support mounting new partitions. However, installing and configuring, say, a *Python* environment requires installing the *Python* interpreter from the repositories and its additional configuration by using scripts, i.e. the process not trivial for non-experts. On the contrary, in a *PaaS* the same effect may be achieved by only one command for installing the respective cartridge (of course, if the required cartridge exists, and the latter observation is true for a great majority of practical cases and *PaaS* platforms). Therefore we propose to use a hybrid approach *IaaS+PaaS*, where a *PaaS* may be deployed within an *IaaS* cloud.

[7] http://www.music-ir.org/mirex/wiki/2013:Main_Page

TABLE III
EXAMPLES OF ERRORS DISCOVERED DURING EXPERIMENTS WITH MIREX 2013 PROJECTS

| Error | Component | Detection method | Recovery action |
|---|---|---|---|
| Incorrect encoding of source file | javac | Pattern matching. Look for “error: unmappable character for encoding” in javac log | Add appropriate javac/maven flag to specify encoding. Encoding can be detected automatically or provided by user |
| copy-maven-plugin runtime exception | maven | Pattern matching. Look for “copy-maven-plugin:0.2.5:copy” and “java.lang.NoClassDefFoundError: Lorg/sonatype/aether/RepositorySystem;” in maven log | Downgrade maven to version 3.0.5 |
| X11 server required | JVM/AWT | Pattern matching. Look for “java.awt.HeadlessException” in JVM classloader log(*) | Install X11 server (we use Xvfb virtual server) |

(*)This exception is especially interesting for two reasons: 1) nobody expected AWT exception in CLI application and 2) This exception is caught in client code, but it is not handled properly (it is ignored). The only way to detect that the exception was thrown is to analyze the classloader log.

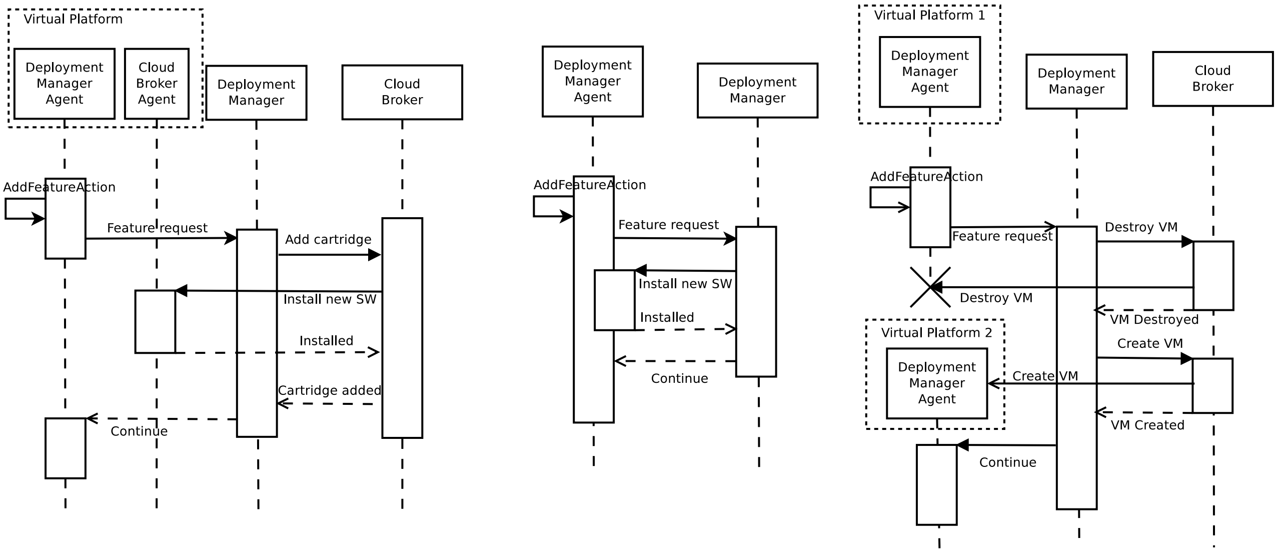


Fig. 9. PaaS/IaaS reconfiguration: (Left) New cartridge installation; (Middle) New software installation; (Right) Virtual machine recreation.

## VI. CONCLUSION

In this paper we have examined the problem of CLI application provisioning in clouds. Being a web application interacting with a cloud broker in order to manage cloud resources and their configuration, the proposed infrastructure provides a *PaaS* platform to deploy CLI applications as *SaaS* services. In contrast to existing solutions we have proposed an architecture where a proxy component and an executor don’t interact directly. Instead of invoking each other, they assert or retract facts in/from an expert system working memory. Thus, the approach draws on the knowledge engineering formalisms used not only for configuration errors recovery, but also for decision making on how to handle requests in order to protect intellectual property (in regards to MIR one can talk about music collections, software implementations, etc.). Although in this research we experimented mostly with automatic deployment of *Java+maven* applications, we are working on knowledge base extensions that allow deploying native *Windows*, *Python*, *MatLab* and *Vamp*[8] applications.

The proposed architecture addresses some of the most complex tasks of client virtual machine automatic reconfiguration, including the following:

1) Installing operating systems on a virtual machine;
2) Installing and configuring a build environment;
3) Installing and configuring a required version of a runtime environment;
4) Installing third party libraries during building;
5) Installing third party runtime libraries;
6) Providing access to machine learning and test data;
7) Providing access to a storage for execution results;
8) Providing user-side access to a client virtual machine.

We have investigated MVC and MVP design patterns as well as the major difficulties of their application to implementing a user interface for a web-based expert system. By introducing a special ontology representing user communication concepts, we have attempted to achieve an MVP-based implementation of an expert system with respect to the loose coupling design requirements, which, in turn, are strongly connected to improving such software quality properties as reusability and changeability. With regards to the demands of scientific communities, we believe that the introduced approach is in good direction to reproducibility, which is

[8] http://www.vamp-plugins.org/

"not an afterthought – it is something that must be designed into a project" [32]. Let us conclude with sharing the idea that research reproducibility might add an overhead (that we attempted to avoid in our approach). However, even "some reproducible practices are better than none – it does not have to be perfect to be a huge improvement" [2].

## Acknowledgment

We would like to express our great appreciation to Nina Popova and Michael Tramontano for the very valuable suggestions. The authors would like to thank Prof. Franck Leprevost for the invitation to present this research within the framework of the 2014 Eastern Europe–Luxembourg Workshop on Cloud Computing, Communications, Security and Services in conjunction with the International Conference on Cloud Networking. An opportunity to have a preliminary discussion of this work has significantly improved our current contribution.

## References

[1] M. D. Plumbley, C. Cannam, and S. Dixon, "Tutorial on reusable software and reproducibility in music informatics research to be presented in to be presented at the 13th ismir conference," Centre for Digital Music, Queen Mary, University of London, 2012.

[2] S. Sufi, N. C. Hong, S. Hettrick, M. Antonioletti, S. Crouch, A. Hay, D. Inupakutika, M. Jackson, A. Pawlik, G. Peru, J. Robinson, L. Carr, D. De Roure, C. Goble, and M. Parsons, "Software in reproducible research: Advice and best practice collected from experiences at the collaborations workshop," in *Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering*, ser. TRUST '14. New York, NY, USA: ACM, 2014, pp. 2:1–2:4. [Online]. Available: http://doi.acm.org/10.1145/2618137.2618140

[3] Y. Janin, C. Vincent, and R. Duraffort, "Care, the comprehensive archiver for reproducible execution," in *Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering*, ser. TRUST '14. New York, NY, USA: ACM, 2014, pp. 1:1–1:7. [Online]. Available: http://doi.acm.org/10.1145/2618137.2618138

[4] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "Openml: Networked science in machine learning," *SIGKDD Explorations*, vol. 15, no. 2, pp. 49–60, 2013. [Online]. Available: http://doi.acm.org/10.1145/2641190.2641198

[5] D. Milne and I. H. Witten, "An open-source toolkit for mining wikipedia," *Artif. Intell.*, vol. 194, pp. 222–239, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.artint.2012.06.007

[6] E. Pyshkin and A. Kuznetsov, "A provisioning service for automatic command line applications deployment in computing clouds," in *2014 IEEE Intl Conf on High Performance Computing and Communications (HPCC)*, Aug 2014, pp. 518–521.

[7] C. Bunch, "Automated configuration and deployment of applications in heterogeneous cloud environments," Ph.D. dissertation, Santa Barbara, CA, USA, 2012, aAI3553710.

[8] Y.-Y. Su, M. Attariyan, and J. Flinn, "Autobash: Improving configuration management with operating system causality analysis," in *In Proceedings of the 21st ACM Symposium on Operating Systems Principles (Stevenson*, 2007, pp. 237–250.

[9] E. Pyshkin, A. Kuznetsov, and V. Klyuev, "Understanding software provisioning: An ontological view," in *Databases in Networked Information Systems*, ser. Lecture Notes in Computer Science, W. Chu, S. Kikuchi, and S. Bhalla, Eds. Springer International Publishing, 2015, vol. 8999, pp. 84–111. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16313-0_7

[10] K. West, A. Kumar, A. Shirk, G. Zhu, J. Downie, A. Ehmann, and M. Bay, "The networked environment for music analysis (nema)," in *Services (SERVICES-1), 2010 6th World Congress on*, July 2010, pp. 314–317.

[11] J. C. Giarratano and G. Riley, *Expert systems: principles and programming*. Brooks/Cole Publishing Co., 1989.

[12] Y. Duan, J. S. Edwards, and M. Xu, "Web-based expert systems: benefits and challenges," *Information & Management*, vol. 42, no. 6, pp. 799–811, 2005.

[13] N. Dunstan, "An interactive webbased expert system degree planner," in *The Second International Conference on Informatics Engineering & Information Science (ICIEIS2013)*. The Society of Digital Information and Wireless Communication, 2013, pp. 302–308.

[14] I. M. Dokas, "Developing web sites for web based expert systems: A web engineering approach." in *ITEE*, 2005, pp. 202–217.

[15] N. Dunstan, "A hybrid architecture for web-based expert systems," *International Journal of Artificial Intelligence and Expert Systems*, vol. 3, no. 4, pp. 70–79, 2012.

[16] R. A. Harrington, S. Banks, and E. Santos Jr, "Development of an intelligent user interface for a generic expert system," in *Online Proceedings of the Seventh Midwest Artificial Intelligence and Cognitive Science Conference*, 1996.

[17] M. Nofal and K. M. Fouad, "Developing web-based semantic expert systems," *IJCSI International Journal of Computer Science Issues*, vol. 11, no. 1, pp. 103–110, Jan. 2014.

[18] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," *User Modeling and User-Adapted Interaction*, vol. 22, no. 4-5, pp. 441–504, 2012.

[19] I. M. Dokas, "Developing web sites for web based expert systems: A web engineering approach," in *In Proceedings of the Second International ICSC Symposium on Information Technologies in Environmental Engineering (Magdeburg*. Shaker Verlag, 2005, pp. 202–217.

[20] R. Morales-Chaparro, M. Linaje, J. Preciado, and F. Sánchez-Figueroa, "Mvc web design patterns and rich internet applications," *Proceedings of the Jornadas de Ingenierıa del Software y Bases de Datos*, 2007.

[21] P. Gupta and M. C. Govil, "Mvc design pattern for the multi framework distributed applications using xml, spring and struts framework," *Int J Comput Sci Eng*, vol. 2, no. 4, pp. 1047–1051, 2010.

[22] G. E. Krasner, S. T. Pope *et al.*, "A description of the model-view-controller user interface paradigm in the smalltalk-80 system," *Journal of object oriented programming*, vol. 1, no. 3, pp. 26–49, 1988.

[23] M. Potel, "Mvp: Model-view-presenter the taligent programming model for c++ and java," *Taligent Inc*, 1996.

[24] R. Kazman, L. Bass, M. Webb, and G. Abowd, "Saam: A method for analyzing the properties of software architectures," in *Proceedings of the 16th International Conference on Software Engineering*, ser. ICSE '94. Los Alamitos, CA, USA: IEEE Computer Society Press, 1994, pp. 81–90. [Online]. Available: http://dl.acm.org/citation.cfm?id=257734.257746

[25] C.-C. Lin, S.-C. Chen, and Y.-M. Chu, "Automatic price negotiation on the web: An agent-based web application using fuzzy expert system," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5090–5100, 2011.

[26] B. Roy and T. N. Graham, "Methods for evaluating software architecture: A survey," *School of Computing TR*, vol. 545, p. 82, 2008.

[27] A. Kipkebut, "An evaluation of web based expert system as a catalyst for maize production in kenya," *Computer Engineering and Intelligent Systems*, vol. 5, no. 3, pp. 86–97, 2014.

[28] K. Tutuncu and M. Koklu, "A new expert system shell in turkish language for training," in *Proceedings of the International Conference on challenges in IT, Engineering and Technology*, ser. ICCIET'2014, 2014, pp. 26–30.

[29] A. Kuznetsov and E. Pyshkin, "An ontology of software building, execution and environment configuration and its application for software deployment in computing clouds," *St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems*, no. 2(193), pp. 110–125, 2014.

[30] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *Int. J. Hum.-Comput. Stud.*, vol. 43, no. 5-6, pp. 907–928, Dec. 1995. [Online]. Available: http://dx.doi.org/10.1006/ijhc.1995.1081

[31] B. McFee, T. Bertin-Mahieux, D. P. W. Ellis, and G. R. G. Lanckriet, "The million song dataset challenge." in *WWW (Companion Volume)*, A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab, Eds. ACM, 2012, pp. 909–916. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/www2012c.html#McFeeBEL12

[32] D. L. Donoho, "An invitation to reproducible computational research," *Biostatistics*, vol. 11, no. 3, pp. 385–388, 2010.

DOI: 10.15439/2015F280
ACSIS, Vol. 5

# Concepts extraction from unstructured Polish texts: a rule based approach

Piotr Szwed
AGH University of Science and Technology
E-mail: pszwed@agh.edu.pl

***Abstract*—We present recently developed solution allowing extraction of concepts from unstructured Polish texts with special focus on correct morphological forms of obtained concept names. As Polish is a highly inflected language, detected names need to be transformed following Polish grammar rules. We propose a user-friendly method for specification of transformation patterns, which is based on a simple annotations language. Annotations prepared by a user are compiled into transformation rules. During the concept extraction process the input document is split into sentences and the rules are applied to sequences of words comprised in sentences. Recognized strings forming concept names are aggregated at various levels and assigned with scores. We report also results of initial experiments performed on a medical text.**

***Index Terms*—NLP, Text Mining, concept extraction, unstructured text, inflection, rules**

## I. Introduction

IN THIS PAPER we investigate the problem of concepts extraction from unstructured Polish texts. The term concept is defined here as a sequence of words (an n-gram) occurring frequently in analyzed documents. However, as primary function of concepts is to represent objects, ideas, events or activities, extracted names should comprise words belonging to specific parts of speech classes (usually nouns or verbal nouns with complements) and also satisfy certain syntactical restrictions stemming from language conventions.

Concepts retrieved from unstructured textual data have several applications. Their usage improves relevance of documents returned by search queries due to more accurate indexing and clustering. They may help to assure data privacy by identifying sensitive information and removing it from published texts (documents anonymization). Models based on concepts can be used to express and structure knowledge existing within an organization, which is often distributed between large number of documents of various types: e-mails, reports or internal regulations.

The goal of our work was to develop a tool allowing to extract referenced concepts from documents in Polish language. The key assumption made was that concept names should have correct morphological forms according to Polish grammar rules. Such feature was selected due to aesthetic reasons (in various applications concept names are presented to end users), as well as the expectation that in many situations using the correct form may improve sense disambiguation. Hence, the designed solution, apart from identifying concepts, should provide automatic translation of n-grams representing concepts to their nominative form. This task is particularly challenging for the Polish language, which is characterized by the high degree of inflection.

We decided to apply a rule based approach to perform transformation of concept names. The rule language, inspired by Petri nets, allows to define patterns of input tokens and required transformations. For this purpose part of speech (POS) information is used. Considering the complexity of POS tags, specification of rules may be a complicated and tedious task. We build the translation rules fully automatically by defining samples of translation patterns (annotations) and compiling them to rules.

The paper is organized as follows: next Section II discusses the problem of concept extraction, as well as several tools dedicated to Polish language processing. It is followed by Section III, which presents general features of our approach. Next Section IV describes shortly Morfologik, a dictionary and POS tagging/lemmatizing library. In Section V we discuss models, algorithms and other details of the solution. Results of initial experiments are given in Section VI. Last Section VII provides concluding remarks.

## II. Related works

Ontology learning is a process of building ontologies from various data representations. It includes such tasks as identification of concepts, their relations and attributes, arranging into hierarchies. Such task is apparently easier in the case of structured or semi-structured data. A challenging problem, however, is building ontologies (usually taxonomies) comprising concepts extracted from unstructured text documents [14]. For this purpose various text mining techniques can be used: syntactical analysis, Formal Concept Analysis [4] and clustering [7].

Concepts (sometimes also referred as compound terms or phrases) are important features used in Text Mining [23]. Compound terms processing is a technique aiming at improving accuracy of search engines by indexing documents according to compound terms, i.e. combinations of 2 or more single words. During query execution searched compound terms are also extracted form queries (which can be phrases in natural language) and then matched with compound terms attributed to documents. A good example can be a compound term “table wine”. A document referencing it is more likely to discuss wines, than pieces of furniture. Such approach outperforms solutions that use in queries keywords combined

978-83-60810-66-8/$25.00©2015, IEEE

with boolean operators. The idea of statistical compound terms processing was proposed and commercialized by Concept Search company [5]. On the other hand, a syntactical technique based on rules defining patterns for various European languages was developed within CLAMOUR project (see reports at http://webarchive.nationalarchives.gov.uk/20040117000117/ http://statistics.gov.uk/methods_quality/clamour/default.asp).

In [9] Dalvi, Kumar et al. from Yahoo Labs. coined an idea of a *web of concepts*. They claimed that the current model of the web in form hyperlinked pages represented by bags of words should be augmented by extracting concepts and creating a new rich view of all available resources for each concept instance. The paper discussed several use cases enabled by the new approach, including: more accurate web search, browsing optimization, session optimization and advertising. The paper also indicated several challenges, related mainly to information extraction, potential uncertainty and changing data. An algorithm for concept extraction following this idea was proposed in [20].

Blake and Pratt [2] used automatically retrieved concepts to build searchable representations of medical texts. In [3] authors also extracted ontologies from medical texts and showed that using concepts improves search results.

Osinski and Weiss described Lingo, a concept driven document clustering algorithm [19]. Concepts (frequent sequences of terms), were used to label clusters in a document-term matrix.

An important task in NLP language classification is *tagging*, i.e. assigning part of speech (POS) information to inflected forms of words. This is a challenging task for a highly inflected languages like Polish. According to [1] words in English language can be described with about 200 tags whereas for Polish their number ranges at 1000.

In our work we used Morfologik stemming library [17], which is discussed in detail in Section IV. The software and the dictionary was used in several projects including Language tool [18], carrot search [19], PSI toolkit [10], PELCRA [21] and Smyrna [11]. Morfologik dictionary was integrated as a part of PoliMorf [27].

Language tool [18], [16] is a proofreading program supporting a number of languages including Polish. It allows to define text correction rules referring to explicit token values, lemmatized words and part of speech tags. Rules can be specified either in XML format or written directly as Java classes. The tool provides also a visual rule editor, which allows to construct a rule in an interactive manner. As the rules yield suggestions, their core functionality is somehow similar to the concepts extraction rules described in this paper.

## III. Towards Polish concepts extraction

Concepts are terms denoting sets of objects sharing common properties. The most general concepts are represented in texts as single nouns (referring to tangible or abstract objects) and verbal nouns describing actions. Examples of both types can be a *car* (a tangible object) or *driving* (an action). Subclasses of general concepts are usually represented by various complements, e.g. a *green car* (a car having the green color) or *car driving* (an action performed on a specific type of vehicle).

Hence, concepts in texts are represented by n-grams (sequences of tokens), whose elements satisfy grammar rules related to correct words ordering and their morphological forms. The rules are language specific; this in particular applies to the Polish language, in which the complements can be declined depending on the noun case, e.g. for the green car: *zielony samochód* (Nominative), *zielonego samochodu* (Accusative), *zielonemu samochodowi* (Dative), etc.

Table I gives examples of several sentences and concepts identified in a supervised manner. Underlined words indicate parts of concept names denoting their superclasses, e.g. *szybki samochód* $\sqsubset$ *samochód*. (The English equivalent is: *FastCar* $\sqsubset$ *Car*).

The concepts listed in Table I constitute typical parts of speech combinations: adjective+noun (*szybki samochód*, *słone jezioro*, *mały ludzik*), nount+noun (*architektura systemu*, *hurtownia danych*) or verbal noun+noun (*leczenie pacjenta*, *zbieranie informacji*). They appear in source sentences as sequences of tokens that after identifying them should be transformed into appropriate nominative forms following the language grammar rules, especially, as regards noun complements. In most cases concepts occur in distinct parts of sentences, however sometimes they may overlap, e.g three concepts can be selected for the entry 6 in the table. Similarly, for the entry 10 showing two adjectives separated by a coma and a noun, three concepts forming a small taxonomy are possible:

- *mały ludzik* $\sqsubset$ *ludzik*
- *zielony ludzik* $\sqsubset$ *ludzik*
- *mały zielony ludzik* $\sqsubset$ *mały ludzik* $\sqcap$ *zielony ludzik*

### A. Rule based approach

The proposed approach to concept extraction consists in defining rules that search for selected patterns in an input text, then apply appropriate morphological transformations to matched words in order to obtain correct forms of concepts.

This is shown in Fig. 1. A window having the length equal to size of a rule input pattern slides through the input text. Sentences are the natural boundaries for the analysis, i.e. the window stops at the sentence end.

If the rule is applicable for a current n-gram appearing in the window, a set of output sequences of tokens is produced. They constitute candidates for concepts, that can be further analyzed and aggregated according to various attributes: sentences, where they occurred, frequency in the whole document, weights describing confidence, etc.

### B. Specification of rules - learn by example

Selection of the language used to specify extraction rules is an important decision, as the language capabilities have strong impact on both on the process on rules definition and obtained results. In our approach we decided to define rules in a semi-formal way, by giving samples of expected translations rather

TABLE I
IDENTIFICATION OF CONCEPTS IN SENTENECES AND EXAMPLES OF ANNOTATIONS.

| # | Sentence | Concept | Annotation | Function |
|---|---|---|---|---|
| 1. | Jeżdżę *szybkim samochodem.* | szybki samochód | @(szybki samochód = szybki @samochód) | extract noun + verb |
| 2. | Morze Kaspijskie jest *jeziorem słonym.* | słone jezioro | @(jeziorem słonym = słone @jezioro) | inversion + Instrumental → Nominative |
| 3. | *Leczenie pacjenta* przebiegało bardzo wolno. | leczenie pacjenta | @(leczenie pacjenta = @leczenie pacjenta) | extract verbal noun + complement |
| 4. | Opracowałem *architekturę systemu.* | architektura systemu | @(architekturę systemu = @architektura systemu) | Accusative → Nominative |
| 5. | *Rozwojowi rolnictwa* towarzyszyło *zanikanie lasów.* | rozwój rolnictwa<br>zanikanie lasów | @(rozwojowi rolnictwa = @rozwój rolnictwa)<br>@(zanikanie lasów = @zanikanie lasów) | Dative → Nominative<br>extract verbal noun + complement in plural form |
| 6. | Następnie poświęcił się *zbieraniu informacji* o *losach misjonarza.* | zbieranie informacji<br>los misjonarza | @(zbieraniu informacji = @zbieranie informacji)<br>@(losach misjonarza = @los misjonarza) | Dative → Nominative<br>Locative plural → Nominative sing. |
| 6. | Obecnie coraz częściej stosowane są *zwinne metodyki zarządzania projektami informatycznymi.* | zwinna metodyka<br>metodyka zarządzania<br>zarządzanie projektami informatycznymi | @(zwinne metodyki = zwinna @metodyka )<br>@(metodyki zarządzania = @metodyka zarządzania)<br>@(zarządzania projektami informatycznymi = @zarządzanie projektami informatycznymi) | Nominative plural → Nominative sing.<br>Nominative plural → Nominative sing.<br>Accusative → Nominative |
| 8. | *Hurtownie danych* stanowią osobną klasę *systemów informatycznych.* | hurtownia danych<br>system informatyczny | @(hurtownie danych = @hurtownia danych)<br>@(systemów informatycznych = @system informatyczny) | Nominative plural → Nominative sing.<br>Genitive plural → Nominative sing. |
| 9. | Małe, *zielone ludziki* rozłożyły się u *podnóża dębu.* | zielony ludzik<br>podnóże dębu | @(zielone ludziki = zielony @ludzik)<br>@(podnóża dębu = @podnóże dębu) | Nominative plural → Nominative sing.<br>Locative → Nominative |
| 10. | *Małe, zielone ludziki* rozłożyły się u podnóża dębu. | mały ludzik, zielony ludzik, mały zielony ludzik | @(małe $, zielone ludziki = mały @ludzik \| zielony @ludzik \| mały zielony @ludzik) | Nominative plural → Nominative sing. |

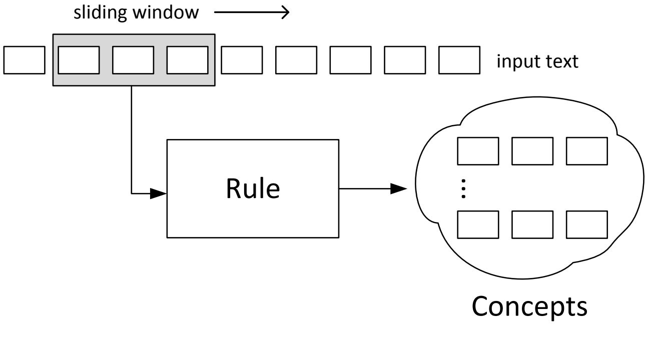


Fig. 1. General rule design

than using formally defined rules. The translation patterns are defined using special, relatively simple textual annotations that can be embedded in a source text or placed in a separate file.

The third column of Table I gives examples of annotation matching the identified concepts. Each annotation starts with '@' (at sign) followed by the annotation body put between two parentheses. The '=' sign separates the input pattern and a list of output sequences, whereas '|' is the output sequence separator. Optional '@' sign within an output sequence identifies a key, i.e. a possible concept superclass.

The last column of Table I specifies the expected function for each annotation example. Usually, annotations define transformations of declination cases and plural forms, which according to Polish grammar rules should be applied both to nouns and their complements. However, they may be used to specify extraction patterns only, not accompanied by morphological transformation. Examples are given in rows 1, 3 and 5.

It should be remarked that, basically, annotations do not specify exact matches of words (for an exact match the dollar sign is used as in row 10). The intent of annotations is to specify indirectly rules applicable to classes of tokens. For example a rule derived from the annotation `@(zwinne metodyki = zwinna @metodyka )` (agile methodologies → agile methodology) in row 6 of Table I applies to feminine nouns in plural form with an adjective. Hence, it should match also the pair of words, which are tagged with the same POS information, e.g. *długie wstążki* (long ribbons), and properly convert it to the singular form *długa wstążka*.

An advantage of the proposed indirect method of rules specification is its simplicity: translation patterns can be defined very quickly, moreover, specifications are not bounded to a particular rule language or execution engine. Depending on a compiler used, rulsets in various rule languages can be obtained.

In many cases defined translation patterns may result in conflicting rules. Let us consider the following annotations:

1) `@(hurtowniom danych=@hurtownia danych)`
2) `@(grubościom pokryw=@grubość pokrywy)`

The first translation results in a noun having a complement in plural form, whereas int the second case the complement is singular. However, they both represent valid transformations. To tackle with such problems the language used to represent rules should allow to attribute results with weights representing likelihood of concept occurrence.

### C. Process of concept extraction

The outline of the process of concept extraction is presented in Fig. 2. A text file containing manually prepared annotations is compiled yielding a set of rules stored in an XML file. The input text is split into sentences and then the rule set is executed for each sentence giving candidate concepts. Finally, the concept occurrences are aggregated and sorted. As different rules may return identical results, the aggregation is at first performed at the sentence level, then for the whole document.

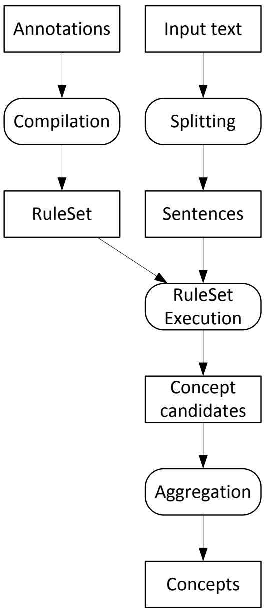


Fig. 2. Process of concept extraction

## IV. Morfologik

During annotations processing and rules execution input words are checked for specific part of speech properties and undergo morphological transformations. This function is provided by the Morfologik.

Morfologik is both a a comprehensive dictionary of Polish inflected forms and a software library written in Java accompanied by a number of utility tools. The main function offered by the Morfologik is *stemming* (lemmatization) of Polish words, i.e. finding a stem (lemma) accompanied by grammar information for an inflected form. Examples of inflected forms and corresponding stems can be: *psu* – *pies* (noun: dog), *czystego* – *czysty* (adjective: clean) or *pisaniu* – *pisać* (verbal noun writing and verb: to write).

Morfologik dictionary can be seen as a relation

$$D \subset IF \times S \times \mathcal{P}, \tag{1}$$

where $IF$ is a set of inflected forms, $S$ is a set of stems, $S \subset IF$, and $\mathcal{P}$ is a set of POS tags defining properties of inflected forms (part of speech, gender, singular vs. plural, declination case, etc.)

A few entries from the Morfologik dictionary are shown in Table II. It can be observed that several stems may be found for an inflected form, e.g. *czarnym* – *czarna*, *czarny*. Moreover, pairs of inflected forms and stems can be attributed with multiple tags (separated with '+' sign).

A tag is a string of symbols (usually abbreviations) separated by colon signs. Examples in the table show typical tag elements: *subst* – noun, *adj*-adjective, *ger*–verbal noun, *sg*–singular, *pl* –plural, *nom*, *gen*, *dat*, *acc*, *inst*, *loc* – declination cases. For a given part of speech class, tag components appear in the same order. In some cases multiple symbols separated by dot sign may occur, e.g. "m1.m2" - various classes of masculine forms.

Based on the Morfologik dictionary scheme two functions can be defined: $stem$ (2) and $synth$ (3). The first takes as input an inflected form and returns a set of stems with accompanying tags, the second synthesizes inflected forms from an input stem and tag.

$$stem \colon IF \to 2^{S \times \mathcal{P}} \tag{2}$$

$$synth \colon S \times \mathcal{P} \to 2^{IF} \tag{3}$$

Morfologik fully supports stemming. The library uses internally an efficient dictionary representation based on Finite State Machine (FSM) model, which is characterized by compact data size and short access times [8]. The current Morfologik dictionary for the Polish language (version 2.0) constitutes a large 317 MB text file, whereas the same data compiled to FSM are stored in 2.7 MB binary file. The precompiled dictionary is a part of Morfologik distribution, however, the library offers also a number of awk scripts to preprocess the dictionary data, as well as tools allowing to compile it into FSM.

Unfortunately, the synthesizing function is not provided directly by the library, although the documentation suggests that it is possible to rebuild (revert) the dictionary and recompile to FSM form. As the guidelines for implementing an FSM

TABLE II
SAMPLE ENTRIES FROM THE MORFOLOGIK DICTIONARY.

| $IF$ - inflected form | $S$ – stem (lemma) | $\mathcal{P}$ - part of speech tags |
|---|---|---|
| czarnym | czarna | subst:pl:dat:f |
| | czarny | adj:pl:dat:m1.m2.m3.f.n1.n2.p1.p2.p3:pos<br>+adj:sg:inst:m1.m2.m3.n1.n2:pos<br>+adj:sg:loc:m1.m2.m3.n1.n2:pos<br>+subst:pl:dat:m1+subst:sg:inst:m1+subst:sg:loc:m1 |
| czystego | czysty | adj:sg:acc:m1.m2:pos+adj:sg:gen:m1.m2.m3.n1.n2:pos |
| psu | pies | subst:sg:dat:m1+subst:sg:dat:m2 |
| pisaniu | pisać | ger:sg:dat.loc:n2:imperf:aff:refl.nonrefl |
| | pisanie | subst:sg:dat:n2+subst:sg:loc:n2 |

based synthesizer seemed quite difficult to follow, we left this possibility for further improvements. Instead, we developed a synthesizing function, which uses the dictionary data stored in a local PostgreSQL database. It was populated with 288657 stems, 1173 tags broken on plus (+) signs and 7410145 triples comprising inflected forms, stems and tags.

The implemented $synth(s, p)$ function comprises two steps:

1) A query to the database is made to select a set of triples matching the stem $s$. As the query result set $IFT = \{(if, s', p') \in D \colon s' = s\}$ is returned. The set is usually small, in most cases it contains up to 30 entries.
2) From the triples in $IFT$ a set of inflected forms $IF = \{if \colon (if, s', p') \in IFT \wedge match(p, p')\}$ is selected, whose tags $p'$ match $p$. The $match(t, t')$ function takes into account order of symbols appearing in tags, as well as alternate properties indicated by the dot separator.

The time efficiency of the implemented $synth$ function is obviously inferior to the $stem$ function offered by Morfologik. Single call to $stem$ (based on FSM representation) takes about 0.085 ms, whereas $synth$ (based on database queries) ranges to 0.75 ms. It should be remarked that we have also implemented an ORM version of $synth$ . In this case the execution time is about 1.25 ms.

## V. METHODS

The language used to define rules is based on Petri nets. Each rule comprises ordered sets of input and output places, which are linked by transitions.

Fig. 3 gives an example of the rule comprising three input places, three output places and four transitions. Multiple transitions, here *t1* and *t2* may link a pair of input and output places. The net layout shows also, how inversion of tokens can be achieved.

Each rule transition is assigned with two sets of tags: input $itag$ and output $otag$. Input tags are used as guards, they allow check if the transition applies to a word tagged with part of speech information. Output tags are used to synthesize target inflected form from a word stem (lemma). Additionally, each transition has assigned weight, that can be used to differentiate less and more likely translation schema.

Below we give a formal definition of the language used to define rules.

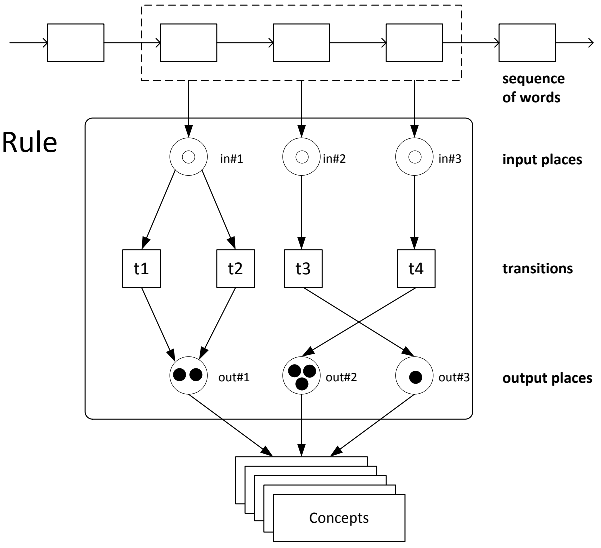


Fig. 3. Detailed rule design

**Definition 1.** Concept extraction rule is a tuple $R = (\mathcal{P}, I, O, T, itag, otag, itoken, \mu)$, where:

- $\mathcal{P}$ is a set of tags
- $I = \{i_1, i_2, \ldots, i_n\}$ is an ordered set of input places,
- $O = \{o_1, o_2, \ldots, o_m\}$ is an ordered set of output places,
- $T \subset I \times (O \cup \{nil\})$ is a set of transitions,
- $itag \colon T \to 2^{\mathcal{P}}$ is a function assigning to a transition a set of *input tags*,
- $otag \colon T \to 2^{\mathcal{P}}$ is a function assigning to a transition a set of *output tags*,
- $itoken \colon I \to \mathcal{A}$ is a function assigning to an input place an exact (possible empty) string form $\mathcal{A}$
- $\mu \colon T \to [0, 1]$ is a transition weight function

The symbol $nil$ used in transition definition denotes a fake output place. It is used when a transition serves as a guard allowing to check if an input place contains a word appertaining to a particular class without translating it. The $itoken$ function is used to assign particular token values to input places. They result from compilation of '\$'-token specifications in input annotations (see Table I row 10).

### *A. Compiling annotations to rules*

An annotation defining multiple outputs $@(\sigma = \pi_1|\pi_2|\dots|\pi_k)$ is split into $k$ rules corresponding to annotations $@(\sigma = \pi_1)$, $@(\sigma = \pi_2)$,..., $@(\sigma = \pi_k)$. (Symbols $\sigma$ and $\pi$ stand here for input and output sequences of words.

For an annotation $@(\sigma = \pi)$, $n$ input places and $m$ output places are created, where $n$ and $m$ denote the lengths of $\sigma$ and $\pi$ respectively.

Then each word in sequences $\sigma$ and $\pi$ is stemmed and transitions between input and output places are derived based on stem matching.

Fig. 4 illustrates this process on an example of annotation `@(informacji przetwarzaniu = @przetwarzanie informacji)`. The equivalent English term is *information processing*. The annotation defines a translation pattern that includes inversion and changing the case of the whole expression from Dative to Nominative form.

Rounded rectangles depict words appearing in the input sequence $\sigma$ (above) and output sequence $\pi$ (below). For each word in the sequence the stemming information is determined with Morfologik $stem$ function. It consists of a stem (lemma) and a set of tags. The word *informacji* is classified as the noun *informacja* tagged as singular or plural form with various declination cases (see the first transition in Fig. 5 for details). The stemming information for *przetwarzanie(u)* gives two options: it is either a noun *przetwarzanie* (Eng. *processing*) or a verbal noun (tagged by Morfologik as *ger*) derived from the verb *przetwarzać* (Eng. *to process*).

Hence, by performing stem/tags matching three transitions are created (indicated by arrows in Fig. 4): one for places corresponding to the mapping *informacji* → *informacji* and two for *przetwarzaniu* → *przetwarzanie*. XML code for the resulting rule is presented in Fig. 5.

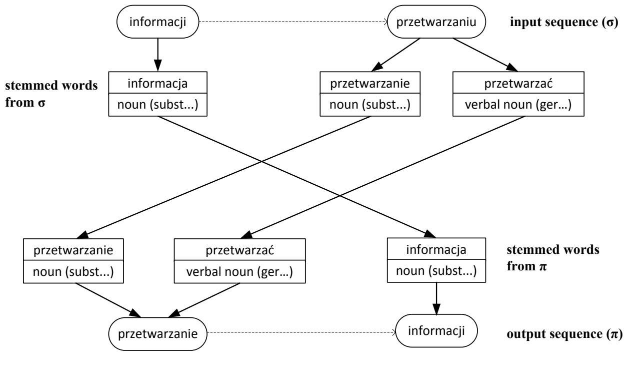


Fig. 4. Translation of annotation `@(informacji przetwarzaniu = @przetwarzanie informacji)`

It is expected that rules obtained as results of the annotations compilation satisfy the following conditions:

- Each output place is a target of a transition: $\forall o \in I . \exists t = (i_s, o_e) \in T \colon o = e_e$
- All transitions leaving an input place must target the same output place: $\forall t_1 = (i_1, o_1), t_2 = (i_2, o_2) \in T . i_1 = i_2 \rightarrow o_1 = o_2$

If the compilation returns a rule, which does not satisfy the above conditions, the source annotation is reported as ambiguous. Obviously, with the presented translation algorithm it is not possible to compile such annotation as `@(danym danym = dane @dane)` although it can be quite easily interpreted as *given data*.

To our consolation, statistical translation functions of Google Translate (as of 2015) also have problems in this case: *danym danym* is translated to *given the*, wheras *dane dane* to *data data*. However, after specifying the input n-gram explicitly, i.e entering *“dane dane”* the correct form *given data* is returned.

```
<rule id="r:13" weight="1.0">
  <source>@( informacji przetwarzaniu  = @przetwarzanie informacji ) @line:34</source>
  <inputPlaces>
    <inputplace isExact="false" ord="0">
      <transitions>
        <transition target="op:30" isGuard="false" weight="1.0">
          <inputTags>
            <intag>subst:pl:gen:f</intag>
            <intag>subst:sg:gen:f</intag>
            <intag>subst:sg:dat:f</intag>
            <intag>subst:sg:loc:f</intag>
          </inputTags>
          <outputTags>
            <outtag>subst:pl:gen:f</outtag>
            <outtag>subst:sg:gen:f</outtag>
            <outtag>subst:sg:dat:f</outtag>
            <outtag>subst:sg:loc:f</outtag>
          </outputTags>
        </transition>
      </transitions>
    </inputplace>
    <inputplace isExact="false" ord="1">
      <transitions>
        <transition target="op:29" isGuard="false" weight="1.0">
          <inputTags>
            <intag>subst:sg:loc:n2</intag>
            <intag>subst:sg:dat:n2</intag>
          </inputTags>
          <outputTags>
            <outtag>subst:sg:nom:n2</outtag>
            <outtag>subst:sg:acc:n2</outtag>
            <outtag>subst:sg:voc:n2</outtag>
          </outputTags>
        </transition>
        <transition target="op:29" isGuard="false" weight="1.0">
          <inputTags>
            <intag>ger:sg:dat.loc:n2:imperf:aff:refl.nonrefl</intag>
          </inputTags>
          <outputTags>
            <outtag>ger:sg:nom.acc:n2:imperf:aff:refl.nonrefl</outtag>
          </outputTags>
        </transition>
      </transitions>
    </inputplace>
  </inputPlaces>
  <outputPlaces>
    <outputplace id="op:29" ord="0" isKey="true"/>
    <outputplace id="op:30" ord="1" isKey="false"/>
  </outputPlaces>
</rule>
```

Fig. 5. A rule in XML format

### *B. Execution of rules*

As the described rules (see Definition 1) follow the Petri net approach, to define their execution such notions as tokens and marking are necessary. The set of tokens $TK$ is defined as $TK = IF \times S \times \mathcal{P} \times \mathbb{R}^{0+}$. (See also formula 1.) Components of a token tuple $(if, s, p, w) \in TK$ are the following: $if$ denotes an inflected form, $s$ is a stem, $p$ is a part of speech tag and $w$ is a non-negative weight.

**Definition 2** (Marking)**.** *Marking* for a rule $R = (\mathcal{P}, I, O, T, itag, otag, itoken, \mu)$ is defined as a function that assigns sets of tokens to places $M : I \cup O \rightarrow 2^{TO}$

Before executing rules, each word in an input sequence $\pi = (\pi_i)$ is submitted to Morfologik $stem$ function (see formula 2) and corresponding sequence of stemming information is obtained $\rho = (\rho_i)$, where $\rho_i \in 2^{S \times \mathcal{P}}$. Hence, $\pi_i$ is an input inflected form of a word and $\rho_i$ a set of possible stem–tag combinations for $\pi_i$.

Execution of a rule comprises the following steps:

1) Input places are filled with tokens starting from position $b$ in sequences $\pi$ and $\rho$ ($b$ defines the beginning of the sliding window). Each $k$-th input place receives tokens from $\pi_{b+k}$ and $\rho_{b+k}$. Its marking is gets: $M(I_k) \leftarrow \{\pi_{b+k}\} \times \rho_{b+k} \times \{0\}$.
2) For each input place $I_k \in I$ it is checked, if there exists at least one enabled transition (or exact word matching, if specified). Hence, for each transition $t \in T(I_k)$ and each token $tk = (if, s, p, w) \in M(I_k)$ it is checked, if the sets of transition input tags $itag(t)$ and token tags $p$ match. The matching function compares tags in both sets, splitting them into smaller parts where needed. A pair $(t, tk)$, such that the transition $t$ is enabled is called an enabled binding.
3) Finally, transitions are executed for all enabled bindings $(t, tk)$. If $O_i$ is the output place for the transition $t$, and $tk = (if, s, p, w)$, then the marking for $O_i$ is modified according to (4):

$$M(O_i) \leftarrow M(O_i) \cup synth(s, otag(t)) \times \{s\} \times \\ \times (p \cap otag(t)) \times weight(t, tk) \quad (4)$$

Interpretation of the formula (4) is the following: a stem (lemma) together with output tags are submitted to the $synth$ function, which returns a set of inflected forms. In consequence, for an input word $if_{in}$ the chain of translations is accomplished:

$$if_{in} \xrightarrow{p_i n \cap itag(t)} \{s\} \xrightarrow{p \cap otag(t)} \{if_{out}\}$$

In the first step $if_{in}$ is converted into a set of stems $\{s\}$. In the second stems are converted back to altered inflected forms $\{if_{out}\}$ according to the set of output tags assigned to the transition.

The $weight(t, tk)$ function is used to assign weight value to tokens. It calculates the Jaccard index of two sets: token tags $p$ and transition input tags $itag(t)$ and then multiplies it by the transition weight $\mu(t)$.

$$weight(t, tk) = \frac{|p \cap itag(t)|}{|p \cup itag(t)|} \cdot \mu(t) \quad (5)$$

Weight calculation according to formula (5) is based on a certain intuition: the Jaccard index allows assessing the similarity between a prototype word appearing in the annotation, from which the rule originates and the input term.

After executing a rule, its output places may contain several tokens. Such situation is shown in Fig 3, where the place *out#1* contains two tokens, the place *out#2* three and the place *out#3* one. For the presented example six separate strings representing concept candidates, each of them having different content and weights are obtained.

### C. Aggregation

Aggregation of results constitutes the final stage of text processing and concepts extraction. Actually, it is rather a chain of aggregations performed at subsequent levels (see Fig. 6). All data processed within the chain are attributed with numerical weights (scores). At each step, apart from data conversions (e.g. combining tokens into strings) the assigned weights are summarized using various norms: $sum$,$min$, $max$, etc.

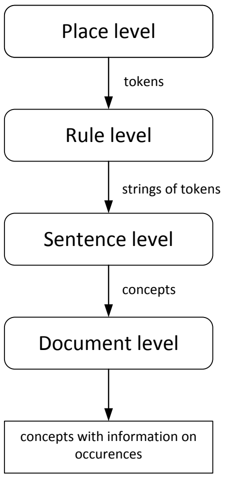


Fig. 6. The chain of aggregations

- During the aggregation occurring at the *place level* identical tokens assigned to an output places combined into one. The default norm is $sum$.
- Aggregation at the *rule level* aims at creating strings from sets of tokens residing in output places. In the same time weights tokens are aggregated and obtained values are assigned to strings. The default norm is $min$.
- Various rules may return identical concepts. At present there are no measures implemented that would allow to manage the ruleset, and in particular remove redundant rules [13]. Aggregation at the *sentence level* combines multiple strings using the selected norm (default is $max$).
- Finally, aggregation at the *document level* keeps track of occurrences of concepts in sentences, counts them and aggregates the weights. (The default norm is $sum$).

### D. Implementation

The software implementing the discussed approach was implemented in Java language, what enables integration with Morfologik stemming libraries and Jena for OWL output handling. It includes the following modules:

- A tool allowing to load the Morfologik dictionary into PostgreSQL database (see Section IV)

- Synthesizer shortly described in Section IV
- Annotations compilation tool, which create rules and serializes them with JAXB libraries
- Sentence scanner allowing to split the input text into sentences and perform additional preprocessing
- Rule execution modules
- Configurable aggregation modules, e.g. it is possible to select various aggregation norms
- Output modules that renders results in various formats including CSV and OWL

## VI. Experiments and results

In this section we present initial results of concept extraction from a large text file specifying medical guidelines for asthma treatment. The file is a Polish translation of a document issued by Global Initiative for Asthma (GINA) in 2011. The file size is 308KB it contains about 40000 words and 2000 sentences.

We selected this document, as it was used in previous works aiming at building ontologies of medical guidelines and performing fuzzy reasoning based on ontological models [24]. Moreover, the presented here solution was intended to be an improvement of a prototype tool (unpublished) that used internally FSM to extract concepts from texts. Both tools were tested on the same file and we wanted to compare the results of the described method and the previous one (available at http://home.agh.edu.pl/~pszwed/en/doku.php?id=ontologies#in_polishgina_guidelines_glossary). The main difference with respect to the previous solution is the focus on correct morphological form of concept names and different approach to evaluation of the accuracy of extracted concepts .

The ruleset used during the experiment resulted from compilation of an annotations file comprising translation patterns for single words and pairs of words only. Single word translations were limited to nouns and verbal nouns, 2-grams included variations comprising nouns, verbal nouns and adjectives in various cases. The resulting ruleset contained 223 rules.

The average processing time (concepts extraction without rules compilation) was about 36.785 seconds (executed on Intel Core i7-2675QM laptop at 2.20 GHz, 8GB memory under Windows 7). The extraction process returned 5151 concepts.

Table III shows selected results ordered by the weight values. It comprises top 20 results, 10 concepts from the middle and 10 concepts with the lowest weights. The word in capital letter indicates the key, i.e. the part of the concept name referring to a prospective superclass. Improperly identified concepts are underlined.

In the whole table the entry 2495 is apparently incorrect according to syntactical rules. The correct form should be: *EKONOMIKA_wdrażania*. Other underlined entries are formed correctly according to grammar rules, however their semantics does not match the document content. Position 18 WIEKO (Eng. lid) should be replaced by WIEK (Eng. age). The word NIEODPOWIEDNI (Eng. inaccurate) is rather an adjective. However, according to Morfologik it can be classified both as adjective and noun. The word SŁUŻĄCY can be interpreted as “a servant” and “serving to”. In this context rather the second meaning was used over the document.

TABLE III
Extracted concepts from the Polish translation of asthma treatment medical guideline

| Pos | Concept | Count | weight |
|---|---|---|---|
| 1 | DROGI_oddechowe | 161 | 644.00 |
| 2 | LECZENIE | 487 | 496.35 |
| 3 | LECZENIE_astmy | 87 | 348.00 |
| 4 | ASTMA | 965 | 289.50 |
| 5 | ZAOSTRZENIE_astmy | 42 | 126.00 |
| 6 | LECZENIE_zaostrzenia | 23 | 92.00 |
| 7 | GLIKOKORTYKOSTEROID_wziewny | 46 | 92.00 |
| 8 | ZAOSTRZENIE | 103 | 90.90 |
| 9 | BADANIE | 87 | 90.75 |
| 10 | ROZPOZNANIE_astmy | 30 | 90.00 |
| 11 | RYZYKO | 98 | 87.30 |
| 12 | POSTĘPOWANIE | 73 | 70.20 |
| 13 | ZATOKI_przynosowe | 17 | 68.00 |
| 14 | ciężkie_ZAOSTRZENIE | 22 | 66.00 |
| 15 | PRZEPŁYW_powietrza | 31 | 62.00 |
| 16 | WYSTĘPOWANIE | 65 | 61.50 |
| 17 | ROZPOZNANIE | 68 | 60.30 |
| 18 | WIEKO | 85 | 59.85 |
| 19 | STOSOWANIE | 192 | 57.60 |
| 20 | GRUPY_wiekowe | 14 | 56.00 |
| 2488 | szkodliwa_CZĄSTKA | 1 | 1.00 |
| 2489 | DZIAŁANIE_glikokortykosteroidu | 1 | 1.00 |
| 2490 | ZAKRESY_zużycia | 1 | 1.00 |
| 2491 | JAMA_nosowa | 1 | 1.00 |
| 2492 | źródłowy_DOKUMENT | 1 | 1.00 |
| 2493 | nowoczesny_LEK | 1 | 1.00 |
| 2494 | charakterystyczna_CECHA | 1 | 1.00 |
| 2495 | EKONOMIKI_wdrażania | 1 | 1.00 |
| 2496 | OCENA_lekarska | 1 | 1.00 |
| 2497 | PRZETWÓRNIA_ryb | 1 | 1.00 |
| 5141 | MOC | 1 | 0.04 |
| 5142 | ODPOWIEDZIALNOŚĆ | 1 | 0.04 |
| 5143 | OKOLICZNOŚĆ | 1 | 0.04 |
| 5144 | NIEODPOWIEDNI | 1 | 0.04 |
| 5145 | POŚCIEL | 1 | 0.04 |
| 5146 | MNIEJSZOŚĆ | 1 | 0.04 |
| 5147 | PRACUJĄCY | 1 | 0.04 |
| 5148 | CAŁOŚĆ | 1 | 0.04 |
| 5149 | SŁUŻĄCY | 1 | 0.04 |
| 5150 | WPŁYW_CFC | 1 | 0.04 |
| 5151 | BETA | 1 | 0.01 |

Fig. 7 shows extracted concepts related to the term *ryzyko* (Eng. risk) arranged into a small hierarchy. The term repeated quite frequently over the document. It can be seen that in two cases 3-grams would be more adequate, because some complements are missing, e.g. *RYZYKO_utraty* (Eng. risk of loosing).

The presented software is still in development phase, hence only preliminary results of concept extraction can be given. During future experiments several parameters controlling the execution should be tuned to provide high accuracy ratio.

The configurable parameters include norms used at various stages of aggregation (see Section V-C), as well as weights assigned to rules and individual transitions. Taking into account the complexity of rule definitions (see Fig. 5) and the number of rules, weights tuning must be performed in an automated way.

We already started to implement such mechanisms. During the experiment described in Section VI rules with single word transformation patterns were attributed with smaller weights (0.3). This value was selected quite arbitrarily.

Another example is related to handling a situation described as *tag conflict*. Quite often Morfologik identifies a word a member of two various part of speech classes, e.g. an adjective vs. a noun or a verbal noun vs. a noun. In this case, while compiling the annotations, we attribute higher weights to transitions derived from adjectives and verbal nouns (0.9), than nouns (0.1).

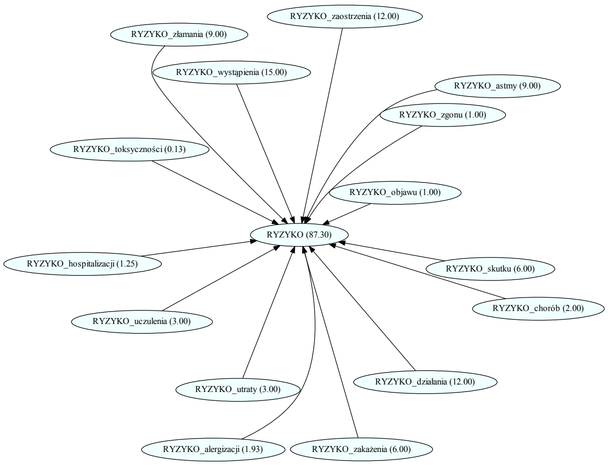


Fig. 7. Concepts related to the term *ryzyko* (risk)

## VII. Conclusions

In this paper we present a solution for concept extraction from Polish texts with special focus on correct morphological forms of obtained concept names. As Polish is a highly inflected language, detected names need to be transformed following Polish grammar rules. We propose a user-friendly method for specification of transformation patterns, which is based on a simple annotations language. Annotations prepared by a user are compiled into transformation rules. During the concept extraction process the input document is split into sentences and the rules are applied to sequences of words comprised in sentences. Recognized strings forming concept names are aggregated as various levels. Although the annotation language is simple, it is flexible enough to specify various translation patterns, for example it is possible to apply inversion or to omit such tokens appearing in the input sequence as commas or prepositions.

The design of rules follow the language of Petri nets, combining elements of colored [12] and fuzzy [6] Petri nets. Rules comprise input places (filled with elements of analyzed n-grams), output places, where results are collected, and transitions linking them. Similarly to colored Petri nets, tokens are tuples – in this case they represent words and part of speech information. Tokens are also equipped with fuzzy weights. The internal behavior exhibited during rules execution borrows from fuzzy Petri nets (i.e. there is no conflict between transitions that may fire simultaneously and produce multiple tokens in output places). A similar approach was used in our previous works related to semantic event recognition and it turned out to be very efficient [26], [25].

Although we did not provide the formalization in the flavor of fuzzy rules and Fuzzy Inference Systems and fuzzy rules [22], [15], the discussed approach follows this direction. A token put in an input place receives a weight based on a kind of membership function (see Section V-B, step 3). This step resembles the fuzzification stage in fuzzy inference systems. Further, after a rule is executed, weights of output token are combined using configurable aggregation norms. This in turn corresponds to defuzzification step. Nevertheless, the transformation rules are not semantic, they are not prepared by experts and do not reference linguistic terms in the sense of fuzzy logic.

The rules are obtained fully automatically. This point also differs the presented approach from the Language tool [18] discussed in Section. II.

The developed software is tightly coupled with the Morfologik dictionary for the Polish language and accompanying library API. However, the proposed approach is quite general and can be applied to texts in other languages, provided that a language specific dictionary, as well as a software supporting lemmatization and synthesizing are available.

There are several avenues for future work. We plan to define 3-gram translation patterns and patterns comprising verbs, .e.g. *he* reads a book → book reading. Further improvement to the extraction accuracy can be achieved by tuning several parameters: rule and transition weights and norms. Another direction may be related to rule management, removing redundant rules, refactoring (e.g. splitting transitions) and purging tag sets assigned to transitions.

## References

[1] S. Acedański, "A morphosyntactic brill tagger for inflectional languages," in *Advances in Natural Language Processing*. Springer, 2010, pp. 3–14.

[2] C. Blake and W. Pratt, "Better rules, fewer features: a semantic approach to selecting features from text," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, 2001, pp. 59–66.

[3] S. Bloehdorn, P. Cimiano, and A. Hotho, "Learning ontologies to improve text clustering and classification," in *From Data and Information Analysis to Knowledge Engineering*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul, Eds. Springer Berlin Heidelberg, 2006, pp. 334–341. [Online]. Available: http://dx.doi.org/10.1007/3-540-31314-1_40

[4] C. Carpineto and G. Romano, *Concept data analysis: Theory and applications*. John Wiley & Sons, 2004.

[5] J. Challis, "Lateral thinking in information retrieval white paper," Concept Searching, Tech. Rep., 2003.

[6] S.-M. Chen, J.-s. Ke, and J.-F. Chang, "Knowledge representation using fuzzy petri nets," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 2, no. 3, pp. 311–319, Sep 1990.

[7] P. Cimiano, A. Hotho, and S. Staab, "Learning concept hierarchies from text corpora using formal concept analysis." *J. Artif. Intell. Res.(JAIR)*, vol. 24, pp. 305–339, 2005.

[8] J. Daciuk, "Incremental construction of finite-state automata and transducers, and their use in the natural language processing," Ph.D. dissertation, Gdansk University of Technology, ETI faculty, Gabriela Narutowicza 11/12, 80-233 Gdansk Poland, 1998.

[9] N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu, "A web of concepts," in *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2009, pp. 1–12.

[10] F. Graliński, K. Jassem, and M. Junczys-Dowmunt, "Psi-toolkit: A natural language processing pipeline," in *Computational Linguistics*, ser. Studies in Computational Intelligence, A. Przepiórkowski, M. Piasecki, K. Jassem, and P. Fuglewicz, Eds. Springer Berlin Heidelberg, 2013, vol. 458, pp. 27–39. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-34399-5_2

[11] D. Janus, "Smyrna prosty konkordancer obsługujący język polski," 2015, accessed: May 2015. [Online]. Available: http://smyrna.danieljanus.pl/

[12] K. Jensen, *Coloured Petri Nets: Basic Concepts, Analysis Methods and Practical Use*. Springer, 1996, vol. 1, no. Basic Concepts.

[13] A. Ligeza, *Principles of Verification of Rule-Based Systems*. Springer, 2006.

[14] A. Maedche and S. Staab, "Ontology learning for the semantic web," *Intelligent Systems, IEEE*, vol. 16, no. 2, pp. 72–79, Mar 2001.

[15] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of ManMachine Studies*, vol. 7, no. 1, pp. 1–13, 1975. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0020737375800022

[16] M. Miłkowski, "Developing an open-source, rule-based proofreading tool," *Software: Practice and Experience*, vol. 40, no. 7, pp. 543–566, 2010.

[17] ——, "Morfologik," 2015, accessed: May 2015. [Online]. Available: http://morfologik.blogspot.com/

[18] D. Naber, "Language tool style and grammar check," 2015, accessed: May 2015. [Online]. Available: https://www.languagetool.org/

[19] S. Osinski and D. Weiss, "A concept-driven algorithm for clustering search results," *Intelligent Systems, IEEE*, vol. 20, no. 3, pp. 48–54, 2005.

[20] A. Parameswaran, H. Garcia-Molina, and A. Rajaraman, "Towards the web of concepts: Extracting concepts from large datasets," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 566–577, 2010.

[21] P. Pęzik, "Wyszukiwarka PELCRA dla danych NKJP," 2012.

[22] T. Ross, *Fuzzy Logic with Engineering Applications*. Wiley, 2009.

[23] A. Stavrianou, P. Andritsos, and N. Nicoloyannis, "Overview and semantic issues of text mining," *ACM Sigmod Record*, vol. 36, no. 3, pp. 23–34, 2007.

[24] P. Szwed, "Application of fuzzy ontological reasoning in an implementation of medical guidelines," in *Human System Interaction (HSI), 2013 The 6th International Conference on*, June 2013, pp. 342–349.

[25] ——, "Video event recognition with Fuzzy Semantic Petri Nets," in *Man-Machine Interactions 3*, ser. Advances in Intelligent Systems and Computing, A. Gruca, T. Czachórski, and S. Kozielski, Eds. Springer International Publishing, 2014, vol. 242, pp. 431–439. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-02309-0_47

[26] P. Szwed and M. Komorkiewicz, "Object tracking and video event recognition with fuzzy semantic petri nets," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, Kraków, Poland, September 8-11, 2013.*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2013, pp. 167–174. [Online]. Available: http://fedcsis.org/2013/

[27] M. Wolinski, M. Milkowski, M. Ogrodniczuk, and A. Przepiórkowski, "Polimorf: a (not so) new open morphological dictionary for polish." in *LREC*, 2012, pp. 860–864.

# AAIA'15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene

AAIA'15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene is a continuation of the last year's competition organized within the framework of International Symposium on Advances in Artificial Intelligence and Applications (AAIA'15, https://fedcsis.org/aaia). It is also an integral part of the 2nd Complex Events and Information Modelling workshop (CEIM'15 https://fedcsis.org/ceim) devoted to the fire protection engeneering. This time, the task is related to the problem of recognizing activities carried out by firefighters based on streams of information from body sensor networks. Prizes worth over 4,000 PLN will be awarded to the most successful teams. The contest is sponsored by Polish Information Processing Society (http://www.pti.org.pl/), with a support from University of Warsaw (http://www.mimuw.edu.pl/) and ICRA project (http://icra-project.org/).

## Introduction

A fire ground is considered to be one of the most challenging decision taking environment. In dynamically changing situations, such as those occurring at a fire scene, all decisions need to be taken in a very short time. Since wrong decisions might have severe consequences, a commander of the response team is forced to act under a huge psychological pressure. This fact, combined with incomplete or inaccurate information about the current situation, sometimes leads to committing serious mistakes [1].

There are several initiatives that investigate this complex problem. One of them is The National Near Miss program in the USA (www.nationalnearmiss.org/). It gathers and analyzes reports describing real-life dangerous situations, and tries to draw some conclusions regarding their causes. Based on several thousand of carefully analyzed reports, experts identified the “lack of situational awareness” as the main factor associated with major accidents among firefighters [2]. This observation is in accordance with the results of the previous edition of our data mining competition [3]. The situational awareness corresponds to the cautiousness of a commander and his understanding of the actual state of the environment. Conditions affecting the situational awareness can be broken down into three groups: a lack of information, a lack of knowledge and a lack of cognition [4]. In this context, it seems that an increase in the situational awareness of commanders would result in taking better decision and thus increasing the safety of firefighters. Studies on the causes for mortal accidents during the actions of firefighters were also conducted by the Department of Homeland Security of the United States [5]. One conclusion of their research is that over 43% of deaths at a fire scene was caused by the stress or overexertion. Therefore, another critical way of increasing the firefighter safety is by monitoring their kinematics and psychophysical condition during the course of fire & rescue actions.

Our research team works on those problems within a frame of ICRA project (www.icra-project.org/). One of prototype tools developed as a result of this project is, so called, a "smart jacket". This device is a wearable set of body sensors that allows to automatically track a firefighter at a fire scene. It also enables real-time screening of firefighter’s vital functions and monitoring of ongoing activities at the scene. The later of those two tasks is the main scope of this year’s AAIA Data Mining Competition. We would like to ask participants to come up with efficient algorithms for labelling activities conducted by firefighters during their training exercises, based on provided data sets from our body sensor network. More details regarding the data and their acquisition process can be found in Task description section. Moreover, a good starting point for research on the problem of activity recognition based on body sensor data was described in [6]. We hope that your expertise and innovative ideas will become a valuable contribution in our effort to increase the safety of brave men and women serving in Polish State Fire Service.

## Special session at CEIM'15

As in the previous year, a special session devoted to the competition will be held at 2nd Complex Events and Information Modelling workshop (CEIM'15 https://fedcsis.org/ceim) which is a part of 10th International Symposium on Advances in Artificial Intelligence and Applications (AAIA'15, https://fedcsis.org/aaia). We will invite authors of selected reports to extend them for publication in the conference proceedings (after reviews by Organizing Committee members) and presentation at the conference. The publications will be treated as regular papers (including indexation in the Thomson Reuters Web of Science, DBLP, Scopus and other portals). The invited teams will be chosen based on their final rank, innovativeness of their approach and quality of the submitted report.

## Awards

Authors of the top ranked solutions (based on the final evaluation scores) will be awarded with prizes:

- First Prize: 3000 PLN + one free FedCSIS'15 conference registration,
- Second Prize: 1000 PLN + one free FedCSIS'15 conference registration,
- Third Prize: one free FedCSIS'15 conference registration.

The award ceremony will take place during the FedCSIS'15 conference (September 13 - 16, Łódź). Traditionally, invited authors who decide to attend the conference will receive a diploma and a competition T-shirt.

## Contest Organizing Committee

**Andrzej Janusz,** University of Warsaw.
**Michał Meina,** University of Warsaw
**Adam Krasuski,** Main School of Fire Service
**Krzysztof Rykaczewski,** University of Warsaw
**Bartosz Celmer,** Main School of Fire Service
**Dominik Ślęzak,** University of Warsaw & Infobright Inc.

## References

[1] A. Krasuski: “A framework for Dynamic Analytical Risk Management at the emergency scene. From tribal to top down in the risk management maturity model”, *FedCSIS 2014,* pp. 323-330.

[2] L. J. Grorud and D. Smith: “The National Fire Fighter Near-Miss Reporting. Annual Report 2008”, in *An exclusive supplement to Fire & Rescue magazine.* Elsevier Public Safety, 2008, pp. 1–24.

[3] A. Janusz, A. Krasuski, S. Stawicki, M. Rosiak, D. Ślęzak, H. S. Nguyen: “Key Risk Factors for Polish State Fire Service: a Data Mining Competition at Knowledge Pit”, *FedCSIS 2014,* pp. 345-354.

[4] A. Krasuski, A. Jankowski, A. Skowron, and D. Ślęzak: “From sensory data to decision making: A perspective on supporting a fire commander”, in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT),* 2013 IEEE/WIC/ACM International Joint Conferences on, vol. 3. IEEE, 2013, pp. 229–236.

[5] United States Fire Administration: “Annual report on firefighter fatalities in the United States”, http://apps.usfa.fema.gov/firefighter-fatalities/.

[6] M. Meina, B. Celmer, K. Rykaczewski: “Towards Robust Framework for On-line Human Activity Reporting Using Accelerometer Readings”, *AMT 2014,* pp. 347-358.

DOI: 10.15439/2015F426
ACSIS, Vol. 5

# Tagging Firefighter Activities at the Emergency Scene: Summary of AAIA'15 Data Mining Competition at Knowledge Pit

Michał Meina, Andrzej Janusz,
Krzysztof Rykaczewski
Institute of Mathematics
University of Warsaw
Warsaw, Poland
{mich,janusza,krykaczewski}@mimuw.edu.pl

Dominik Ślęzak
Institute of Mathematics
University of Warsaw
and
Infobright Inc.
Warsaw, Poland
slezak@mimuw.edu.pl

Bartosz Celmer and Adam Krasuski
Section of Computer Science
The Main School of Fire Service
Warsaw, Poland
bart.celmer@gmail.com
krasuski@inf.sgsp.edu.pl

***Abstract*—In this paper, we summarize AAIA'15 data mining competition: Tagging Firefighter Activities at a Fire Scene, which was held between March 9 and July 6, 2015. We describe the scope and background of the competition. We also reveal details regarding the data set used in the competition, which was collected and tagged specifically for the purpose of this data challenge. We explain the data acquisition process which involved using a body sensor network system consisting of several inertial measurement units and a physiological data sensor. Finally, we briefly discuss submitted results with respect to their possible real-life application in our decision support system.**

## I. Introduction

The emergency scene is considered to be one of the most dangerous and stressful working environments [3]. Each year the numbers of firefighters die or are injured during the fire & rescue operations. There are many contributing factors which lead to the unsafe events. Those factors are investigated by special commissions and according to [10] 25% of accidents are caused by bad decision making, 29% by bad situational awareness and 10% by bad communication.

The key aspect (regardless of the type of the Incident Management System) comes down to modelling of perception and evaluation of the emergency scene by the Incident Commander (IC) [5], [9], [6]. The relevant incident assessment methodology increases the safety of the rescuers and the chances for success. Information plays a pivotal role in the perception and evaluation. Therefore, there is a strong demand on increasing the sensor density in order to provide more information to the IC. The information reported to the IC must satisfy the *information triangle* rule [5]. It means that it should be *relevant, accurate* and *timely*. All of those aspects — in dynamically changing environment such as a fire incident is — is difficult to satisfy.

During emergency firefighters are mostly concentrated on their principal activities. For the safety reasons they are not able to share in real-time with the IC the information about the current conditions. Paradoxically, as described above, activities currently performed by the firefighters, their physical condition as well as temporal and spatial relation to the fire are crucial for the IC in order to ensure safety operating of the firefighters.

Since the information at the emergency scene is crucial for the operating safety and the personal real-time reporting from firefighters are hampered, the new ways of indirect communication and information sharing are needed. The computer-based systems for human activity recognition may help to reduce the unsafe events, improving the communication and increasing the efficacy of the incident management. Moreover, due to the dangerous environment for human, it could be expected that in the future more and more robot will be involved in operating at the emergency scene. The human activity recognition system may create a bridge for human robot cooperation at the scene.

Human activity recognition using Body Sensor Networks (BSN) is a non-invasive system that is able to deliver information about person locomotion patterns, current posture and specific action performed. In the system that is considered in this paper, a network of inertial measurement units are used. They are able to gather kinematic (motion) data from different parts of the body. This information is then processed using classification techniques [16] (by the body-worn computation unit) in order to estimate activity and passed via radio link to the IC. In such systems like described training data needs to be prepared beforehand and classification procedure needs to take into account power consumption on mobile processing unit and low-radio link throughput.

This paper describes a data mining competition which was organized at the Knowledge Pit platform [14]. Special dataset collected during training exercises of firefighters cadets using custom-built BSN was prepared and made available publicly to encourage a research community to work on new methods and classification models in this particular application. The competition was a part of a broader research addressing a sensory data acquisition at the incident place. The submitted

The research was partially supported by the Polish National Science Centre - grant 2012/05/B/ST6/03215 and by the Polish National Centre for Research and Development (NCBiR) — grant No. O ROB/0010/03/001 in the frame of Defence and Security Programmes and Projects: "Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in the buildings".

results forms very valuable input into the field and forms a comparable platform for future research.

The paper starts from a detailed discussion of the data acquisition process, including our hardware setup, collection of the data and the process of tagging. The third section describes the preprocessing of the data set for the competition and the evaluation scenario of solutions submitted by competitors. This section also includes a brief summary of the competition results. The paper ends with concluding remarks and a draft of our plans for future work.

## II. Data Acquisition

This section describes data acquisition part of the competition. In the first paragraph we will discuss our hardware design and processing challenges connected with on-line processing the data. Second paragraph will summarize recording sessions and the final paragraph will cover process of annotating the data set.

### A. Hardware setup

The data acquisition hardware is consisted of sensors (seven IMUs and physiological data sensor) and Data Acquisition Unit (DAU). In order to overcome clock synchronization issues all IMUs are connected via physical link to real-time embedded system that streams the data to DAU. Physiological sensor has it own clock synchronization system, therefore information can be streamed via Bluetooth. Moreover, DAU provides simple web interface (accessible via Wi-Fi connection) to control recording trials (starting, stopping and monitoring data collection process).

In our setup we use:

- 7×IMU (Polulu AltIMU-9 rev-4) — 3-axis accelerometer ($\pm 16\,g$ dynamic range) and gyroscope ($\pm 2000\,°/s$ maximum angular rate) with 16 bit signed integer resolution, two IMUs for legs, two for hands and arms and one for back,
- Real-time system embedded — based on Arduino Micro prototype platform, connected via USB interface to DAU,
- Data Acquisition Unit — based on Odroid-U3+ with external battery, additional Bluetooth and Wi-Fi module,
- Physiological data sensor — Equivital EQ02 SEM,
- Communication nodes — XBee-PRO® 868.

In a low level point of view, IMU are two different sensors, one for measuring accelerations one for angular rates. Both, accelerometer and gyroscope provide 16 bit signed integer for each axis (horizontal, vertical and altitudinal), and communicate via $I^2C$ bus. In order to provide the data with zero-time offset between different IMUs, we used real time system (microcontroller) and we have prepared parallel communication version of $I^2C$ library. The library enables us to read the sensor readings from all IMUs at the same time with 1000 samples per second. We decided, however, to gather the data with the frequency about 200 Hz which is a sufficient for classification purposes, while providing low noise and enough accuracy for the most of the sudden movements.

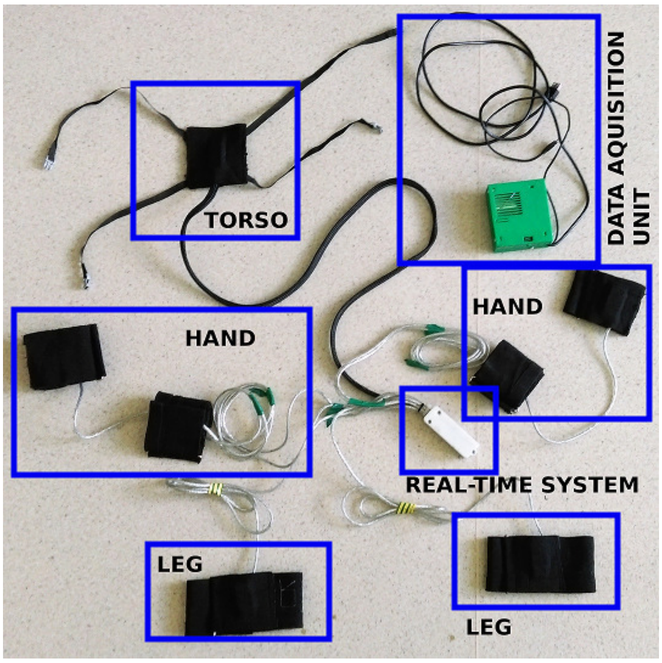


Fig. 1. Hardware setup that is able to gather motion data from different parts of the body. IMUs are kept in a custom 3D printed casing and mounted to the body using elastic bands.

Physiological data (ECG waveform, heart rate, breathing rate and skin temperature) was recorded by Equivital[1] on internal storage and redundantly sent via Bluetooth to DAU.

During recording session the raw data was stored on DAU using binary format. Single sample, consist of readings from all sensor which sums up to 680 bits ($16$ bit $\cdot\ 2{\cdot}7$ sensors $\cdot\ 3$ axis $+\ 8$ bit time stamp). In the real application, however, the minimum required bandwidth for the data transmission is higher because of network maintaining routines overhead, so final minimal bandwidth was more than $49.8\,kB/s$ for each subject. Reliable transmission of the raw data through radio for longer distances (1–2 km) while maintaining reasonable power consumption is very challenging or almost impossible task. LTE or Wi-Fi connections enables us to stream such amount of data, but while using these technologies problems with the power consumption arise (see [13], [7]). Although low frequency radios have low bandwidth, the power consumption is sufficient — this setup, however, imposes the need of processing the data locally and transmission partially preprocessed data.

Most intuitive and robust design is that the data is processed on DAU and only the classification result should are sent to remote server. On the other hand there is no implementation of more sophisticated classification system dedicated to work on embedded system (due to floating point precision and lack of processing power). Therefore, we have used Odroid-U3+ platform based on quad-core Cortex-A9. This platform has a two scaling frequencies options and possibility of choosing the

[1] http://www.equivital.co.uk/

number of cores being used. This setup gives us ability to scale power consumption depending on classification algorithm.

### B. Training session recording

Recording sessions took place from January till February 2015 at the Main School of Fire Service in Warsaw. Sixteen cadets were recorded during a training session of fire incident. There were two scenarios of the trial and each of them was repeated three times. However, we did not use breathing apparatus and real-life obstacles which changes significantly kinematics of a subject in real incidents.

Recording procedure was as follows: before session each sensor was calibrated on-place and checked for errors using gravitational force measurements in different sensor orientations. Two cadets participated in each session, whereas only one of them was wearing the sensors. Second cadet was later responsible for maintaining equipment on field. Those preparations lasted for 30 minutes. Each session was recorder using two cameras. Quality of the kinematic recordings was monitored with respect to throughput of the data using mobile device. At the beginning of each trial subject was asked to jump for three times to estimate precise time shift between video and kinematic data. Cadet performed two significantly different scenarios in three trials (each of them lasted 3–5 minutes), after that wearable sensor was transferred to the second cadet. At the end of the trial, the sensor on-body mounting was checked for displacement and the data was archived.

Recording of two cadets usually lasts for three hours and it requires commitment of three additional persons. After manual examination and verification of data quality we have chosen only eight out of fifteen sets of recordings. In the most of cases the reason for the elimination of a recording was either a sensor malfunction or a displacement of the sensor during the recorded training session.

### C. Tagging the data set

Tagging was prepared using video material on which acceleration from torso sensor was overlaid with precise time synchronization using marker (three jumps that was very easily observable in sensory data). Time series was divided into episodes of actions and there was always left a small gap between episodes in order not to introduce noisy transition phases (for example, deceleration from running to walking).

We decided to prepare two set of labels (tags), namely:

- **posture** which included: crawling, crouching, jumping, movement, standing, stooping,
- **activity** — few tags specifying short and specific actions. It included: carrying, carrying_hammer, carrying_nozzle, carrying_hose, hammer_striking, hose_throwing, ladder_down, ladder_up, manipulating, mounting_hose, nozzle_usage, running, searching, signal_hose_pullback, signal_water_first, signal_water_main, stairs_down_fast, stairs_up, stairs_up_fast, standing_up, starting, step, stopping, taking_equipment, taking_hammer, waiting, walking.

TABLE I. OCCURRENCES OF POSTURE TAG IN FINAL DATA SET.

| Posture | Number of occurrences | Total time [s] |
|---|---|---|
| crawling | 27 | 461.75 |
| crouching | 125 | 1109.0 |
| jumping | 96 | 242.75 |
| movement | 613 | 3083.0 |
| standing | 423 | 1524.25 |
| stooping | 267 | 645.0 |

TABLE II. MOS POPULAR TAGS IN SECOND GROUP.

| Activity | Number of occurrences | Total time [s] |
|---|---|---|
| ladder_down | 24 | 230.5 |
| ladder_up | 23 | 237.5 |
| manipulating | 482 | 1787.75 |
| no_action | 51 | 122.5 |
| nozzle_usage | 48 | 715.5 |
| running | 337 | 1800.25 |
| searching | 25 | 439.0 |
| signal_hose_pullback | 2 | 4.25 |
| signal_water_first | 33 | 63.0 |
| signal_water_main | 30 | 51.25 |
| signal_water_stop | 4 | 9.75 |
| stairs_down | 32 | 175.0 |
| stairs_up | 69 | 294.0 |
| starting | 49 | 119.5 |
| stopping | 47 | 123.25 |
| striking | 68 | 339.75 |
| taking_hammer | 47 | 68.75 |
| throwing_hose | 84 | 208.5 |
| walking | 94 | 273.25 |

However, not all of these tags were fully independent. For example, carrying always occurs after one of carrying_hammer, carrying_nozzle, carrying_hose. Therefore, not all combinations exist.

First tag in the activity labels specifies main action and other introduce detailed activities. The idea behind this was to show hierarchy of events.

### D. The acquired collection of data

The data set consist of 16 recording session (one for each cadet), while only 8 is most suitable for analysis. Each session is composed from kinematic data divided into six trials (three trials for two scenarios) accelerometer and gyroscope in three axes from seven mounting points described before (see Fig. 2 as example of such session). There are 3 065 664 records as a whole summing up to 4.25 h of data.

TABLE III. EIGHTEEN MOST POPULAR COMBINATION OF TAGS.

| Posture | Activities | Total time [s] |
|---|---|---|
| movement | running | 727.25 |
| movement | running carrying_hose carrying | 537.50 |
| crawling | searching | 439.00 |
| movement | running carrying_hammer carrying | 436.00 |
| crouching | nozzle_usage | 432.25 |
| crouching | manipulating | 389.00 |
| stooping | manipulating | 350.00 |
| standing | striking | 339.75 |
| crouching | manipulating mounting_hose | 261.75 |
| standing | nozzle_usage | 260.00 |
| standing | manipulating carrying_nozzle carrying | 252.50 |
| standing | manipulating | 248.50 |
| movement | ladder_up carrying_nozzle carrying | 237.50 |
| movement | ladder_down carrying_nozzle carrying | 197.75 |
| movement | stairs_down | 168.50 |
| stooping | throwing_hose | 158.75 |
| movement | stairs_up carrying_nozzle carrying | 137.25 |
| movement | stairs_up carrying_hammer carrying | 132.75 |

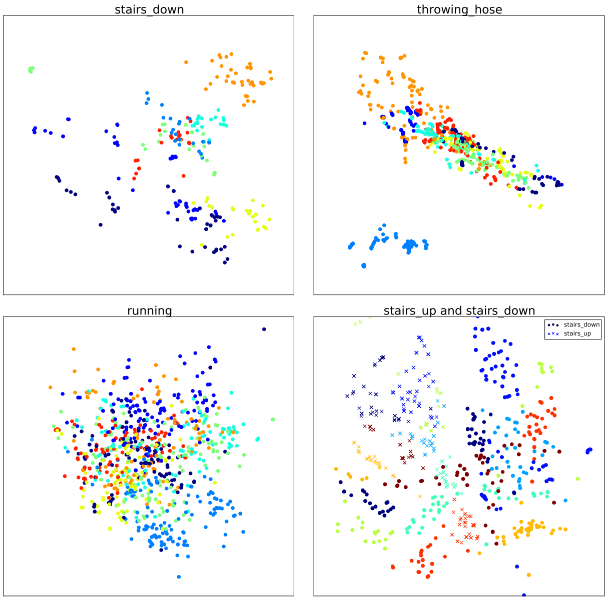


Fig. 3. MDS embedding for selected activities (colours denotes different test subjects). Notice that in the last figure dots denotes stairs_down tag, whereas crosses stairs_up.

Physiological data is composed from ECG waveforms from two leads filtered using proprietary algorithm. From this time series heart rate (estimated in the periods of 5 seconds) and R-R intervals are computed and enclosed into the data set. Additionally, breathing rate was estimated using elastic band and skin temperature are enclosed.

The main challenge for classification scenario in this particular data set is to generalize the same activities between different subjects. Certain physical activities can be performed in a very different way — this problem arises from number of reasons: starting from handedness, physical disposition, weight and height, etc. Those differences could be even so significant that they can be used for biometric authentication [8]. Moreover, sensor placement can be slightly different between session (although we put extra effort to ensure the same mounting places).

In Fig. 3 we depicted similarities of randomly selected $\sim 2\,s$ sequences of data. To be more precise, we see there multidimensional scaling on the set of features of the data, i.e. solution to the following minimization problem

$$\min_{x_1,\ldots,x_N\in\mathbb{R}^2} \sum_{1\leq i<j\leq N} \left(\|x_i - x_j\| - \|f_i - f_j\|\right)^2, \qquad (1)$$

where $f_i \in \mathbb{R}^K$ denotes feature vector of the $i$th window. The feature vector is composed using statistics commonly used in solutions described in section III-D.

The bottom-right plot of Fig. 3 illustrates that the certain activity can be more similar to different activity than the same activity of different subject. Of course this strictly depends on classification model but, all in all, this shows that physical activity classification needs to be performed with respect to subject identification. Another problem is illustrated on top-right plot of the same figure. In the dataset there was left-handed subject and certain activities (for example throwing_hose) was performed in very distinctive way.

## III. AAIA'15 Data Mining Competition

AAIA'15 Data Mining Competition took place between March 9, 2015 and June 5, 2015 at Knowledge Pit on-line platform. It was a continuation of the contest initiated during the previous edition of the data challenge associated with International Symposium on Advances in Artificial Intelligence and Applications (the AAIA conference series) [14]. This year's topic was related to real-time screening of firefighters' vital functions and monitoring of ongoing physical activities at the incident scene.

### A. *The task description*

The objective in this competition was to devise efficient methods for automatic labelling of short series of the sensory data with basic activities of a firefighter. On the one hand, this task was very challenging due to a fact that different people tend to perform the same activities in different ways. On the other hand, automatically generated and accurate activity labels would facilitate monitoring of firefighters' safety and contribute to development of efficient command supporting systems [13], [14].

For the purpose of the competition we provided the acquired data in a tabular format as two separate sets. The training data set contained 20,000 rows and 17,242 columns. Each row corresponded to a short time series (approximately $1.8\,s$ long) of sensory readings. Such a short time period was dictated by the applicability requirements and the available hardware setup (see Section II-A)

The first 42 columns represented aggregations of data from sensors monitoring firefighter's vital functions. The remaining columns were divided into 400 chunks that represented consecutive readings from the sets of kinetic sensors attached to firefighter's torso, hands, arms and legs (see the more detailed description in Section II). Therefore, a single chunk of columns consisted of 43 numeric values, from which the first one was time from the beginning of the series and the following 42 values represented the readings from the accelerometers (measured in $m/s^2$) and gyroscopes (measured in $deg/s$).

An average time difference between consecutive sensory readings in the data was $4.5\,ms$. Labels for the training data were provided in a separate file. Each row in this file contained two labels for a corresponding row in the training data. The first label described a posture of a firefighter and the second described his current activity. In the original dataset the activity was described by 1 to 4 tags. For the competition we have provided only the main activity. Test data file was in the same format as the training data set, however, the labels for the test series were hidden from participants.

It is important to note that the training and test data sets consisted of recordings which were obtained from different groups of firefighters. Each of the sets contained the data acquired from only four persons and the scenario of the training exercises which they conducted was slightly different. There were no firefighter identifiers available in the data. By hiding this information from participants we wanted to promote

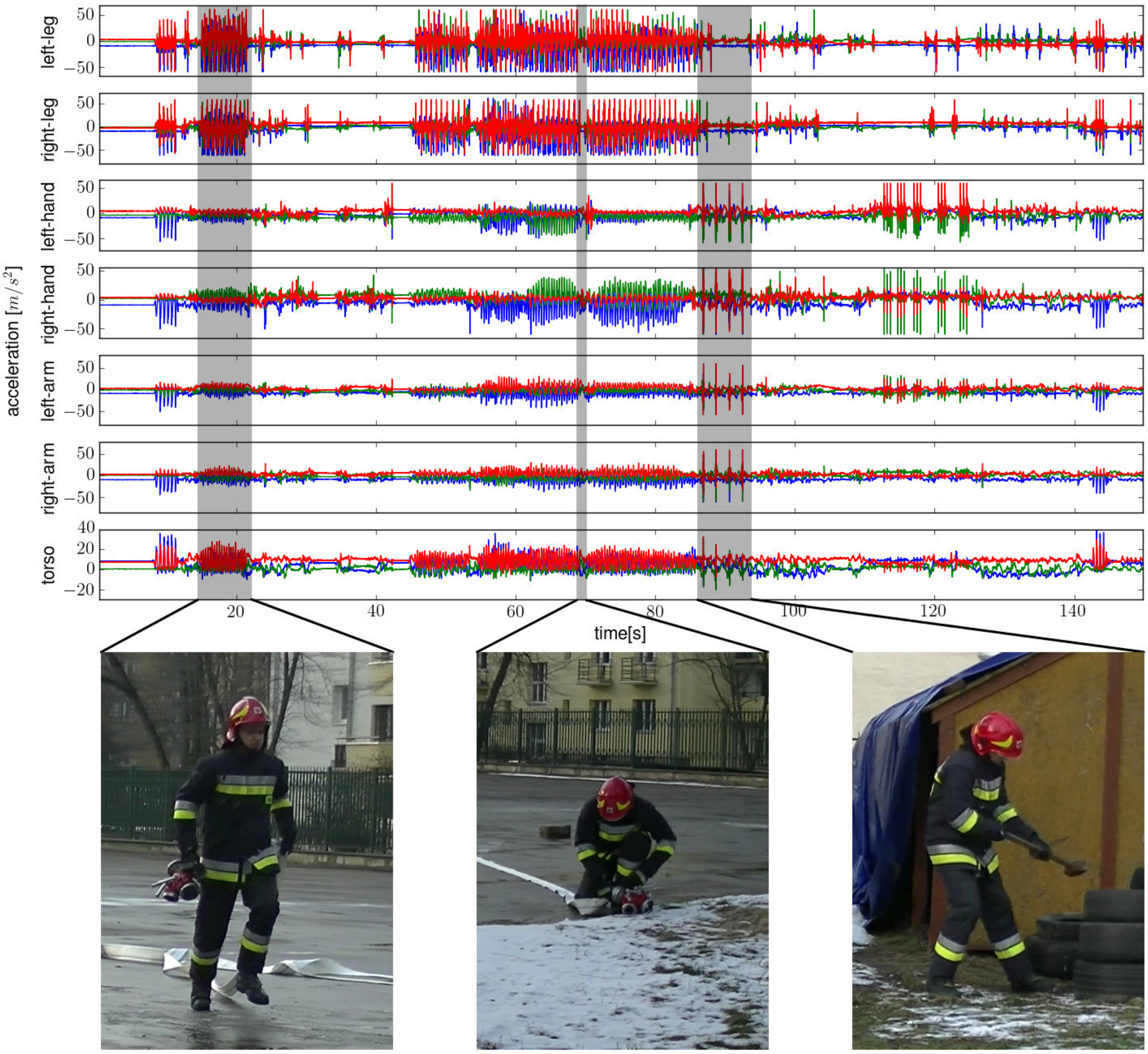


Fig. 2. Accelerations from BSN. Three actions were selected and marked.

solutions which are insensitive to individual characteristics of particular firefighters and are able to cope with disturbances of the data. Those prerequisites were the main reason for the unbalanced distribution of decision labels and had a huge impact on the shape of the most successful solutions.

### B. Evaluation procedure

The quality criterion in the competition was devised so that it reflected the specific requirements for the task. Since this task involved labelling sequences with tags that could take many different values with unbalanced representation in the data, the quality of submitted solutions was assessed using the balanced accuracy measure. This particular criterion had been already used in several others data mining competitions, e.g. [18]. It is insensitive to skewed distribution of decisions and thus promotes classifiers which are able to robustly identify labels of cases from minority classes.

The balanced accuracy (BAC) is typically defined as an average accuracy within all decision classes. In particular, if we denote a vector with predictions returned by a classifier by $preds$ and a vector of true decision labels by $labels$, we may define BAC as:

$$ACC_i(preds, labels) = \frac{|\{j : preds_j = labels_j = i\}|}{|\{j : labels_j = i\}|}, \quad (2)$$

$$BAC(preds, labels) = \frac{1}{l}\sum_{i=1}^{l} ACC_i(preds, labels), \quad (3)$$

where $l$ is the total number of possible labels. In our compe-

TABLE IV. THE FINAL AND PRELIMINARY RESULTS OF THE TOP-RANKED TEAMS.

| Rank | Team name | Preliminary | Final score |
|---|---|---|---|
| 1 | jan | 0.85768 | 0.83912 |
| 2 | zagorecki | 0.85184 | 0.82985 |
| 3 | nitekna | 0.85015 | 0.8261 |
| 4 | mathurin | 0.82523 | 0.80408 |
| 5 | lp319499 | 0.80318 | 0.79137 |
| ... | ... | ... | ... |
| 38 | baseline | 0.61414 | 0.60361 |

tition, a value of BAC measure was separately computed for the two decision attributes (the posture and activity). Due to the fact that the decision attributes considerably differed in the number of possible values, we decided to assign them different weights which would compensate for the increased difficulty of predicting the activity label. As a result, the BAC values for the posture and activity attributes had a different impact on the evaluation score. If we denote the BAC value for the posture as $BACp$ and the value for the activity as $BACa$, we can determine the evaluation score of a solution $s$ as:

$$score(s) = \frac{BACp(s) + 2 \cdot BACa(s)}{3}. \tag{4}$$

The evaluation procedure in the competition was two-fold. During the course of the contest an on-line evaluation system was providing a constant feedback for the participants in a form of a publicly available leaderboard — a dynamic ranking of participant's best results. However, the scores displayed on the leaderboard were only a preliminary assessment of the solution's quality. They were computed using only $10\,\%$ of available test data. After completion of the competition there were the second evaluation round. It was available only for those teams which had provided a description of their solution in a form of a short competition report. The final evaluation was carried out independently from the preliminary one, using the remaining part of the test data.

### C. Summary of the competition results

AAIA'15 Data Mining Competition attracted skilled data mining practitioners from around the world. Comparing to the previous year's edition of the challenge there were noticeably more registered teams (152 in total, an increase by 36 teams), from which 79 actively participated in the challenge by submitting at least one solution (an increase by 22 teams). We received 1,840 correctly formatted solutions (an increase by $42\,\%$) and the top-ranked participants have beaten the baseline solution by nearly 24 percentage points. Additionally, 50 teams provided a brief report describing their approach. Table IV shows scores obtained by the top-ranked teams.

The solutions submitted by participants proved to be a valuable source knowledge. They not only provided an insightful view on the state-of-the-art in multidimensional time series analysis but also contained inspiring new ideas, design specifically for the considered problem. The most interesting of these ideas are described by their authors in separate papers submitted for the competition track of AAIA'15 conference [2], [11], [15], [17], [19], [20].

### D. Summary of the most successful submissions

The top-ranked participants of AAIA'15 Data Mining Competition managed to test effectiveness of a wide range of approaches to classification of high dimensional sensory data. However, nearly all of the successful solutions had a common denominator. They all utilized some sort of attribute engineering or feature extraction methods [12] in order to represent the time series data by a new set of attributes. In all cases, the main purpose of this data transformation was to define an attribute space which on the one hand was insensitive to differences in movement patterns between different firefighters and on the other hand, was characterized by a much lower dimensionality in comparison to the tabular representation of the competition data.

Among the new features used in the top solutions a large share corresponded to summary statistics, commonly used to describe sample distributions. These characteristics include typical location statistics (i.e. mean, centiles, time window minimum and maximum), shape statistics (i.e. skewness, kurtosis) and dispersion statistics (i.e. standard deviation, energy, range and difference between centiles). Several solutions also made use of characteristics measuring the dynamics of data, such as the mean, minimum and maximum difference between values of consecutive sensor readings. Finally, the two best solutions in the competition were using features derived from the Fast Fourier Transform of the data.

Having defined a suitable representation of data, the participants were employing standard machine learning algorithms to perform the classification. Among the most popular classifiers were the Random Forest [4] and Support Vector Machines [1]. The best results were obtained by teams which carefully tuned parameters of their learning algorithm. Interestingly, the top-ranked participants were using different approaches to tackle the problem of two decision attributes in the data. Some of the teams constructed two independent classifiers, whereas the others merged the two class labels into a single one and trained a single model for a multi-class prediction problem. However, the best performing classification model was firstly trained to predict the first label (the posture) and then, the obtained prediction was used as a new feature for prediction of the second label [15].

## IV. CONCLUSION

Physical activity classification based on inertial data from body sensor networks is a very challenging task. The main difficulty that needs to be addressed when solving this problem is the fact, that same activity can be performed in very different ways by different subjects. Moreover, due to severe hardware limitations, any classification scenario in a real-life application needs to find a trade-off between a classification accuracy and power efficiency. So far very little work has been done to address this issue and devise new methods that would be suitable for mobile, long-running platforms.

In this paper we described a data set that can be used in future studies on this important subject. We also summarized a data mining competition which we had organized in order to draw attention of data mining community and stimulate research in this type of application areas. In particular, solutions submitted by participants of the competition constitute a valuable insight regarding the state-of-the-art in the on-line activity recognition field.

## References

[1] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92,* pages 144–152, New York, NY, USA, 1992. ACM.

[2] M. Boull'e. Tagging Fireworkers Activities from Body Sensors under Distribution Drift. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of FedCSIS 2015.* IEEE, 2015.

[3] B. Brehmer. *Strategies in Real-Time, Dynamic Decision Making. Insights in decision making,* pages 262–279, 1990.

[4] L. Breiman. *Random Forests. Machine Learning,* 45(1):5–32, 2001.

[5] Department of Communities and Local Goverment. *Fire Service Operations, Incident Command.* Fire Service Manual. London TSO, third edition, 2008.

[6] Emergency Management Institute. *Introduction to Incident Command System,* ICS-100. http://training.fema.gov/EMIWeb/IS/courseOverview.aspx?code=IS-100.b, 2013. Access: 22.02.201.

[7] R. Friedman, A. Kogan, and Y. Krivolapov. On power and throughput tradeoffs of wifi and bluetooth in smartphones. *Mobile Computing, IEEE Transactions on,* 12(7):1363–1376, July 2013.

[8] D. Gafurov, K. Helkala, and T. Søndrol. *Biometric gait authentication using accelerometer sensor,* 2006.

[9] A. Graeger, U. Cimolino, H. de Vries, and J. Sümersen. *Einsatzund Abschnittsleitung: Das Einsatz-Führungs-System (EFS).* Ecomed Sicherheit, 2009.

[10] L. J. Grorud and D. Smith. The National Fire Fighter Near-Miss Reporting. Annual Report 2008. In *An exclusive supplement to FireRescue magazine,* pages 1–24. Elsevier Public Safety, 2008.

[11] M. Grzegorowski and S. Stawicki. Feature Extraction from Machine Generated Data. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of FedCSIS 2015.* IEEE, 2015.

[12] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, editors. *Feature Extraction: Foundations and Applications.* Studies in Fuzziness and Soft Computing. Springer, 2006.

[13] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. A close examination of performance and power characteristics of 4g lte networks. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, MobiSys '12,* pages 225–238, New York, NY, USA, 2012. ACM.

[14] A. Janusz, A. Krasuski, S. Stawicki, M. Rosiak, D. Ślęezak, and H. S. Nguyen. Key Risk Factors for Polish State Fire Service: a Data Mining Competition at Knowledge Pit. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of FedCSIS 2014,* pages 345–354. IEEE, 2014.

[15] J. Lasek and M. Gagolewski. The Winning Solution to the AAIA?15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of FedCSIS 2015.* IEEE, 2015.

[16] M. Meina, B. Celmer, and K. Rykaczewski. Towards robust framework for on-line human activity reporting using accelerometer readings. In D. Ślęzak, G. Schaefer, S. Vuong, and Y.-S. Kim, editors, *Active Media Technology,* volume 8610 of Lecture Notes in Computer Science, pages 347–358. Springer International Publishing, 2014.

[17] S. Wawrzyniak and W. Niemiro. Clustering Approach to the Problem of Human Activity Recognition using Motion Data. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of FedCSIS 2015.* IEEE, 2015.

[18] M. Wojnarski, A. Janusz, H. S. Nguyen, J. Bazan, C. Luo, Z. Chen, F. Hu, G. Wang, L. Guan, H. Luo, J. Gao, Y. Shen, V. Nikulin, T.-H. Huang, G. J. McLachlan, M. Bošnjak, and D. Gamberger. RSCTC'2010 discovery challenge: Mining DNA microarray data for medical diagnosis and treatment. In M. S. Szczuka et al., editor, *Proceedings of RSCTC'2010,* volume 6086 of LNAI, pages 4–19, Heidelberg, 2010.Springer.

[19] A. Zagorecki. A Versatile Approach to Classification of Multivariate Time Series Data. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of FedCSIS 2015.* IEEE, 2015.

[20] E. Zdravevski, P. Lameskiy, R. Mingov, A. Kulakov, and D. Gjorgjevikj. Robust histogram-based feature engineering of time series data. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of FedCSIS 2015.* IEEE, 2015.

DOI: 10.15439/2015F418
ACSIS, Vol. 5

# The Winning Solution to the AAIA'15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene

Jan Lasek
Interdisciplinary PhD Studies Program,
Institute of Computer Science,
Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
Email: janek.lasek@gmail.com

Marek Gagolewski
Systems Research Institute,
Polish Academy of Sciences,
ul. Newelska 6, 01-447 Warsaw, Poland
and
Faculty of Mathematics and Information Science,
Warsaw University of Technology,
ul. Koszykowa 75, 00-662 Warsaw, Poland
Email: gagolews@ibspan.waw.pl

***Abstract*—Multi-sensor based classification of professionals' activities plays a key role in ensuring the success of an his/her goals. In this paper we present the winning solution to the *AAIA'15 Tagging Firefighter Activities at a Fire Scene* data mining competition. The approach is based on a Random Forest classifier trained on an input data set with almost 5000 features describing the underlying time series of sensory data.**

## I. Introduction

HUMAN activity recognition based on sensor inputs, cf., e.g., [1], [7], [14], is essential in many practical applications. In particular, a fire scene constitutes a dynamic environment in which valid, precise, and fast human decisions play a key role. Here, the aim is to achieve success in an emergency rescue mission, having in mind safety of the involved firemen [4] and his/her ability to save other peoples' lives and – in the second place – property, wealth, etc. It is worth noting that an automated decision support system may be used to increase the widely-conceived quality of an agents' behavior. One of its most fundamental components relies on a proper detection of an action a fireman is actually performing at a given moment. The topic of *AAIA 2015 Data Mining Competition: Tagging firefighters' activities at a fire scene* [8] aimed to deliver accurate model for recognising firefighters movements and activities based on multi-sensor data. In consecutive sections we explain the winning approach in very detail. The proposed solution was implemented in the `R` environment for statistical computing [11]. The solution is available on–line as a Git repository at https://github.com/janekl/AAIA15_Data_Mining_Contest.

The paper is organized as follows. In the section to follow, we describe the analyzed data set and define the evaluation metric used. In Section III we discuss main challenges that the data set brought. In Section IV we present the winning solution in detail and indicate its advantages, limitations, and possible extensions for future work. Finally, Section V concludes the paper.

## II. Problem Statement

The main purpose here is to design a model for a classification problem with two class attributes. The first class denotes the main activity of a firefighter. This class is referred to as *posture* class and it has 5 distinct labels (`crawling`, `crouching`, `moving`, `standing` and `stooping`). The second class, called *action* class denotes a particular activity of a firefighter and consists of 16 labels (4 labels associated with movement along ladder or stairs: `ladder_down`, `ladder_up`, `stairs_down`, `stairs_up`, 2 labels regarding forward movement: `walking` and `running`, labels describing firefighters' operational movements: `manipulating`, `nozzle_usage`, `signal_hose_pullback`, `signal_water_first`, `signal_water_main`, `signal_water_stop`, `striking` and `throwing_hose` and a `no_action` label).

The evaluation metric employed in the competition is the weighted average of *balanced accuracy* for the two classes. Below we recall the definition of this measure. For each label $l_i$ within a class attribute we define *classification accuracy* as

$$\mathrm{acc}(l_i) = \frac{|\{j : l(x_j) = l_i \wedge p(x_j) = l_i\}|}{|\{j : l(x_j) = l_i\}|},$$

where $l(x_j) = l_i$ denotes the true label for instance $x_j$ and $p(x_j)$ denotes the label assigned by a classifier. If a class attribute $C$ assumes $L$ possible labels, then the balanced accuracy score for that class is defined as

$$\mathrm{BAC}(C) = \frac{1}{L}\sum_{i=1}^{L}\mathrm{acc}(l_i).$$

Now, we may consider the weighted average of balanced accuracy scores for *posture* and *action* classes, which is given by

$$\mathrm{EvaluationMetric} = \frac{1}{3}\mathrm{BAC}(\mathtt{posture}) + \frac{2}{3}\mathrm{BAC}(\mathtt{action}).$$

978-83-60810-66-8/$25.00©2015, IEEE

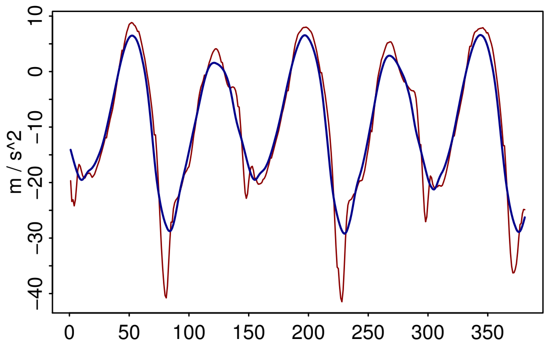


Fig. 1. Plot of the raw series (red) along $x$-axis of the accelerometer recordings at right hand for pair of labels (moving, running) and the smoothed series with 20-moving average filter (blue).

During the competition, the solutions were evaluated against approximately 10% of test data. An evaluation metric is the essence of a contest for both its organizers and participants. Through design of an evaluation metric, the organizers define their goal that they want to achieve. On the other hand, the participants need to tailor their models to optimize a given evaluation metric.

To train a statistical model, a training set consisting of 20.000 instances, each tagged with a pair of labels for *posture* and *action* class, was used. Each instance consists of basic statistics on the vital functions of a firefighter and a set of 42 time series which came from x/y/z–axis recordings from gyroscopes and accelerometers attached at 7 points on the body of a firefighter (left hand, right hand, left arm, right arm, left leg, right leg, torso). Each time series consists of 400 recordings (every 4-5 ms) over ca. 2 seconds. Test set consists of 20.000 instances as well. The goal was to develop a model for tagging instances in the test set with a pair of labels for the two class attributes. Both the training and test data set are of size approximately 2.4 GB (uncompressed *csv* files).

## III. Main challenges

### A. The same action, different results

Among one of the many challenges we find that the data set was inherently noisy. Moreover, the samples of activities in the training and test sets were due to activities of different firefighters. We observed that this had a significant impact on the classifier's score: our scores in terms of the evaluation metric were as high as 98% on a hold-out validation set. This is a considerably high score bearing in mind that the given classification problem is presumably not an easy task. However, in related studies as high accuracy scores were reported [10], [5]. The scores on the official leader-board were significantly lower – with the best scores being equal to ca. 85% during preliminary evaluation. The fact that an instance may come from different source is a great challenge in any application domain.

### B. Imbalance of labels distribution

Another problem which required proper handling was related to the imbalance of labels proportion within each class attribute. Table I presents the pair of labels within *action* and *posture* for the test set.

Since the evaluation metric in the competition was the discussed balanced accuracy score, no label was distinguished and misclassification rate has equal weight for every label within each class. This metric treats each label within a class as being equally important (equal weights), regardless of its *a priori* distribution in the data. This distribution of labels varied significantly on the training set. For example, only about 0.5% of all instances constitute for the `signal_hose_pullback` label while about 32% for the `manipulating` label within the *action* class. This means that we are given over 60 times more instances having the former label. Such an uneven distribution of labels requires proper handling by a model. To overcome the problem of imbalanced label distribution, we trained individual classifiers in an ensemble (to be precise, using the below-discussed Random Forest method) based on a stratified subsamples of training set in which each label was represented in an equal amount. The proper balancing of the training set enabled to tailor a model for the evaluation metric employed in the competition.

## IV. The winning solution

Let us describe the implemented approach towards feature extraction and model building for activity tagging problem. The model used was based on the Random Forest classifier which is an ensemble of decision trees. It is observed that in practical situations it often yields high accuracy scores [2], [6]. Another advantage of the Random Forest classifier is that it is a fast method: its training and prediction phase can be parallelized. Is is also relatively easy to handle (i.e., its parameter setup) as compared to other advanced ensemble methods. Both the described below feature extraction and the final model training procedures (included in the GitHub repository) can be performed on a single machine within a couple of hours. In our computations we used a 4–core 2.0 GHz CPU 16 GB RAM machine. The described parameter optimization steps were performed on a cluster of 10 8–core 3.40 GHz CPU 16 GB RAM machines to speed up the computations.

### A. Feature extraction

Our approach was particularly focused on the phase dealing with features' extraction. The extracted features were based on literature [5], [9], [10] as well as the authors' experimental ideas. The processed training/test dataset is of size about 1.4 GB. For each activity we derived over 4700 features describing a particular activity. First of all, we filtered the data with a moving average window of size 20, see Figure 1. Since the sensor recordings were gathered at a 4.5 ms. resolution, this roughly corresponds to averaging the arriving over a window of 0.1 second. This step was not crucial for the model performance, however, it allowed to filter out the noise slightly.

TABLE I
COUNTS FOR PAIRS OF LABELS FOR THE TWO CLASSES – TRAINING SET.

| | crawling | crouching | moving | standing | stooping |
|---|---|---|---|---|---|
| ladder_down | 0 | 0 | 465 | 0 | 0 |
| ladder_up | 0 | 0 | 476 | 0 | 0 |
| manipulating | 0 | 1764 | 331 | 2356 | 1898 |
| no_action | 0 | 87 | 0 | 491 | 0 |
| nozzle_usage | 0 | 492 | 0 | 443 | 0 |
| running | 0 | 0 | 4324 | 0 | 0 |
| searching | 459 | 0 | 0 | 0 | 0 |
| signal_hose_pullback | 0 | 0 | 0 | 98 | 0 |
| signal_water_first | 0 | 0 | 41 | 496 | 0 |
| signal_water_main | 0 | 46 | 0 | 405 | 0 |
| signal_water_stop | 0 | 0 | 0 | 277 | 0 |
| stairs_down | 0 | 0 | 644 | 0 | 0 |
| stairs_up | 0 | 0 | 1157 | 0 | 0 |
| striking | 0 | 0 | 0 | 1022 | 0 |
| throwing_hose | 0 | 0 | 0 | 234 | 930 |
| walking | 0 | 0 | 1064 | 0 | 0 |

Next, for each of the time series, we derived basic summary statistics: quantiles (denoted with qx in Tables III and IV for x $\in \{0.01, 0.05, 0.1, 0.2, \ldots, 0.9, 0.95, 0.99\}$), standard deviation (sd), skewness, kurtosis, amplitude (defined as the difference between 0.99-quantile and 0.01-quantile of the series), the signal energy (ener; defined as the sum of squares of consecutive recordings), the ratio between its maximal absolute value and the median and minimal and maximal of the first differences of a series (deriv1min and deriv1max). We extracted a set of quantiles and standard deviations on the time series processed by the Fast Fourier Transform, to its real, imaginary and modulus (ModFFT) parts independently. Additionally, we recorded first 5 Fourier coefficients of the real and imaginary part of the transformed series. We also extracted quantiles and standard deviation of the periodogram (Period) of each time series. Further, for each pair of time series we computed the linear correlation coefficients (cor).

We also extracted several experimental features for counting the number of peaks in the series based on their sub-chunks in which they exceeded the mean by one or two standard deviations. We imposed a constraint that the minimal length of a sub-chunk is 5 (for the filtered series). Finally, we counted the number of times a given series crosses 0 and its mean.

Another property of Random Forest model is that it has an inherent method of evaluation of feature relevance. Tables III and IV present the 50 most important features for two individual classification tasks for the two class attributes. The criterion of our choice according to which features are evaluated is the mean decrease in Gini Impurity Index for classification (column M.D.Gini). As far as the vital functions are concerned, median respiratory rate reading med.rr is present in the top 50 list for *action* classification problem.

Let us note that the number of features derived is large and some of them do not posses a clear interpretation. However, due to Random Forest model described in the next section – which includes an inherent method of selecting relevant attributes – we were able to handle and select relevant content from this rich set of features.

### B. Classification model

For the purpose of tagging the activities we used the balanced Random Forest [2]. By the balanced Random Forest classifier we mean an ensemble of trees that are trained on subsamples of training set in which every label within a given class is represented in an equal amount. This model was used in a stepwise approach. In the first step, we trained the model which aimed to recognize the *posture* of a firefighter. In the second one, we trained the model to recognise the main *action* of a fireman, given *posture* class attribute. This approach is analogous to the classifier chaining method in multilabel classification tasks [12]. In the tagging phase for new data we plug the predicted *posture* labels by the first model as an input for the model for the *action* class. The combined predictions complete the tagging phase for test set.

The idea behind such a chaining method was driven by the fact that some combinations of activities and posture labels are mutually exclusive. For example, the posture cannot be equal to standing when the main activity is equal to ladder_up. When individual classifiers were trained, such inconsistencies were very common. We managed to reduce them by employing the mentioned stepwise approach. However, we did not succeed in eliminating them at all: our final submission still contained some fraction of prediction labels that were mutually exclusive. By mutually exclusive pairs of labels we mean such a combination of pairs of labels that were not observed in the training set (see Table II; these pairs are given in bold). Another way of reducing conflicts was to aggregate different submissions by, e.g., majority voting. We observed that aggregating individual submissions often produced a new submission with a higher preliminary evaluation score than each of the individual ones. This serves as a method for providing more stable and accurate predictions since, e.g., they are based on a larger number of trees.

We also experimented with one-vs-all and single class (i.e., a single class was obtained by mapping each pair of labels within (*posture*, *action*) classes to an individual class) versions of the model. However, the best results were achieved by

TABLE II
COUNTS FOR PAIRS OF PREDICTED LABELS FOR THE TWO CLASSES – TEST SET.

| | `crawling` | `crouching` | `moving` | `standing` | `stooping` |
|---:|---:|---:|---:|---:|---:|
| `ladder_down` | 0 | **1** | 459 | **209** | 0 |
| `ladder_up` | 0 | **2** | 452 | **118** | 0 |
| `manipulating` | 0 | 1576 | 12 | 1639 | 2438 |
| `no_action` | 0 | 71 | 0 | 467 | **31** |
| `nozzle_usage` | 0 | 454 | 0 | 1060 | 0 |
| `running` | 0 | **13** | 3974 | 0 | **2** |
| `searching` | 513 | **42** | 0 | 0 | 0 |
| `signal_hose_pullback` | 0 | 0 | 0 | 96 | 0 |
| `signal_water_first` | 0 | **3** | 10 | 580 | 0 |
| `signal_water_main` | 0 | 55 | 0 | 174 | 0 |
| `stairs_down` | 0 | 0 | 533 | 0 | 0 |
| `stairs_up` | 0 | 0 | 1442 | 0 | 0 |
| `striking` | 0 | **13** | **7** | 1026 | **49** |
| `throwing_hose` | 0 | 0 | 0 | 196 | 982 |
| `walking` | 0 | **2** | 1251 | **46** | **2** |

the described chaining method.

### C. Parameter tuning

Due to the discussed issue of performing activities by different people, the model's parameter tuning process poses a real challenge. We were primarily interested in the parameters responsible for balancing the classifier (parameter `sampsize` in `R`'s `randomForest` package), the minimal number of instances in the leafs (parameter `nodesize`) and the number of sampled features to perform a test split (parameter `mtry`). In our methodology we experimentally set parameters by monitoring out–of–bag error accuracy estimates for each of the classes. Additionally, we run 3–fold cross validation for different pairs of parameters (`mtry`, `nodesize`). Our conclusions was that the parameter `nodesize` should be set to 1, i.e., the trees should be grown to maximal depth. Moreover, given `nodesize = 1`, setting parameter `mtry` to a couple of hundreds already provided stable and high accuracy scores. Finally, to balance training sets, in our initial trials we sampled more instances of most represented labels, i.e., `moving` within *posture* class and `manipulating` and `running` within *action* class as indicated by lower out–of–bag error estimates for those labels. However, significantly better preliminary scores were obtained just by sampling each of the labels in an equal amount. Although choosing parameter values based on leader–board score is quite a dangerous way of tuning them, we took this risk for those parameters as the test set instances differ significantly from the training set ones. In any case, sampling each of the labels equally appears to be a reasonable setup. The described sampling procedure was a crucial step of achieving high evaluation scores. Another advantage of this methodology is that the models are training using fewer instances from the training set. This may in turn prove useful to a reduce overfitting of the model to the training data. The number of trees in the forest was set to 700, i.e., a relatively large number accounting for the computation time of the model.

### D. Final submission

During the competition, we submitted over 100 proposals, which were based on different ideas and changes in model parameters and enrichment of the training data with new features. Most of them relied on experiments with different setup of Random Forest model, but we also tried the Gradient Boosting Machine (GBM) model (tree-based) [3], [13]. However, the performance of a less complex Random Forest model was satisfactory and we devoted more time for optimising this model. Moreover, we primarily focused on the feature extraction step.

The best performing model consists of 700 trees, it has the number of attributes for performing test split equal to 300, stratified over class attribute with sample size of 400 for each *posture* and 90 for *action* label. The final submission was derived by majority voting of three classifiers (in fact, two–stage classifiers) with weights 1.5 (to avoid ties), 1, 1 respectively:

1) Random forest model with the minimum size a leaf in a single tree equal to 1 (attribute `nodesize = 1`)
2) Random forest model with `nodesize = 3` and
3) Random forest model with `nodesize = 1` trained on the dataset with exclusion of features associated with left arm of sensory data and a subset of quantiles $(0.01, 0.05, 0.2, 0.4, 0.6, 0.8, 0.95, 0.99)$.

The preliminary evaluation scores were 0.858, 0.8573, and 0.8567 for consecutive models. The averaging step was aimed to reduce the variance of a single model as well as to resolve the mentioned conflicts due to contradictory labels. We used this method as we concluded by our previous experimentation with averaging that it yields higher evaluation scores. However, in case of our final submission, it yielded a not significantly lower preliminary evaluation score than the best one used for aggregating. In any case, we believed that it would produce more stable and accurate predictions on the whole test set. The third model was trained on a subset of attributes since we observed that some recordings in the test set have constant values of these which we interpret as missing values. We also excluded a part of quantiles with the aim to reduce overfitting of the model to the training sample (however, this appeared not to be of help in this case). The final submission yielded score of $0.8577$ during preliminary evaluation and $0.8391$ on the whole test set – it was ranked the first among 79 submitted proposals.

### *E. Unsolved puzzles a.k.a. future work*

By the end of the challenge, we were still left with some unsolved problems that became evident after the solutions were submitted.

First of all, our final submission still contained some mutually exclusive pairs of labels e.g. `ladder_down` and `standing` or `walking` and `standing`. This problem was limited to some extent by the two–stage classification as well as submissions averaging.

The other problem with our submissions was that our model never predicted the activity `signal_water_stop`. Perhaps feeding the classifier with more instances with this particular label could resolve this issue. This could also possibly apply to `signal_hose_pullback` label within *action* class as there where merely 98 instances tagged with this activity. Finally, as we already mentioned, our preliminary evaluation scores based on out–of–bag predictions from the Random Forest model were overly optimistic: the scores on training data were about $98\%$ of the evaluation metric while the evaluation of our solution on the whole training data yielded much lower score of about $84\%$. This issue could be addressed by, e.g., performing evaluation and optimisation of a model via cross-validation, where the validation folds would contain activities performed by different firefighters. However, this could not be performed as the information on, e.g., firefighters identifiers performing a given action was not made available to the participants. Another possibility would be to derive more robust features with better generalisation properties for different people performing the same activities.

## V. Conclusions

The *AAIA'15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene* contest was a very interesting and absorbing event. Taking part in such competitions requires some persistence as only few tested ideas prove to give an improvement for the classification score. It is enough to mention that our winning solution was submitted within 24 hours of the competition's deadline.

In our approach, we employed Random Forest classifier and spend much more time on pre-processing data and engineering new features. We believe that the key to success were good data. Using a more sophisticated model may constitute for improvement in classification, however, as in this competition raw time series data needed to be processed, we regarded feature engineering as a more important step. Moreover, proper balancing of training sample provided major gains in the evaluation metric employed in this competition.

The code for our submission is available at GitHub. Further enhancements of the proposed solution are possible. We hope that it will serve as a benchmark for even better performing models for the task of tagging activities at a fire scene.

## Acknowledgments

Jan Lasek would like to acknowledge the support by the European Union from resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications", agreement UDA-POKL.04.01.01-00-051/10-00 via the Interdisciplinary PhD Studies Program.

## References

[1] Atallah, L., Lo, B., Ali, R., King, R., Yang, G.-Z. 2009. Real-time activity classification using ambient and wearable sensors, IEEE Transactions on Information Technology in Biomedicine 13:1031–1039, http://dx.doi.org/10.1109/TITB.2009.2028575.

[2] Breiman, L. 2001. Random forests, Machine Learning 45:5–32, http://dx.doi.org/10.1023/A:1010933404324.

[3] Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine, Annals of Statistics 29:1189–1232, http://dx.doi.org/10.1214/aos/1013203451.

[4] Krasuski A. 2014. A framework for Dynamic Analytical Risk Management at the emergency scene. From tribal to top down in the risk management maturity model. Proc. Federated Conference on Computer Science and Information Systems (FedCSIS'14), IEEE, pp. 323–330, http://dx.dox.org/10.15439/2014F371.

[5] Leutheuser H., Schuldhaus D., Eskofier B.M. 2013. Hierarchical, multi-sensor based classification of daily life activities: Comparison with state-of-the-art algorithms using a benchmark dataset, PLoS ONE 8:e75196, http://dx.doi.org/10.1371/journal.pone.0075196.

[6] Liaw, A., Wiener M. 2002. Classification and Regression by randomForest. R News 2:18–22, urlhttp://cran.r-project.org/doc/Rnews/.

[7] Maurer, U., Smailagic, A., Siewiorek, D.P., Deisher, M. 2006. Activity recognition and monitoring using multiple sensors on different body positions, Proc. International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06), IEEE, pp. 113–116, http://dx.doi.org/10.1109/BSN.2006.6.

[8] Meina, M., Janusz, A., Rykaczewski, K., Ślęzak D. Celmer, B., Krasuski, A. 2015. Tagging Firefighter Activities at the Emergency Scene: Summary of AAIA'15 Data Mining Competition at Knowledge Pit, Proc. Federated Conference on Computer Science and Information Systems (FedCSIS'15), IEEE, pp. 379–385, http://dx.doi.org/10.15439/2015F426.

[9] Mörchen, F. 2003. Time series feature extraction for data mining using Discrete Wavelet Transform and Discrete Fourier Transform, Technical Report No. 33, Philipps-University Marburg, Germany.

[10] Preece, S.J., Goulermas, J.Y., Kenney, L.P.J., Howard, D. 2009. A Comparison of Feature Extraction Methods for the Classification of Dynamic Activities From Accelerometer Data, IEEE Transactions on Biomedical Engineering 56:871–879, http://dx.doi.org/10.1109/TBME.2008.2006190.

[11] R Core Team. 2015. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

[12] Read, J., Pfahringer, B., Holmes, G.J., Frank, E. 2009. Classifier Chains for Multi-label Classification, Machine Learning and Knowledge Discovery in Databases 5782:254–269, http://dx.doi.org/10.1007/978-3-642-04174-7_17.

[13] Ridgeway, G. et al. 2015. `gbm`: Generalized Boosted Regression Models, R package version 2.1.1, http://CRAN.R-project.org/package=gbm.

[14] Safonov, I., Gartseev, I., Pikhletsky, M., Tishutin, O., Bailey M.J.A. 2015, An approach for model assessment for activity recognition, Pattern Recognition and Image Analysis 25:263–269, http://dx.doi.org/10.1134/S1054661815020224.

TABLE IV

Evaluation of feature importance according to mean decrease in Gini impurity index for the **action** class.

| Rank | Feature name | M.D.Gini |
|---|---|---|
| **1** | cor.acc_left_leg_x.gyr_left_leg_y | 12.03 |
| **2** | cor.acc_right_leg_x.gyr_right_leg_y | 10.56 |
| **3** | cor.gyr_left_leg_y.gyr_right_leg_y | 10.50 |
| **4** | q50.acc_left_hand_x | 8.71 |
| **5** | q60.acc_left_hand_x | 8.40 |
| **6** | q30.acc_left_hand_x | 8.31 |
| **7** | q40.acc_left_hand_x | 7.95 |
| **8** | q70.acc_left_hand_y | 7.87 |
| **9** | q20.acc_right_hand_y | 7.85 |
| **10** | q30.acc_right_leg_x | 7.54 |
| **11** | deriv1max.acc_left_arm_z | 7.17 |
| **12** | ModFFT.sd.gyr_left_leg_y | 7.15 |
| **13** | q20.acc_right_arm_z | 7.15 |
| **14** | Period.sd.acc_left_leg_z | 7.12 |
| **15** | ener.gyr_left_leg_y | 7.12 |
| **16** | Period.sd.gyr_left_leg_y | 6.97 |
| **17** | q20.acc_torso_x | 6.92 |
| **18** | q20.acc_left_hand_x | 6.90 |
| **19** | q70.gyr_left_leg_y | 6.74 |
| **20** | q10.gyr_left_leg_y | 6.70 |
| **21** | q70.acc_left_hand_x | 6.57 |
| **22** | q20.acc_right_leg_x | 6.56 |
| **23** | q40.acc_left_leg_x | 6.27 |
| **24** | sd.gyr_left_leg_y | 6.21 |
| **25** | q10.acc_right_arm_z | 6.20 |
| **26** | q30.acc_left_leg_x | 6.06 |
| **27** | q50.acc_left_leg_x | 5.91 |
| **28** | q10.acc_left_hand_x | 5.83 |
| **29** | q40.acc_right_arm_z | 5.76 |
| **30** | ModFFT.sd.acc_left_leg_z | 5.64 |
| **31** | q30.acc_right_arm_z | 5.59 |
| **32** | q95.acc_left_hand_y | 5.48 |
| **33** | q80.gyr_right_hand_y | 5.40 |
| **34** | q80.acc_left_hand_x | 5.37 |
| **35** | ener.acc_right_arm_z | 5.37 |
| **36** | q20.acc_left_leg_x | 5.31 |
| **37** | q90.acc_right_leg_z | 5.04 |
| **38** | q80.gyr_left_leg_y | 4.98 |
| **39** | ener.gyr_right_hand_y | 4.85 |
| **40** | Period.sd.gyr_right_arm_x | 4.81 |
| **41** | q70.acc_torso_x | 4.78 |
| **42** | ener.acc_left_leg_x | 4.68 |
| **43** | deriv1min.acc_left_arm_z | 4.68 |
| **44** | ener.acc_left_hand_x | 4.48 |
| **45** | q60.acc_left_leg_x | 4.43 |
| **46** | q40.acc_left_arm_x | 4.43 |
| **47** | q95.acc_torso_x | 4.41 |
| **48** | q05.gyr_right_hand_y | 4.38 |
| **49** | sd.acc_left_leg_z | 4.33 |
| **50** | Period.sd.acc_right_leg_z | 4.28 |

TABLE III

Evaluation of feature importance according to mean decrease in Gini impurity index for the **posture** class.

| Rank | Feature name | M.D.Gini |
|---|---|---|
| **1** | q40.acc_right_leg_x | 38.16 |
| **2** | q20.acc_right_leg_x | 36.25 |
| **3** | q30.acc_right_leg_x | 34.59 |
| **4** | q01.acc_torso_x | 34.23 |
| **5** | q10.acc_torso_x | 32.50 |
| **6** | q70.acc_left_leg_z | 30.99 |
| **7** | q05.acc_torso_x | 29.25 |
| **8** | q20.acc_torso_x | 29.12 |
| **9** | q30.acc_torso_x | 27.50 |
| **10** | q50.acc_right_leg_x | 27.04 |
| **11** | q80.acc_left_leg_z | 25.88 |
| **12** | ener.acc_right_leg_x | 25.79 |
| **13** | q90.acc_left_leg_z | 24.90 |
| **14** | q60.acc_left_leg_z | 23.37 |
| **15** | q10.acc_right_leg_x | 22.35 |
| **16** | q40.acc_torso_x | 21.73 |
| **17** | q50.acc_torso_x | 20.85 |
| **18** | q40.acc_left_leg_x | 19.65 |
| **19** | q30.acc_left_leg_x | 18.86 |
| **20** | q60.acc_right_leg_x | 18.42 |
| **21** | q95.acc_left_leg_z | 17.89 |
| **22** | q60.acc_torso_x | 16.08 |
| **23** | q20.acc_left_leg_x | 14.87 |
| **24** | q70.acc_right_leg_x | 14.65 |
| **25** | q50.acc_left_leg_x | 14.09 |
| **26** | q60.gyr_left_leg_y | 13.17 |
| **27** | med.rr | 12.71 |
| **28** | q70.gyr_left_leg_y | 12.48 |
| **29** | q80.gyr_right_leg_y | 12.24 |
| **30** | q50.acc_right_hand_x | 12.13 |
| **31** | cor.acc_torso_x.acc_torso_z | 11.33 |
| **32** | q70.acc_torso_x | 10.53 |
| **33** | q99.acc_left_leg_z | 10.24 |
| **34** | q40.acc_right_hand_x | 9.84 |
| **35** | q60.acc_right_hand_x | 9.57 |
| **36** | q80.gyr_left_leg_y | 9.46 |
| **37** | q50.acc_left_leg_z | 9.14 |
| **38** | q40.acc_left_leg_z | 8.95 |
| **39** | q90.acc_torso_x | 8.71 |
| **40** | q95.acc_torso_x | 8.62 |
| **41** | ener.acc_left_leg_x | 8.44 |
| **42** | q60.acc_left_leg_x | 8.31 |
| **43** | Period.sd.acc_right_leg_x | 8.09 |
| **44** | q30.acc_left_leg_z | 7.74 |
| **45** | ener.acc_right_hand_x | 7.74 |
| **46** | q80.acc_torso_x | 7.63 |
| **47** | sd.acc_right_leg_x | 7.55 |
| **48** | ModFFT.sd.acc_right_leg_x | 7.22 |
| **49** | q80.acc_right_leg_x | 6.73 |
| **50** | q90.gyr_right_leg_y | 6.65 |

DOI: 10.15439/2015F420
ACSIS, Vol. 5

# Robust histogram-based feature engineering of time series data

Eftim Zdravevski*,Petre Lameski†, Riste Mingov¶, Andrea Kulakov‡ and Dejan Gjorgjevikj§
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: *eftim.zdravevski@finki.ukim.mk, †petre.lameski@finki.ukim.mk,
‡andrea.kulakov@finki.ukim.mk, §dejan.gjorgjevikj@finki.ukim.mk
NI TEKNA - Intelligent Technologies, Negotino, Macedonia
Email: ¶riste.mingov@ni-tekna.com

***Abstract*—Collecting data at regular time nowadays is ubiquitous. The most widely used type of data that is being collected and analyzed is financial data and sensor readings. Various businesses have realized that financial time series analysis is a powerful analytical tool that can lead to competitive advantages. Likewise, sensor networks generate time series and if they are properly analyzed can give a better understanding of the processes that are being monitored. In this paper we propose a novel generic histogram-based method for feature engineering of time series data. The preprocessing phase consists of several steps: deseansonalyzing the time series data, modeling the speed of change with first derivatives, and finally calculating histograms. By doing all of those steps the goal is three-fold: achieve invariance to different factors, good modeling of the data and preform significant feature reduction. This method was applied to the AAIA Data Mining Competition 2015, which was concerned with recognition of activities carried out by firefighters by analyzing body sensor network readings. By doing that we were able to score the third place with predictive accuracy of about 83%, which was about 1% worse than the winning solution.**

***Keywords*—feature engineering, feature reduction, time series classification, temporal data mining***

## I. Introduction

THE introduction of lightweight and low-cost sensors has increased the potential for real time measurements of different activities. The advancements in microelectronics, wireless communications and other scientific areas has introduced the possibility of placing tiny sensor nodes on specific places of the body in order to monitor the health of patients or human body activities in general [1]. These sensors generate large amounts of data that need to be processed often in real-time. Most of the data, like temperature, accelerometer readings, GPS locations, etc can be presented as a time series data and processed as such. Time series data analysis allows detection of patterns in the data and making assumptions about the current activities or even predict future activities based on the past data. The operations that can be performed based on the time series data are mainly directed towards pattern discovery, clustering, classification and rule discovery [2]. Due to the density of the available data that can be collected by the sensors and the nature of the time series data, there are three main tasks that need to be defined so that the previously mentioned operations can be performed [3]: Dimensionality reduction and Data representation, Distance measurement and Indexing.

Dimensionality reduction and Data representation is one of the most important tasks that need to be performed when analyzing the time series data. This process is taken for granted when we use our sensory organs to obtain the data and then our brain processes it so we can make conclusions. The time series data is usually noisy and too large to be processed by a computer in an acceptable time frame. The human brain processes have learned to ignore the noisy data when generating conclusions. This is why a good representation and dimensionality reduction is crucial before we can continue with the decision making based on the given or obtained data when using a machine learning method. The representation must consider the assumption that time series data is not always aligned properly [4], that it is noisy and that it should comply to the constraints of time and space for its processing. By choosing the right data representation we are able to engineer good features for any given dataset.

The distance measurement is one of the main things that need to be defined in order to make a successful distinction between different time series and be able to correctly classify or identify similar patterns in the data. There are several well known distance measurements that can be used to identify difference between time series. The distance measurement must be invariant to many transformations of the time series data such as amplitude or time shifting, uniform amplification, additive noise, time scaling, etc [3]. For this reason many types of distance measurements have been proposed in the literature and each have their advantages and drawbacks. They can be divided in several groups. There are distance measurements based on the time series shape that use the direct signal properties to give the distance between two series. Then, there are measurements based on the operations needed to make the signals similar with each-other. Also there are measurements based on features extracted from the signal and finally there are measurements based on finding some higher level structure so the series can be compared.

The Indexing problem is related to improving the retrieval speed for a given series when searching trough a database of time series data.

In this paper we propose a histogram-based method for feature engineering for time series data. We use Support Vector Machine to generate the classification model for the data and present the obtained results.

978-83-60810-66-8/$25.00©2015, IEEE

The paper is organized as follows. In section II we describe the problem that is addressed by this paper. Then, in section III we give overview of the process used to generate the features and the needed transformations of the data. Next, in section IV we address the machine learning method for generating classification models based on the obtained features. Thereupon, in section V we present the obtained results after applying the proposed methods on the competition dataset. Finally, in section VI we discuss the findings of this research and make some conclusions about the applicability of the proposed methods for feature engineering of time series in general.

## II. Problem description

The topic of the AAIA'15 Data Mining Competition [5] was Tagging Firefighter Activities at a Fire Scene. In particular, the task is related to the problem of recognizing activities carried out by firefighters based on streams of information from body sensor networks. A fire ground is considered to be one of the most challenging decision taking environment. In dynamically changing situations, such as those occurring at a fire scene, all decisions need to be taken in a very short time [6]. Several initiatives and research projects analyze various aspect of this complex problem [6, 7, 8]. The lack of situational awareness is listed there as the main factor associated with major accidents among firefighters. The research presented in these papers aims to increase the firefighter safety by monitoring their kinematics and psychophysical condition during the course of fire and rescue actions. The following paragraph is extracted from the competition website [5] and describes the task in more detail.

During the course of the ICRA project [9], a so called "smart jacket" have been developed. This device is a wearable set of body sensors that allows to automatically track a firefighter at a fire scene. It also enables real-time screening of firefighter's vital functions and monitoring of ongoing activities at the scene. The later of those two tasks is the main scope of this AAIA'15 Data Mining Competition. The goal was participants to come up with efficient algorithms for labeling activities conducted by firefighters during their training exercises, based on provided data sets from a body sensor network. The data were obtained during training exercises conducted by a group of eight firefighters from The Main School of Fire Service. The sensors were registering firefighter's vital functions (i.e. ECG, heart rate, respiration rate, skin temperature) and movement (i.e. seven sets of accelerometers and gyroscopes placed on torso, hands, arms and legs). Each exercise session was also captured on video and the recordings were synchronized with time series representing the sensor readings. All this data were presented to experts who manually labeled it with activities. The objective in this competition is to devise efficient methods for automatic labeling of short series of the sensory data with basic activities of a firefighter. On the one hand, this task is very challenging due to a fact that different people tend to perform the same activities in different ways. On the other hand, however, automatically generated and accurate activity labels would facilitate monitoring of firefighter's safety and contribute to development of efficient command support systems.

The submitted solutions were evaluated on-line and the preliminary results were published on the competition leaderboard. The preliminary score was computed on a random subset of the test set, fixed for all participants. It corresponded to approximately 10% (about 2000 instances) of the test data. The final evaluation was performed after completion of the competition using the remaining part of the test data (about 18000 instances). Those results was also published on-line. The assessment of solutions was done using the balanced accuracy measure which is defined as an average accuracy within all decision classes. It was computed separately for the labels describing the posture and main activities of firefighters. The final score in the competition was a weighted average of balanced accuracies computed for those two sets of labels. Namely, for a vector of predictions *preds* and a vector of true labels *labels*, the balanced accuracy is defined with eq. (1) and (2). Here $BAC_p$ is the balanced accuracy for labels describing the posture, and $BAC_a$ is the balanced accuracy for labels describing the main activity.

$$ACC_i(preds, labels) = \frac{|j : preds_j = labels_j = i|}{|j : labels_j = i|} \tag{1}$$

$$BAC(preds, labels) = \left( \sum_{i=1}^{l} ACC_i(preds, labels) \right) / l \tag{2}$$

The final score in the competition for a solution $s$ was be computed with eq. (3):

$$score(s) = \frac{BAC_p(s) + 2 \times BAC_a(s)}{3} \tag{3}$$

The instances of the available training and test datasets are comprised mostly of sensor readings as time series and 42 values representing some aggregations of data from sensors monitoring firefighter's vital functions. There are totally 7 sensor locations and 2 sensor types, and each sensor is providing readings for the 3 axes, so the total number of time series is $7 \times 2 \times 3 = 42$. Each of those 42 time series has 400 samples, and one additional series of the timestamps of the readings relative to the start of the series. Each instance is labeled with two labels: one representing the posture of the body, and one representing the main activity of the firefighter. After analyzing the datasets, it is evident that there are 4 classes of the first label and 16 classes of the second label, while there are totally 24 different class combinations of the first and second label. Table I displays the classes for each of the two labels and their distribution in the training set.

Obviously the first challenge in this task is the feature engineering of the time series, so that they can be powerful predictors in relation to the labels, but also to be invariant to several things:

- Invariant to the range of values of the sensor readings because different firefighters can perform the same actions differently.
- Invariant to the alignment of the interval represented by the time series. In the current case, let us consider an action that is being performed longer the duration of the intervals. For example, the firefighter might be running for 20 seconds. From those 20 seconds we

TABLE I. DISTRIBUTION OF CLASSES PER LABEL

| Label 1 | Label 2 | Training Instances | Distribution (%) |
|---:|---:|---:|---:|
| crawling | searching | 459 | 2.3 |
| crouching | manipulating | 1764 | 8.8 |
| crouching | no_action | 87 | 0.4 |
| crouching | nozzle_usage | 492 | 2.5 |
| crouching | signal_water_main | 46 | 0.2 |
| moving | ladder_down | 465 | 2.3 |
| moving | ladder_up | 476 | 2.4 |
| moving | manipulating | 331 | 1.7 |
| moving | running | 4324 | 21.6 |
| moving | signal_water_first | 41 | 0.2 |
| moving | stairs_down | 644 | 3.2 |
| moving | stairs_up | 1157 | 5.8 |
| moving | walking | 1064 | 5.3 |
| standing | manipulating | 2356 | 11.8 |
| standing | no_action | 491 | 2.5 |
| standing | nozzle_usage | 443 | 2.2 |
| standing | signal_hose_pullback | 98 | 0.5 |
| standing | signal_water_first | 496 | 2.5 |
| standing | signal_water_main | 405 | 2.0 |
| standing | signal_water_stop | 277 | 1.4 |
| standing | striking | 1022 | 5.1 |
| standing | throwing_hose | 234 | 1.2 |
| stooping | manipulating | 1898 | 9.5 |
| stooping | throwing_hose | 930 | 4.6 |

could extract thousands of different 1.8s subintervals that represent the action running. Ideally, the feature representation should describe all of those subintervals with almost the same feature vector.

- Invariant of the actions that precede the action that is currently being predicted. To put it differently, the feature descriptor should be invariant to Markov properties of the time series. Firefighter actions have these characteristics so they should be properly modeled. From the actions listed in table I, it seems highly unlikely that the firefighter can perform all pairs of actions (i.e. states) in sequence with the same probability. On the contrary, some state transitions seem very unlikely.

After the features are engineered and the dataset is processed, the next challenge is how to perform feature selection. This is essential because there is significant amount of features that can have negative impact on the used classification algorithms.

Finally, the last challenge is how to build classification models for the different labels given the training dataset. Table I reveals another significant challenge - the number of labels is large and their distribution is highly unbalanced. In the following sections we describe how we have coped with the above challenges and which results were obtained using different techniques. Even though at this point we have mentioned specific numbers like the number of time series or the number of samples in a time series, the approach is generic and in the remaining of the paper we use parameters for them.

## III. FEATURE ENGINEERING

In order to address the feature engineering for any problem, first we need to understand the nature of the time series data. Time series data can have different time sampling intervals and different scales of the values, however, most of the data can be useful when building a classification model. In the following subsections we describe which methods were used to address different types of challenges in modeling the time series.

### A. *Capturing the sensitivity to change with first derivatives*

By definition first derivatives are used to capture the speed of changes of some function. When instead of a continuous function we have a discrete sample, like a time series, finding an analytic solution is difficult. Nevertheless, we can estimate them. For a time series with $K$ readings that are collected at times $t[i], 0 \leq i < K$, we can calculate $K-1$ first derivatives. In this case $K = 400$, but we also have the timestamps of the readings which show that usually the interval between readings is 4.5ms. Nevertheless, we decided to use the original time stamp in order to calculate the first derivatives more accurately. Eq. 4 shows how the first derivative $fd$ of time series $j$ and at time $t(i), 0 < i \leq N$ can be estimated.

$$fd_j(i) = \frac{reading_j(i) - reading_j(i-1)}{t(i) - t(i-1)} \tag{4}$$

### B. *Modeling seasonality*

Often in time series there are seasonal components that can consist of periodic, repetitive, and generally regular and predictable patterns in the levels of its values. This is especially evident in business data where things like the holidays, days of week, months, quarters have impact on the values in a business time series (e.g. sales, profit, etc). Seasonal effects can conceal both the true underlying movement in the series, as well as certain non-seasonal characteristics which may be of interest to analysts. There are several main reasons for studying seasonal variation:

- Describing the seasonal effect can provide a better understanding of its impact on a time series.
- Eliminating the seasonal component from time series can aid studying other components such as cyclical and irregular variations.
- Use it to build better models for forecasting and prediction of future seasonal trends.

Keeping in mind that many body movements are also periodic, it occurred to us that maybe we should try to discover and model the seasonality in the current problem. We believe that seasonality in this domain should capture the individual characteristics and style of a particular firefighter and/or the context when some action is performed. By context we mean whether the firefighter is rested (e.g. at the beginning of an exercise), tired (after some time performing various actions), etc.

Before continuing to describing methods for modeling seasonality, some components need to be defined:

- The irregular component (sometimes also known as the residual) is what remains after the seasonal and trend components of a time series have been estimated and removed. It results from short term fluctuations in the series which are neither systematic nor predictable. In a highly irregular series, these fluctuations can dominate movements, which will mask the trend and seasonality.
- The trend is defined as the long term movement in a time series without irregular effects (like calendar related effects in business data), and is a reflection of

the underlying level. In financial data it is the result of influences such as population growth, price inflation and general economic changes.

To model a seasonal component, as described in [10], the following methods are usually used:

1) In an additive time-series model, the seasonal component is estimated as defined with eq. (5). There $S$ stands for the seasonal values, $Y$ is for actual data values of the time-series, $T$ is for trend values, $C$ is for cyclical values and $I$ is for irregular values.
$$S = Y - (T + C + I) \tag{5}$$
2) In a multiplicative time-series model, the seasonal component is expressed in terms of ratio and optionally with percentage as in eq. (6):
$$S = \frac{T \times S \times C \times I}{T \times C \times I} \times 100 = \frac{Y}{T \times C \times I} \times 100 \tag{6}$$
3) The deseasonalized time-series data will have only trend ($T$) cyclical($C$) and irregular ($I$) components and is expressed with eq. (7) and (8), respectively:
$$Y - S = (T + S + C + I) - S = T + C + I \tag{7}$$
$$\frac{Y}{S} \times 100 = \frac{T \times S \times C \times I}{S} 100 = (T \times C \times I) \times 100 \tag{8}$$

Additionally there is a pseudo-additive model that can be used, but as it was not explored in this research we do not provide more information about it.

The main challenge is discovering the seasonal index that describes the seasonality. Some of the methods for measuring it are: methods of simple averages, ratio to trend, and ratio to moving average. In order to apply them we would have needed experimentation about the length of the sliding window, for which we did not have time, so we decided to do something simpler. Namely, we have opted to use the additive model because we believed that it will correspond to the nature of the data. In particular, although there are difference between the same actions performed by different individuals, they are only linearly shifted values.

Next, we assumed that the seasonal component for each of the $N$ time series ($N = 42$ in the case-study) in one sample can be modeled with the mean value of the $K$ values ($K = 400$ in the case-study). In like manner, we calculated the deseasonalized values (refereed to as deltas in the remaining of the paper) in each training and test instance as defined with eq. (9), where $0 \leq i < 42$ and $0 \leq j < 400$.
$$Delta_i(j) = reading_i(j) - \frac{1}{K} \sum_{j=0}^{K-1} reading_i(j), \tag{9}$$

The motivation behind this approach came from the competition problem. Namely, the observation that some movements should have higher amplitudes than other (e.g. acceleration during running compared to acceleration during walking or standing). With this approach we hope to capture the characteristics of each movement in a more invariant way. Namely, the logic is that if some firefighter performs the same action with higher amplitudes than another, the actual sensor readings for the two series would differ more then the delta values (see eq. (9)).

After we have modeled the characteristics of each movement and posture with the first derivatives and the deltas, as explained in section III the set of available features in the datasets is comprised of:

- 42 values representing the vital functions (specific only to the current problem).
- $N \times K$ which corresponds to the number of series times the number of samples in the series. In this case, this is $42 \times 400 = 16800$ values for the deltas, which represent the original readings from the sensors. Although we are not including the original time series in the model, we calculate some statistics based on them, as described in the following paragraph.
- $N \times (K - 1)$ which corresponds to the number of series times the number of first derivatives in the series. In this case this is $42 \times 399 = 16758$ values for the first derivatives.
- $N \times K$ which corresponds to the number of series times the number of samples in the series. In this case this is $42 \times 400 = 16800$ values for the deltas.

To summarize, after modeling the sensitivity to change with first derivatives as described in subsection III-A, and modeling the seasonality described in the current section, there are 3 time series: the original sensor readings, the series of first derivatives and the time series of deltas. For each of them we can calculate the minimum, maximum, mean and standard deviation, a total of 4 metrics, which results in $N \times 3 \times 4$ values (504 in this case). With this the total number of available features adds up to $N \times K + N \times (K - 1) + N \times 3 \times 4$. In general, we could enrich the feature set by adding other statistics like first quartile, median, third quartile, interquartile range, skewness, kurtosis, etc, but in this research we have not explored this.

The first obvious problem is the number of features, which is way to high for most machine learning algorithms. More importantly, these features are dependent on the start and alignment of the time series. In section V we discuss how these considerations apply to the competition dataset. In the next subsection III-C we propose a robust method that can address these issues.

### C. Histogram-based modeling of time series

In order to address the issues with the features that are available after modeling the sensitivity to change of the time series and modeling the seasonality, it is evident that we need to perform some transformation. Discrete Fourier Transformation (DFT) [11] converts a finite list of equally spaced samples of a function into the list of coefficients of a finite combination of complex sinusoids that has those same sample values. It is usually used to transform the sampled function from its time to the frequency domain. The obtained list of coefficients would be used as a descriptor for the time series, however, we needed a simpler approach that would be more useful for real time applications.

Nevertheless, DFT lead us to the idea to discretize the values in the time series and then to compute histograms based on it. There are multiple ways in which we can discretize the data, but we decided to apply the simplest one, which is uniform discretization. In order to do that, we needed the minimum and maximum values of series and the number of discretization intervals (referred to as bins in the remaining of the paper). Namely, after calculating the minimum and maximum values of each of the $N$ series of first derivatives and deltas, we only needed to decide how many discretization bins to use. In this case we have $N = 42$ series of first derivatives and $N = 42$ series of deltas. Using more bins means results in more values in the histograms and finer grained granularity, but yields more features. In our tests we have tried using 30, 50 and 100 bins. Somewhat surprisingly, the number of bins did not have a significant impact on the classification results. Our analysis showed that with proper classification model one can achieve good performance using a reasonably small number of bins. We provide more details regarding the influence of the number of discretization bins on the classification performance in section V.

To better explain how the histogram-based approach works let us consider only one time series instance with $N$ values. The following steps should be performed prior applying the transformation:

- Determine the minimum and maximum values of each time series. We need them in order to find on which interval the discretization should be performed.
- Determine the step between discreet values $ds$ based on the number of discretization bins $B$ and the $min$ and $max$ values for the particular time series. This can be calculated with eq. (10)

$$ds = \frac{max - min}{B} \tag{10}$$

After the discretization step, $ds$, is calculated for a particular time series, all training and test instances can then be discretized. For each original value $V$ we can calculate the discrete value $DV$ with eq. (11).

$$DV = -|min| + round(\frac{|min| + V}{ds}, 0) \times ds \tag{11}$$

The next step is to calculate the histogram for the time series. If we use $B$ bins, then the histogram for a particular training instance will have $B$ values. By doing this, from a time series with $N$ values we obtain a histogram of $B$ values, where $N > B$. The $B$ values represent the transformed features which are robust, but are also a significantly reduced representation of the original time series.

To illustrate the transformation let us consider the exemplary time series shown at Fig. 1. Let us assume that we want to discretize the values of this time series (red line) to $B = 7$ bins on the interval $[-3, 3]$. After we calculate the discretization step according to eq. (10) we determine the discrete values according to (11). Using the discrete values (blue line) we can calculate the histogram displayed on Fig. 2. Consequently starting from a time series with 20 values, a histogram of 7 values is obtained.

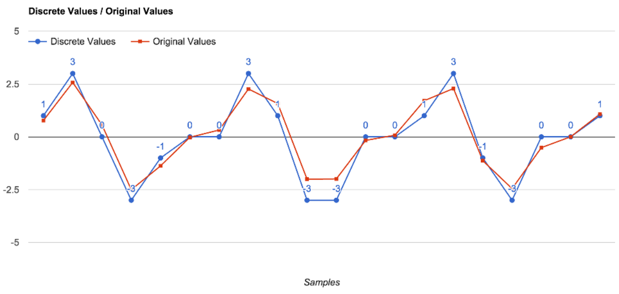


Fig. 1. Original and discrete values of an exemplary time series

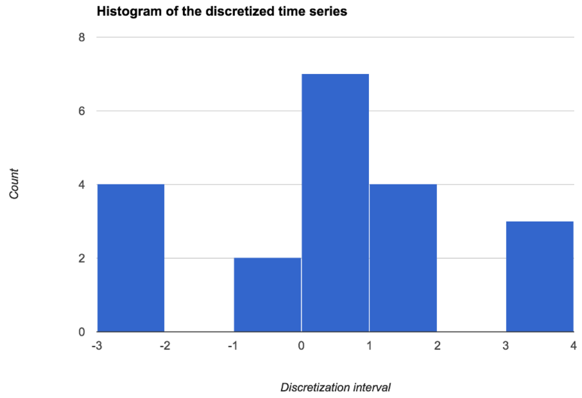


Fig. 2. Histogram of the discrete values of an exemplary time series

After applying the histogram transformation of the time series described above, when using $B$ discretization bins, the following dataset is obtained:

- 42 values representing the vital functions (specific to this case only).
- $2 \times N \times B$ for the first derivatives and the deltas. In this case $N = 42$, so $Hist = 2 \times 1260$ for $B = 30$, $Hist = 2 \times 2100$ for $B = 50$, and $Hist = 2 \times 4200$ for $B = 100$.
- $N \times 3 \times 4$ for the 4 aggregated values of the 3 types of series, which in this case is 504 additional features.

To summarize, after performing the histogram transformation, the total number of features in the dataset of the current problem was 3066, 4746 and 8946 when using 30, 50 and 100 bins, respectively.

### D. Feature selection

After inspection of the transformed dataset it was evident that for some features almost all training instances had the same value. In order to address this, we calculated the variance of each feature in the training set. Using the variance for discarding non-informative features is a simple baseline approach to feature selection. It removes all features whose variance does not meet some threshold. If a feature has the same value in all training samples, then its variance is 0.

When discarding the features with different thresholds for the variance, we have obtained the results in the following table. It can be noted that using a threshold greater than 0.05 does not significantly reduce the number of features. On the contrary, the cross validation results showed decline in performance. Table II shows the number of retained features per discretization bins and variance threshold on the competition datasets.

TABLE II. RETAINED FEATURES PER DISCRETIZATION BINS AND VARIANCE THRESHOLD

| Discretization bins | Features | Variance threshold | Retained features |
|---:|---:|---:|---:|
| 30 | 3066 | 0.05 | 1780 |
| 50 | 4746 | 0.001 | 3196 |
| 50 | 4746 | 0.01 | 2601 |
| 50 | 4746 | 0.03 | 2272 |
| 50 | 4746 | 0.05 | 2137 |
| 50 | 4746 | 0.1 | 1927 |
| 100 | 8946 | 0.1 | 5569 |

In the case when we used 100 bins we applied 0.1 as a variance threshold aiming to discard more features, but still over 5000 features remained. That is why we applied PCA (Principal Component Analysis) [12] and limited the selected features to 2000 explicitly. We have chosen the value of 2000 because it was close to the number of retained features obtained by the variance filter when using smaller number of bins.

When doing feature selection with PCA the results were similar in terms of the actual selected features, but for that the computation time was greater. We acknowledge that with more sophisticated feature selection methods like wrappers we might further lower the dimensionality and improve the classification performance. Nevertheless, the nature of the features is such that the features are different from each other because they represent different things. With this in mind, the expectation is that there are no redundant features in the dataset. Additionally, some machine learning algorithms like Support Vector Machines (SVMs) are able to cope with small number of redundant features without significant degrading of performance.

### *E. Data normalization*

Prior building any models we normalized the data using mean and standard deviation calculated from the training set. This process notably helps the next step while building prediction models. The process is recommended as part of the data preprocessing when using Support Vector Machines (SVM) [13] to make the classification model and is known to decrease the classification error [14] .

## IV. MODEL BUILDING

Prior building any models for the competition challenge one needs to define how the two different labels will be handled. One idea is to build a separate model for each label and then to merge the predictions, but also making sure that there are not any contradictions, i.e. combinations that are not possible. Another idea is to perform hierarchical multi-label classification, so first we would classify the training data based on the first label, and then to perform classification based on the second label. When building the second-level models one can perform one-vs-all classification based on the possible outcomes of the second label assuming the first label was correctly predicted. This approach helps by not needing to explore all possible classes of the second label. Nevertheless, given that the combinations of the second label (16) were close to the number of total combinations (24) and due to the limited time for the competition we did not explore these approaches significantly. Instead we used a one-level classification. The label of the instances is the combination of the two original labels, meaning that our classification model tries to predict both labels at the same time.

The set of features after the feature selection (i.e. preprocessed descriptor) is used to generate a classification model using SVM. SVM generates a classification model by finding the optimal support vectors that divide the feature space with the highest margin between the vector and the nearest points, so that all instances on one side of the support vectors belong to one class, and all other instances belong to other classes. For the purpose of our task we are using SVM with Radial Based Function (Gaussian function) that uses two parameters in the optimization process, C and gamma. Since most of the classification problems are not linearly separable, we are using the Gaussian SVM so that the separation is done in a higher dimension vector space generated with the nonlinear Gaussian kernel. To obtain the best parameters on the given data, we are using grid search as suggested in [15]. During the grid search we use exponentially increasing parameters. We have searched both C and gamma parameters in the interval $10^{-5}$ to $10^{5}$ evaluating $11 \times 11 = 121$ combinations. Then in the intervals that were giving best results we conducted finer grained grid search with smaller steps for gamma and C. Depending on the different training datasets (varied by the number of discretization bins and the variance threshold for feature selection) we have obtained different optimal values. For C the optimal values ranged from 1 to 1000 and for gamma they were usually from $10^{-5}$ to $10^{-4}$. Also important to realize is the unbalanced distribution on the classes. In order to address this problem, we used the class distribution to automatically adjust the weights inversely proportional to class frequencies. The estimated weights are then multiplied by the C parameter and those weighted C parameters are used when building the one-vs-all classification models for multi-class classification. This significantly improved the performance of the classifiers when using cross validation, but more importantly on the leaderboard set.

## V. EVALUATION OF THE PROPOSED METHODS ON THE COMPETITION DATASET

For the current challenge, we began our analysis with plotting and visual inspection of some of the sensor readings for random training samples. Before inspecting the data the impression was that the interval of 4.5 milliseconds between different readings could be too short for the sensors to output different values. On the contrary, the analysis showed that there are evident differences between consecutive readings, meaning that indeed all values in the series are potentially useful for building prediction models.

After we have modeled the characteristics of each movement and posture with the first derivatives and the deltas, as explained in subsections III-A and III-B the set of available features in the datasets is comprised of:

TABLE III. SCORE ON THE LEARDERBOARD DATASET DEPENDING ON VARIOUS DISCRETIZATION BINS AND SVM CONFIGURATIONS

| Id | Leaderbord Score | Bins | Variance | Retained features | SVM with RBF kernel |
|---|---|---|---|---|---|
| 1 | 0.8502 | N/A | N/A | N/A | Combination of configurations in rows 2 and 3 (below) |
| 2 | 0.8381 | 50 | 0.05 | 2137 | C=10 gamma=0.00002 weight=auto |
| 3 | 0.8380 | 30 | 0.05 | 1780 | C=10 gamma=0.000035 weight=auto |
| 4 | 0.8379 | 50 | 0.05 | 2137 | C=10 gamma=0.000015 weight=auto |
| 5 | 0.8376 | 30 | 0.05 | 1780 | C=10 gamma=0.00004 weight=auto |
| 6 | 0.8371 | 50 | 0.05 | 2137 | C=10 gamma=0.000025 weight=auto |
| 7 | 0.8371 | 50 | 0.05 | 2137 | C=20 gamma=0.00001 weight=auto |
| 8 | 0.8349 | 50 | 0.05 | 2137 | C=10 gamma=0.00004 weight=auto |
| 9 | 0.8342 | 30 | 0.05 | 1780 | C=10 gamma=0.00002 weight=auto |
| 10 | 0.8341 | 50 | 0.05 | 2137 | C=20 gamma=0.00002 weight=auto |
| 11 | 0.8337 | 50 | 0.1 | 1927 | C=10 gamma=0.00002 weight=auto |
| 12 | 0.8313 | 50 | 0.05 | 2137 | C=10 gamma=0.00001 weight=auto |
| 13 | 0.8308 | 30 | 0.05 | 1780 | C=10 gamma=0.00007 weight=auto |
| 14 | 0.8302 | 30 | 0.05 | 1780 | C=10 gamma=0.00008 weight=auto |
| 15 | 0.8288 | 100 | 0.1 | 2000 (reduced from 5569 with PCA) | C=10 gamma=0.0001 weight=auto |
| 16 | 0.8279 | 100 | 0.1 | 2000 (reduced from 5569 with PCA) | C=10 gamma=0.00004 weight=auto |
| 17 | 0.8257 | 30 | 0.05 | 1780 | C=10 gamma=0.0001 weight=auto |
| 18 | 0.8251 | 50 | 0.1 | 1927 | C=10 gamma=0.00004 weight=auto |
| 19 | 0.8225 | 100 | 0.1 | 2000 (reduced from 5569 with PCA) | C=10 gamma=0.00002 weight=auto |
| 20 | 0.8198 | 100 | 0.1 | 2000 (reduced from 5569 with PCA) | C=10 gamma=0.00001 weight=auto |
| 21 | 0.8191 | 30 | 0.05 | 1780 | C=10 gamma=0.0001 weight=auto |
| 22 | 0.8167 | 50 | 0.05 | 2137 | C=10 gamma=0.0001 weight=auto |
| 23 | 0.8110 | 100 | 0.1 | 2000 (reduced from 5569 with PCA) | C=1.0 gamma=0.0001 weight=auto |
| 24 | 0.8077 | 50 | 0.05 | 2137 | C=1000 gamma=0.0001 weight=auto |
| 25 | 0.8062 | 50 | 0.05 | 2137 | C=10 gamma=0.0001 |

- 42 values representing the vital functions (specific to this case only).
- $2 \times N \times B$ for the first derivatives and the deltas. In this case $N = 42$, so $Hist = 2 \times 1260$ for $B = 30$, $Hist = 2 \times 2100$ for $B = 50$, and $Hist = 2 \times 4200$ for $B = 100$.
- $N \times 3 \times 4$ for the 4 aggregated values of the 3 series, which in this case is 504 additional features.

The first obvious problem is the number of features, which is way to high for most machine learning algorithms. More importantly, these features are dependent on the start of the time series. For instance, let us assume that we have an action (e.g. running) that is being performed with a total duration of 10000 milliseconds (10 seconds). If we extract two samples (training or test instances) from it, sample A spreading from $t_1 = 0ms$ to $t_2 = 1800ms$, and sample B spreading from $t_3 = 9ms$ to $t_4 = 1809ms$. The logic is that those two samples are nearly identical and should be treated accordingly in a good feature space. However, in the current representation they will be different because they are shifted by 2 periods of 4.5ms. Obviously this needs to be addressed, and this is performed by the histogram-based method described in subsection III-C. Then with the method described in subsection III-D we have reduced the number of features as shown with table II.

We have obtained many similar results based on the different alternatives we tried (different C and gamma parameters, number of discretization bins, etc.) on both the leaderboard dataset and with 5-fold cross validation with the training set. Table III shows some of the more significant configurations ordered by the score on the leaderboard dataset in descending order. Before explaining the best score, which is shown in the first line in this table, we first want to discuss the scores of the individual classifier configurations shown in all other rows.

The first obvious thing to notice is that when applying weights (denoted as “weight=auto” in table III to the C parameters improves the performance of the classifiers. Namely row 25 does not have weights applied to the C parameter, meaning the value of C=10 is used for all classes in the one-vs-all multi-class SVM. When applying weights proportional to the class frequencies in the training dataset the performance is improved as shown in row 22. We have noticed the same pattern in other configurations (i.e. different bin size, different C and gamma parameters) which are not shown in this table.

Next, we have noticed that when varying the values of the C and gamma parameters the classification score also varies when using the same feature set. This was expected and confirms the need for grid search, as described in section IV, to find the optimal values for those parameters.

Finally, the most important realization is that the greater number of discretization bins does not necessarily mean the score will be better. Most compelling evidence to this statement are the two best configurations shown in rows 2 and 3. Row 2 uses 50 bins, while row 3 uses 30 bins, but the difference between their score is 0.0001 in favour of the 50 bins feature set. On the other hand, the 30 bins feature set has 357 features less, which in turn leads to less training and test time. Another evidence that supports this claim is the case when we use 100 bins (rows 15, 16, 19, 20, 23). In those cases we obtain a much greater number of features than when using 50 or 30 bins, so we have to perform additional feature selection in order to reduce the training time. If we have retained more features and performed more detailed grid search for those feature sets, the results might have been better. Our initial tests showed that this in fact leads towards over-fitting and generating too many support vectors. For this reason this did not seem like worth doing, especially because we have managed to find better solutions with significantly less features.

The score difference between different configurations may seem negligible, but when looking at the actual predictions made by the 2 best models (rows 2 and 3) we came to an important realization. Namely, it was evident that both cross-validation and test predictions made by the 2 models are significantly different, yet the score of the 2 models was similar. This gave us an idea to train a second-level (ensemble) model. The features for this model were the predictions and probabilities of the 2 individual classifiers made with cross validation on the training dataset and the class was combination of the two labels. By applying this we further improved the score on the leaderboard dataset, and our final solution had a score of 0.8502. It was the third-placed score on the leaderboard and was only 0.001 behind the second-placed solution, 0.008 behind the first place, and 0.025 better then the fourth-placed solution. After the final results were published our score was still third - 0.8261, while the first was 0.8391, the runner-up was 0.82985 and the fourth placed score was 0.80408. We acknowledge that combining more individual models should be explored with a more generic approach.

## VI. Conclusion and future work

Based on the obtained results, we can conclude that the proposed descriptor gives a good model for the movements, while providing some invariance to deviation of the input values and also performing significant feature reduction. During the analysis of the predictions made by the different classification models we observed that most of the miss-classification errors were made when the classification model was trying to distinguish between very similar tasks for the given dataset, such as *moving* + *stairs-up* vs *moving* + *stairs-down* and *moving* + *manipulating* vs *crouching* + *manipulating*. These errors were inevitable since our approach for generating the descriptors is amplitude based and some specific movements have very similar amplitude characteristics. We believe that there is room for further improvement if we add additional descriptors to the feature space like: other statistics to describe the time series, amplitudes of the acceleration across all 3 axes, discovering and modeling the trend in the series, trying logarithmic transformations, adding second order derivatives, etc. Performing a better feature selection may also improve the performance. Finally, a better modeling that takes into account the hierarchical nature of the classification problems, as discussed previously, can further improve the results.

The main contribution of this paper is the proposed method for invariant modeling the time series by using first derivatives and deltas. Moreover, the novelty of our approach is in the proposed histogram-based method for feature reduction of the time series. Even though it was developed during the competition, it is applicable to time series regardless from their domain. In our future research we plan to affirm this method by analyzing time series from various domains and comparing it to other methods for modeling time series.

## Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss.Cyril and Methodius University, Skopje, Macedonia and by NI TEKNA - Intelligent technologies, Negotino, Macedonia.

## References

[1] S. Movassaghi, M. Abolhasan, J. Lipman, D. Smith, and A. Jamalipour, "Wireless body area networks: A survey," pp. 1658–1686, Third 2014.

[2] T. chung Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164 – 181, 2011. doi: http://dx.doi.org/10.1016/j.engappai.2010.09.007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197610001727

[3] P. Esling and C. Agon, "Time-series data mining," *ACM Comput. Surv.*, vol. 45, no. 1, pp. 12:1–12:34, Dec. 2012. doi: 10.1145/2379776.2379788. [Online]. Available: http://doi.acm.org/10.1145/2379776.2379788

[4] B. Hu, Y. Chen, and E. Keogh, "Classification of streaming time series under more realistic assumptions," *Data Mining and Knowledge Discovery*, pp. 1–35, 2015. doi: 10.1007/s10618-015-0415-0. [Online]. Available: http://dx.doi.org/10.1007/s10618-015-0415-0

[5] M. Meina, A. Janusz, K. Rykaczewski, D. Slezak, B. Celmer, and A. Krasuski, "Tagging firefighter activities at the emergency scene: Summary of AAIA'15 data mining competition at Knowledge Pit," in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2015, in print September 2015.

[6] A. Krasuski, "A framework for dynamic analytical risk management at the emergency scene. from tribal to top down in the risk management maturity model," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*. IEEE, 2014, pp. 323–330.

[7] A. Krasuski, A. Jankowski, A. Skowron, and D. Slezak, "From sensory data to decision making: A perspective on supporting a fire commander," in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. IEEE, 2013, pp. 229–236.

[8] M. Meina, B. Celmer, and K. Rykaczewski, "Towards robust framework for on-line human activity reporting using accelerometer readings," in *Active Media Technology*. Springer, 2014, pp. 347–358.

[9] "ICRA' project," http://www.icra-project.org/, accessed: 2015-06-05.

[10] A. G. Barnett and A. J. Dobson, *Analysing Seasonal Health Data*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2010. ISBN 9783642107481 3642107486

[11] R. Agrawal, C. Faloutsos, and A. N. Swami, "Efficient similarity search in sequence databases," in *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*. Springer-Verlag, 1993, pp. 69–84.

[12] I. Jolliffe, "Principal Component Analysis," in *Wiley StatsRef: Statistics Reference Online*, N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels, Eds. Chichester, UK: John Wiley & Sons, Ltd, Sep. 2014. ISBN 9781118445112. [Online]. Available: http://doi.wiley.com/10.1002/9781118445112.stat06472

[13] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011. doi: 10.1145/1961189.1961199. [Online]. Available: http://doi.acm.org/10.1145/1961189.1961199

[14] D. M. Tax and R. P. Duin, "Feature scaling in support vector data descriptions," Technical report, Technical report, American Association for Artificial Intelligence, Tech. Rep., 2000.

[15] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification."

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 389–396

DOI: 10.15439/2015F423
ACSIS, Vol. 5

# Tagging Fireworkers Activities from Body Sensors under Distribution Drift

Marc Boullé
Orange Labs,
2 avenue Pierre Marzin, 22300 Lannion, France
http://www.marc-boulle.fr
Email: marc.boulle@orange.com

***Abstract*—We describe our submission to the AAIA'15 Data Mining Competition, where the objective is to tag the activity of firefighters based on vital functions and movement sensor readings. Our solution exploits a selective naive Bayes classifier, with optimal preprocessing, variable selection and model averaging, together with an automatic variable construction method that builds many variables from time series records. The most challenging part of the challenge is that the input variables are not independent and identically distributed (i.i.d.) between the train and test datasets. We suggest a methodology to alleviate this problem, that enabled to get a final score of 0.76 (team marcb).**

## I. Introduction

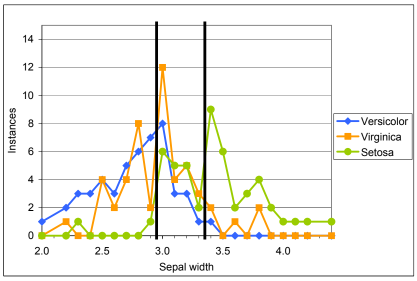
The AAIA'15 Data Mining Competition [1] is related to a problem of activity tagging. Firefighters are equipped with body sensors that register vital functions and movements. Vital function records are summarized by statistics (minimum, maximum, median...) using a fixed number of input variables, whereas movement records are available as 42 times series of length 1.8 s with measures every 4.5 ms. Train data consists of 20,000 samples for activities recorded from four firefighters, whereas the test data contains 20,000 samples coming from a different group of four firefighters. The objective is to tag the activity of firemen, among 24 activities, and the evaluation criterion is the balanced accuracy (BAC). In this paper, we present our submission to the challenge. It exploits a Selective Naive Bayes classifier together with an automatic variable construction method (Section II). We motivate the choice of this classification framework and describe its application to the challenge in Section III. A good classifier trained on the train data obtains a disastrous leaderboard score. This is not caused by over-fitting, but by a severe distribution drift between the train and test data. We suggest in Section IV a methodology to alleviate this problem. In Section V, we present related work and discuss our approach. Finally, Section VI summarizes the paper.

## II. Supervised Classification Framework

We summarize the Selective Naive Bayes (SNB) classifier introduced in [2]. It extends the Naive Bayes classifier [3] owing to an optimal estimation of the class conditional probabilities, a Bayesian variable selection and a Compression-based Model Averaging. We also describe the automatic variable construction framework presented in [4], used to get a tabular representation from times series.

### A. Optimal discretization

The Naive Bayes (NB) classifier has proved to be very effective in many real data applications [3], [5]. It is based on the assumption that the variables are independent within each class, and solely relies on the estimation of univariate conditional probabilities. The evaluation of these probabilities for numerical variables has already been discussed in the literature [6], [7]. Experiments demonstrate that even a simple equal width discretization brings superior performance compared to the assumption using a Gaussian distribution per class. Using a discretization method, each numerical variable is recoded as a categorical variable, with a distinct value per interval. Class conditional probabilities are assumed to be piecewise constant per interval, and obtained by counting the number of instances per class in each interval. These class conditional probabilities are used as inputs for the naive Bayes classifier. Figure 1 shows an example of discretization into three intervals, for the *Sepal width* input variable of the Iris dataset [8].



| Total | 57 | 57 | 36 |
|---|---|---|---|
| Versicolor | 34 | 15 | 1 |
| Virginica | 21 | 24 | 5 |
| Setosa | 2 | 18 | 30 |
| | ]-∞, 2.95[ | [2.95, 3.35[ | [3.35, +∞[ |

Fig. 1. Number of instances per class in the Iris dataset, for a discretization of the *Sepal width* input variable into three intervals

978-83-60810-66-8/$25.00©2015, IEEE

In the MODL approach [9], the discretization is turned into a model selection problem and solved in a Bayesian way. First, a space of discretization models is defined. The parameters of a specific discretization model $M$ are the number of intervals, the bounds of the intervals and the class frequencies in each interval. Then, a prior distribution is proposed on this model space. This prior exploits the hierarchy of the parameters: the number of intervals is first chosen, then the bounds of the intervals and finally the class frequencies. The choice is uniform at each stage of the hierarchy. Finally, the multinomial distributions of the class values in each interval are assumed to be independent from each other. A Bayesian approach is applied to select the best discretization model, which is found by maximizing the maximum a posteriori (MAP) model. Owing to the definition of the model space and its prior distribution, the Bayes formula is applicable to derive an exact analytical criterion to evaluate the posterior probability of a discretization model. The optimized criterion is $p(M)p(D|M)$, where $p(M)$ is the prior probability of a preprocessing model and $p(D|M)$ the conditional likelihood of the data given the model.

Efficient search heuristics allow to find the most probable discretization given the data sample. Extensive comparative experiments report high performance.

*Univariate informativeness evaluation:* A 0-1 normalized version of the optimized criterion provides a univariate informativeness evaluation of each input variable. Taking the negative log of the MAP criterion, $c(M) = -(\log p(M) + \log p(D|M))$, the approach receives a Minimim Description Length (MDL) [10] interpretation, where the objective is to minimize the coding length of the model plus that of the data given the model. The null model $M_\emptyset$ is the preprocessing model with one single interval, which represents the case with no correlation between the input and output variables. We then introduce the *I(V)* criterion in Equation 1 to evaluate the informativeness of a variable $V$.

$$I(V) = 1 - \frac{c(M)}{c(M_\emptyset)}. \tag{1}$$

The value of $I(V)$ grows with the informativeness of an input variable. It is a between 0 and 1, 0 for irrelevant variables uncorrelated with the target variable and 1 for variables that perfectly separate the target values.

### B. Bayesian Approach for Variable Selection

The naive independence assumption can harm the performance when violated. In order to better deal with highly correlated variables, the Selective Naive Bayes approach [11] exploits a wrapper approach [12] to select the subset of variables which optimizes the classification accuracy. Although the Selective Naive Bayes approach performs quite well on datasets with a reasonable number of variables, it does not scale on very large datasets with hundreds of thousands of instances and thousands of variables, such as in marketing applications or text mining. The problem comes both from the search algorithm, whose complexity is quadratic in the number of variables, and from the selection process which is prone to overfitting. In [2], the overfitting problem is tackled by relying on a Bayesian approach, where the best model is found by maximizing the probability of the model given the data. The parameters of a variable selection model are the number of selected variables and the subset of variables. A hierarchic prior is considered, by first choosing the number of selected variables and second choosing the subset of selected variables. The conditional likelihood of the models exploits the Naive Bayes assumption, which directly provides the conditional probability of each class. This allows an exact calculation of the posterior probability of the models. Efficient search heuristic with super-linear computation time are proposed, on the basis of greedy forward addition and backward elimination of variables.

### C. Compression-Based Model Averaging

Model averaging has been successfully exploited in bagging [13] using multiple classifiers trained from re-sampled datasets. In this approach, the averaged classifier uses a voting rule to classify new instances. Unlike this approach, where each classifier has the same weight, the Bayesian Model Averaging (BMA) approach [14] weights the classifiers according to their posterior probability. In the case of the Selective Naive Bayes classifier, an inspection of the optimized models reveals that their posterior distribution is so sharply peaked that averaging them according to the BMA approach almost reduces to the MAP model. In this situation, averaging is useless. In order to find a trade-off between equal weights as in bagging and extremely unbalanced weights as in the BMA approach, a logarithmic smoothing of the posterior distribution, called Compression-based Model Averaging (CMA), is introduced in [2]. The weighting scheme on the models reduces to a weighting scheme on the variables, and finally results in a single Naive Bayes classifier with weights per variable. Extensive experiments demonstrate that the resulting Compression-based Model Averaging scheme clearly outperforms the Bayesian Model Averaging scheme. In the rest of the paper, the classifier resulting from model averaging is called Selective Naive Bayes (SNB).

### D. Automatic Variable Construction for Multi-Table

In a data mining project, the data preparation phase aims at constructing a data table for the modeling phase [15], [16]. The data preparation is both time consuming and critical for the quality of the mining results. It mainly consists in the search of an effective data representation, based on variable construction and selection. Variable construction [17] has been less studied than variable selection [18] in the literature. However, learning from relational data has recently received an increasing attention. The term Multi-Relational Data Mining (MRDM) was initially introduced in [19] to address novel knowledge discovery techniques from multiple relational tables. The common point between these techniques is that they need to transform the relational representation. Methods named by propositionalisation [20], [21], [22] try to flatten the relational data by constructing new variables that aggregate the information contained in non target tables in order to obtain a classical tabular format.

In [4], an automatic variable construction method is proposed for supervised learning, in the multi-relational setting using a propositionalisation-based approach. Domain knowledge is specified by describing the multi-table structure of the data and choosing construction rules. The formal description of the data structure relies on a root table that contains the main statistical units and several secondary tables in 0 to 1 or 0 to n relationship with the root table. For example, Figure 2 describes the structure of the data for the challenge. The construction rules available for automatic construction of variables are detailed below:

- *Selection(Table, Num)→Table*: selection of records from a secondary table according to a conjunction of selection terms (membership in a numerical interval of a variable *Num* in the secondary table),
- *Count(Table)→Num*: count of records in a table,
- *Mean(Table, Num)→Num*: mean value of variable *Num*,
- *Median(Table, Num)→Num*: median value,
- *Min(Table, Num)→Num*: min value,
- *Max(Table, Num)→Num*: max value,
- *StdDev(Table, Num)→Num*: standard deviation,
- *Sum(Table, Num)→Num*: sum of values.

The space of variables that can be constructed is virtually infinite, which raises both combinatorial and over-fitting problems. When the number of original or constructed variables increases, the chance for a variable to be wrongly considered as informative becomes critical. A prior distribution over all the constructed variables is introduced. This provides a Bayesian regularization of the constructed variables, which allows to penalize the most *complex* variables. An effective algorithm is introduced as well to draw samples of constructed variables from this prior distribution. Experiments show that the approach is robust and efficient.

## III. Applying the Framework for the Challenge

We motivate our choice of the classification framework[1], then describe how we apply it on the challenge dataset.

### A. Choice of the classification framework

In all our challenge submissions, we exploit the framework described in Section II to train a selective naive Bayes classifier, with optimal discretization, variable selection and model averaging. The classifier is trained on a flat data representation, obtained using the automatic variable construction method (Section II-D) that builds many variables from the time series records data. Once the data schema is specified, the only parameter is the number of variables to construct. The method is fully automatic, scalable and highly robust, with test performance mainly equivalent to train performance.

The SNB classifier is resilient to noise and to redundancies between the input variables, but it is blind to non-trivial interactions between the variables. This can be leveraged by feature engineering, relying on domain expertise rather than on statistical expertise. More accurate classification methods are available, such as random forests, gradient boosting methods, support vector machines or neural networks. However, these methods require intensive feature engineering to get a flat input data table representation, are prone to over-fitting, are mainly black-box, not suitable for an easy interpretation of the models and finally require fine parameter tuning, both time consuming and expertise intensive. In an industrial context like the Orange telecommunication operator, the major issue is to quickly provide an accurate, robust and interpretable solution to many data mining problems, rather than a very accurate solution to few problems. In this context, the generic framework described in Section II and used in this challenge offers a good solution.

### B. Application to the challenge dataset

For the AAIA'15 Data Mining Competition, firefighters are described using a root table that contains the target activity as well as the summary variables for the vital function sensors and a secondary table for the movement sensors readings. An identifier variable *Id* is added in each record of both tables, to enable the join between the root and secondary tables.

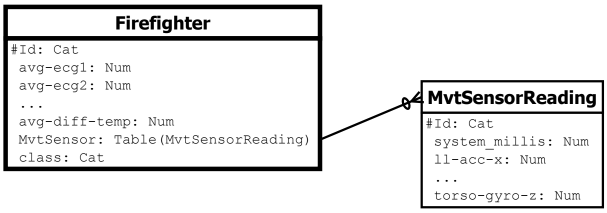


Fig. 2. Multi-table representation for the data of the AAIA'15 challenge

The multi-table representation of the challenge data is presented in Figure 2. The root table (*Firefighter*) contains 20,000 instances, with 44 variables: *Id*, the 42 vital function input variables (*avg-ecg1*, *avg-ecg2*, ... *avg-diff-temp*) and the class variable. The secondary table (*MvtSensorReading*) contains 20,000 × 400 records, with 44 variables: *Id* as a join key and the 43 time series variables for the movement sensor readings (*system_millis*, *ll-acc-x*, ... *torso-gyro-z*). Using the data structure presented in Figure 2 and the construction rules introduced in Section II-D, one can for example construct the following variables ("name" = *formula*: comment) to enrich the description of a *Firefighter*:

- "StdDev(MvtSensor.rl-acc-x)" =
  *StdDev(MvtSensor, rl-acc-x)*:
  standard deviation of rl-acc-x in the sensor readings,
- "Count(MvtSensor) where ll-gyro-y $>$ 60" =
  *Count(Selection(MvtSensor, ll-gyro-y $>$ 60))*:
  number of sensor readings where ll-gyro-y $>$ 0,
- "Max(MvtSensor.torso-acc-z) where torso-acc-x $>$ 5" =
  *Max(Selection(MvtSensor, torso-acc-x $>$ 5), torso-acc-z)*:
  max of torso-acc-z for sensor readings where torso-acc-x $>$ 5.

[1] Available as a shareware at http://www.khiops.com

The number of variables to construct is the only user parameter. An input flat data table representation is then obtained from the initial input variables coming from the root table and the set of all automatically constructed variables. All these variables are then preprocesses using the optimal discretization method (cf. SectionII-A) to assess their informativeness and evaluate their class conditional probabilities, before training the SNB classifier. For advanced use, it is possible to impose a constraint on the granularity of the discretizations (cf. Section II-A): instead of obtaining the optimal number of intervals, the preprocessing method output discretizations with at most $I_{Max}$ intervals, where $I_{Max}$ is a user parameter.

## IV. Challenge submission

In this section, we describe our submissions to the challenge and suggest a methodology to alleviate the problem of the drift between the train and test distributions of the challenge dataset.

### A. *First trials*

*a) Submission 1:* To get familiar with the challenge evaluation protocol, we made a first quick trial, using only the 43 vital function variables. We obtained a surprisingly high train accuracy, with few over-fitting: 0.9840 and 0.9680 on a 70%-30% split of the train dataset. However, our first submission obtained only a 0.1859 score on the challenge leaderboard. This dramatic drop of performance was not caused by overfitting, but by a drift between the train and test distributions (based on two different groups of four firefighters).

*b) Submission 2:* We then generated 100 additional variables to summarize the movement sensors times series, using the framework described in Section II-D. As we observed that the optimal discretizations (see Section II-A) were very fine grained, with up to hundred of intervals, we decided to constrain the discretization method to build at most 10 intervals. We obtained a 0.9499 train accuracy (on the test split of the train dataset), and a 0.4566 leaderboard score. Constraining the discretisations thus reduced the drift effect.

*c) Submission 3:* Using discretizations with at most two intervals, we obtained a 0.8603 train accuracy and a 0.6372 leaderboard score. This confirmed the benefit of the constrained discretisations to reduce the drift effect.

*d) Submission 4.:* Still using discretizations with at most two intervals, we generated 1,000 variables from the movement sensors times series. We obtained a 0.9254 train accuracy and a 0.6951 leaderboard score.

Table I summarizes the performance obtained for each preliminary submission as well as the related user parameters: number of constructed variables (cf. Section II-D) and constrain on the maximum number of intervals in the discretizations (cf. variable preprocessing in SectionII-A).

These preliminary trials took only one hour and gave interesting insights.

### B. *Analysis, trials and errors*

Let us consider two tasks: classification of the activity and detection of the drift. The drift detection task can be turned into a classification task as in [23], by merging the train and test datasets and using the dataset label ('train' or 'test') as the target variable. Intuitively, if we are able to select an input representation with good classification accuracy on the train dataset but poor drift detection, we expect that our classifier will be less sensitive to drift and that the performance drop on the test dataset will be reduced.

TABLE I
Method parameters and performance per submission

| Submission | Constructed variables | Interval max nb | Train accuracy | Leaderboard score |
|---|---|---|---|---|
| Submission 1 | 0 | | 0.9680 | 0.1859 |
| Submission 2 | 100 | 10 | 0.9499 | 0.4566 |
| Submission 3 | 100 | 2 | 0.8603 | 0.6372 |
| Submission 4 | 1000 | 2 | 0.9254 | 0.6951 |

The objective is then to explore varying input representations and select the one with the best classification accuracy together with the poorest drift detection. To do so, we mainly considered the following dimensions: max number of intervals for discretizations, representation of times series, selection of variables, both for the root and secondary tables, choice of construction rules, number of constructed variables. Exploiting the informativeness of variables both from the classification and drift detection tasks, we made many trials and errors and finally obtained a solution with a leaderboard score of 0.7892. This solution mainly consists of:

1) discretizations with at most two intervals,
2) no use of any vital function variables, selection of part of the movement variables (mainly, the *acc-x* and *acc-z* for the *torso*, the *acc-x, acc-z, gyro-x, gyro-y* for the *legs* and *acc-x, gyro-x, gyro-z* for the *hands*),
3) construction of 10,000 variables using only the *Count* and *Selection* construction rules.

### C. *A methodology to reduce the drift problem*

While the approach described in the preceding section was insightful and allowed to improve significantly the leaderboard score, it heavily relies on human expertise and is time consuming. We now suggest a methodology that aims at automatizing the approach. First, we use all the initial input representation and all the available construction rules to build 10,000 variables. Using discretizations with at most two intervals, we evaluate the informativeness of each input variable (cf. Formula 1), both for the drift detection and the classification tasks. The results, displayed in Figure 3, show that there are variables with large drift informativeness and small classification informativeness (top-left of the figure), or on the contrary variables with small drift informativeness and large classification informativeness (bottom-right). The interesting variables are those close to the X axis, with small drift informativeness.

We then sort the variables by increasing drift informativeness and select subsets of variables of increasing sizes, for a list of thresholds of drift informativeness (0, 0.0001, 0.00025...). Figure 4 displays the number of informative

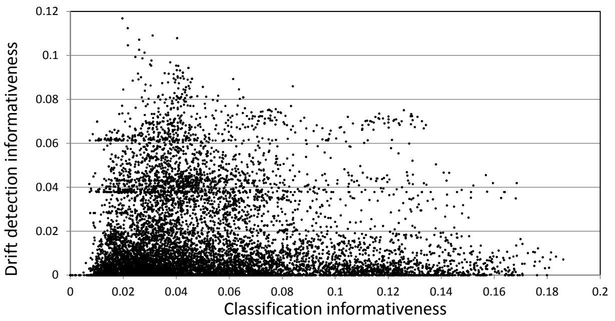


Fig. 3. Informativeness of 10,000 variables

variables for both the classification and detection tasks, for each drift threshold. For example, using a threshold of 0 that excludes any variable with drift informativeness, the smallest selected subset contains around 2,000 variables that are informative for the classification task.

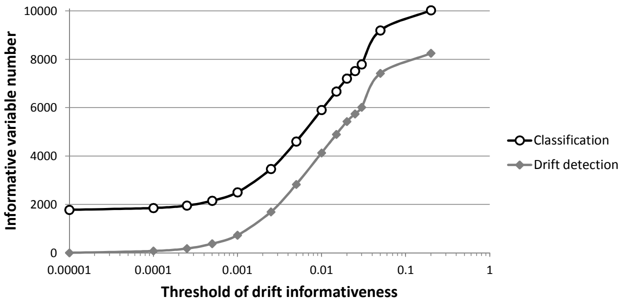


Fig. 4. Number of informative variables per drift threshold for the classification and drift detection tasks

For each of these subsets, we train the selective naive Bayes classifier both for the classification and drift detection tasks, collect the resulting accuracies as well as the leaderboard score obtained with the related submissions. For both tasks, we use a 70%-30% split of the data to evaluate the robustness of the classifier, and always observe a small difference ($\approx 1\%$) between the train and test performance.

Training times are more than ten times longer for the classification task (24 target values) than for the drift detection task (two labels). The largest classification task consists of 14,000 train instances (70%*20,000) and 10,000 constructed variables, each summarizing 400 movement sensor readings. On a PC Windows with Intel Xeon 2.3 Ghz processor, it took about half an hour for the data preparation and four hours to train the SNB classifier.

Figure 5 shows that the accuracy of drift detection rapidly decreases with smaller number of variables while the decrease is slower for the train classification accuracy. Meanwhile, the leaderboard score increases, from 0.7083 using all the variables to a plateau with a leaderboard score of about 0.78 between 2,500 and 6,000 variables . Using this methodology, our final chosen submission obtained a leaderboard score of 0.7856 (using around 4,500 variables) and a final score of 0.76.

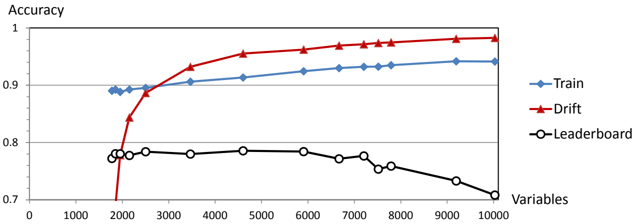


Fig. 5. Train classification, drift detection and leaderboard accuracies w.r.t. number of input variables

### D. *Insights on relevant variables*

Looking at Figure 4, we can see that the selected subset that contains around 4,500 variables is related to a drift threshold of 0.005, which is very small. Figure 3 shows that this corresponds to variables very close from the X axis, with potentially large class informativeness but very small drift detection informativeness.

For interpretability purposes, it is interesting to further investigate on which kind of variables are kept in the best classifier. In Figure 6, we reuse the same kind of plot as in Figure 3 and focus on the initial data representation, per family of variables. The 42 vital function variables are summarized in five families: EGC, heart rate, breath rate, respiration and temperature. The figure shows that these variables (especially the heart rate ones) have large drift informativeness and small classification informativeness. They are excluded from the best classifier. As for the 42 movement sensor time series, we divided the analysis using separate plots for the 7 body parts (torso, left and right hand, arm and legs) with 6 families per body part: x, y and z readings for the accelerometer and gyroscope. From the 10,000 automatically constructed variables, we collected the subset of variables related to each body part per family. The results are summarized in the 7 body part plots in Figure 6. This brings interesting insights w.r.t. to the relevance of each movement sensor for the classification task. For example, the arm sensors are the least interesting, since they are related to many variables with large drift informativeness. On the opposite, the other body part sensors are related to many constructed variable close to the X axis, with small drift informativeness and large classification informativeness. Overall, the torso and leg movement sensor bring the most useful information. The dissimilarity between the left and right body part sensors appears clearly. For example, for the left hand, the z accelerometer is too sensible to the drift, and for the left leg, this the case for the y accelerometer. All these insights may be useful to optimize data collection and pre-processing in order to improve the performance of the classifier.

### E. *Limits of the approach*

Using the leaderboard to chose the best solution is likely to overfit the test dataset, and the solution might not be reliable

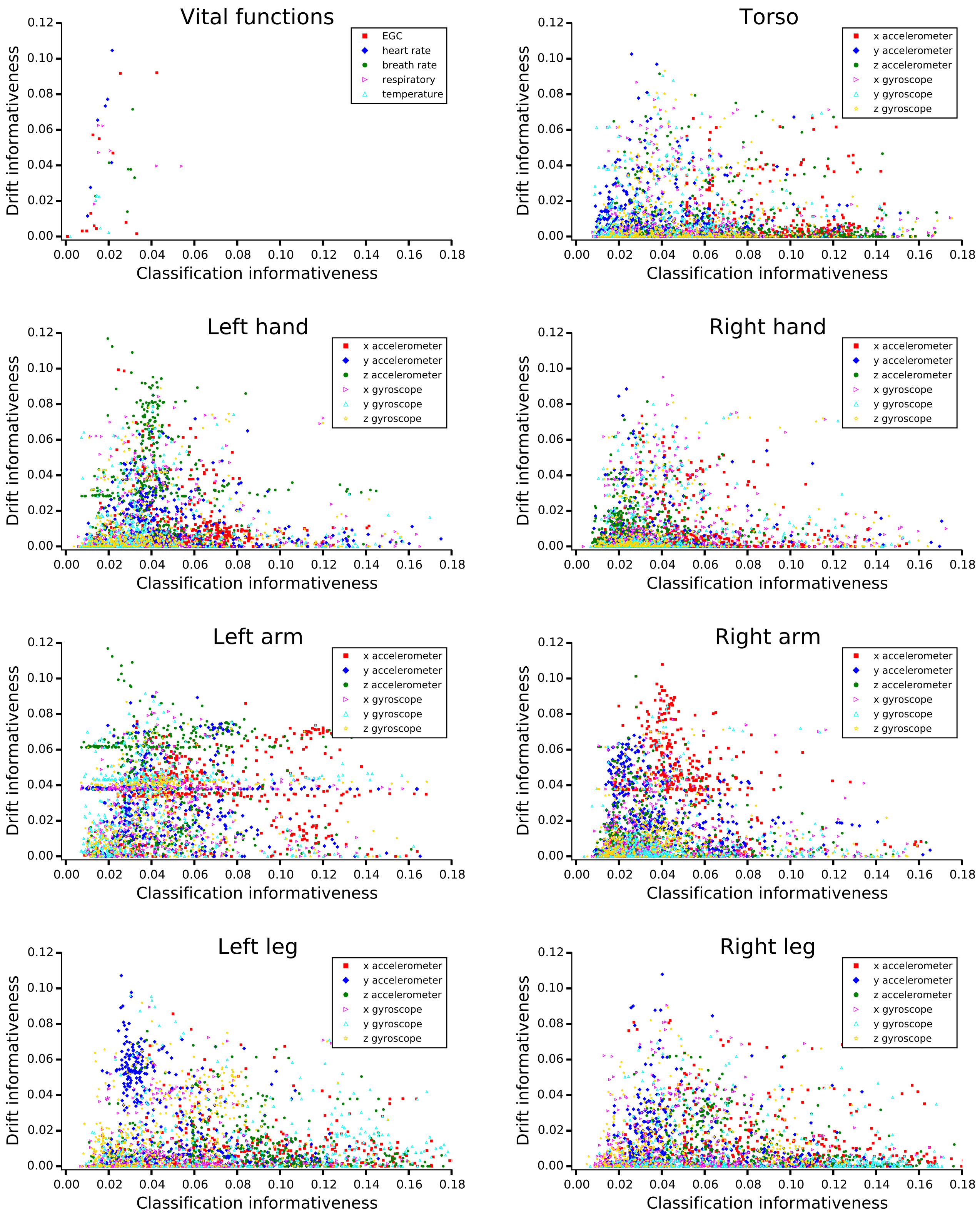


Fig. 6. Informativeness per family of variables

when applied to new firefighters not in the train nor test datasets. Ideally, data should be collected from more distinct firefighters to improve the reliability and performance of the solution. Also, the firefighter identifiers should be available in the datasets. This would enable a cross-validation process with splits based on distinct firefighters, to select the best solution using the train dataset only.

## V. Related work and discussion

The challenge settings rely on a train and test datasets with the same task. As in supervised learning, labels are available in the train dataset and not available in the test dataset. However, the train and test data do not come from the same distribution (different firefighters in train and test).

In semi-supervised learning, [24] the objective is to exploit both labeled and unlabeled data, as in our approach, but the distributions of the labeled and unlabeled data are assumed to be the same. The case of transfer learning (see [25] for a comprehensive survey) is close to the challenge settings. Transfer learning aims at exploiting two source and target domains and tasks, with or without available labels for each task. It has been studied under different names (learning to learn, knowledge transfer, inductive transfer, multitask learning...) and covers a variety of settings. The closest one is called *Transductive Transfer Learning* (also named Domain Adaptation, Sample Selection Bias, Co-variate shift...), where the source and target domains are different but related, the tasks are the same, and the labels are available only in the source domain. In this setting, our approach is related to *Feature-Representation Transfer* approaches, where the objective is to find a good feature representation that reduces the difference between the source and target domains and improves the accuracy for the target task. According to Pan and Yang [25], most feature-representation transfer approaches to the transductive transfer learning settings are under unsupervised learning frameworks. In structural correspondence learning (SCL) [26], a set of domain specific *pivot* features is defined and treated as a new label vector. The corresponding classification problems are assumed to be solved by linear classifiers. SCL then learns a matrix of parameters and applies a singular value decomposition on this matrix. This allows to create new features that encodes a correspondence between the source and target domains. The tricky part is how to well design the pivot features. This can be done heuristically or using mutual information [27]. Many approaches focus on the natural language processing (NLP) domain. In [28], a kernel-mapping function is proposed to map the data from both the source and target domains. However, the kernel is domain knowledge driven and is not easy to generalize to other applications. In [29], a co-clustering based approach is proposed to propagate the labels across the domains. In [30], an algorithm named *bridged refinement* is proposed to correct the labels predicted by a shift-unaware classifier toward a target distribution and take the mixture distribution of the training and test data as a bridge to better transfer from the training data to the test data. Other approaches summarized in [25] extend traditional approaches in the NLP domain, such as spectral analysis, probabilistic latent semantic analysis (PLSA) or dimensionality reduction.

Our approach is clearly related to the settings of Transductive Transfer Learning based on Feature-Representation Transfer. One main difference is that our method relies on a multi-table data representation of the data and exploits the automatic variable construction framework summarized in Section II-D to explore many representations. Still, in the flat table case, our approach could be applied in the NLP domain where many input variables are available using the bag of words representation of texts.

Compared to related transfer learning approaches, another difference is that our approach focuses on the methodology rather than on new specific modeling techniques. The unlabeled train and test data are first exploited jointly to sort the input variables by decreasing drift informativeness. Then, a list of embedded classifiers is build on subsets of variables of increasing size, with expected increasing classification performance but also increasing sensibility to drift. The final classifier can be found either using a tolerance threshold compared to the best classification performance in the source domain, or using labels in the target domain if they are partly available.

Let us now focus on some settings that could benefit from our approach or more generally from transfer learning:

- Like in the AAIA'15 Data Mining Competition, there are many problems where the train and test data are not governed by the same distributions, and where the task is to train a classifier from many train instances coming from few sources (like the many train samples coming from only four firefighters). This is the case in domains where data is hard to collect and relies on few volunteers, each contributing to several train instances. For example, the MNIST database of handwritten digits [31] contains 60,000 train instances and 10,000 test instances, with 250 train writers and 250 other test writers. The UCI repository [8] contains many other such datasets, including for example the *Australian Sign Language signs Data Set* with 95 signs were collected from five signers for a total of 6650 sign samples, or the *Spoken Arabic Digit Data Set* with 10 digits collected from 88 speakers for a total of 8800 samples (represented using timeseries of mel-frequency cepstrum coefficients).
- When the train and test data come from different time periods, the assumption of i.i.d. distribution is violated as soon as the data are not stationary. This applies to many times series prediction problems, where samples are collected from one single source in a train period and where the trained model is applied to a future test period. For example, the IJCRS'15 Data Mining Competition[2] is related to a problem of prediction of methane outbreaks in a coal mine equipped with 28 sensors of different types (barometer, anemometer, temperature meter, humidity

[2] https://knowledgepit.fedcsis.org/contest/view.php?id=109

meter, methane meter...), with 51,700 train samples and 5,076 test samples with time periods that do not overlap with those in the train data.

- A variant of the preceding settings usually occurs in the marketing field, where there is abundant data from millions of customers, which allows extracting train datasets with one sample per customer. However, the marketing tasks are not classification, but prediction on a future period in a non-stationary market environment (for tasks such as churn, fraud or up-selling for example). In the Orange telecommunication operator, we have many such problems and plan to evaluate our transfer learning approach.

## VI. Conclusion

Whereas most data mining methods rely on i.i.d. data, this is not the case in AAIA'15 Data Mining Competition, where the train and test data where collected from two different groups of four firefighters. In this case, a robust classifier was able to achieve 100% accuracy in a 70%-30% split of the train data, with a dramatic drop of the test performance down to 18% leaderboard score. This is not an overfitting problem, but a problem of distribution drift between the train and test datasets. In this paper, we have suggested a methodology to alleviate this problem by evaluating the informativeness of each variable for the classification and drift detection tasks. We follow the intuition that the classifiers that exploit input variables with high class informativeness and low drift informativeness are more likely to be resilient to drift. Applying this methodology, we were able to build a classifier with 0.76 final score, which is a tremendous improvement compared to our initial solution. In future work, we plan to refine the methodology and to evaluate it on new problems, in particular for Orange marketing prediction tasks, where the data are abundant, complex and governed by non stationary distributions.

## References

[1] M. Meina, A. Janusz, K. Rykaczewski, D. Ślęzak, B. Celmer, and A. Krasuski, "Tagging firefighter activities at the emergency scene: Summary of AAIA'15 data mining competition at Knowledge Pit," in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., 2015,* in print September 2015.

[2] M. Boullé, "Compression-based averaging of selective naive Bayes classifiers," *Journal of Machine Learning Research,* vol. 8, pp. 1659–1685, 2007.

[3] P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifiers," in *10th National Conference on Artificial Intelligence.* AAAI Press, 1992, pp. 223–228.

[4] M. Boullé, "Towards automatic feature construction for supervised classication," in *ECML/PKDD 2014,* 2014, pp. 181–196.

[5] D. Hand and K. Yu, "Idiot's bayes? not so stupid after all?" *International Statistical Review,* vol. 69, no. 3, pp. 385–399, 2001.

[6] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Proceedings of the 12th International Conference on Machine Learning.* Morgan Kaufmann, San Francisco, CA, 1995, pp. 194–202.

[7] H. Liu, F. Hussain, C. Tan, and M. Dash, "Discretization: An enabling technique," *Data Mining and Knowledge Discovery,* vol. 4, no. 6, pp. 393–423, 2002.

[8] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[9] M. Boull′ e, "MODL: a Bayes optimal discretization method for continuous attributes," *Machine Learning,* vol. 65, no. 1, pp. 131–165, 2006.

[10] J. Rissanen, "Modeling by shortest data description," *Automatica,* vol. 14, pp. 465–471, 1978.

[11] P. Langley and S. Sage, "Induction of selective Bayesian classifiers," in *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence.* Morgan Kaufmann, 1994, pp. 399–406.

[12] R. Kohavi and G. John, "Wrappers for feature selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[13] L. Breiman, "Bagging predictors," *Machine Learning,* vol. 24, no. 2, pp. 123–140, 1996.

[14] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky, "Bayesian model averaging: A tutorial," *Statistical Science,* vol. 14, no. 4, pp. 382–417, 1999.

[15] D. Pyle, *Data preparation for data mining.* Morgan Kaufmann Publishers, Inc. San Francisco, USA, 1999.

[16] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: step-by-step data mining guide," *The CRISP-DM consortium,* Tech. Rep., 2000.

[17] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection: A Data Mining Perspective.* Kluwer Academic Publishers, 1998.

[18] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Eds., *Feature Extraction: Foundations And Applications.* Springer, 2006.

[19] A. J. Knobbe, H. Blockeel, A. Siebes, and D. Van Der Wallen, "Multi-Relational Data Mining," in *Proceedings of Benelearn '99,* 1999.

[20] S. Kramer, P. A. Flach, and N. Lavrač, "Propositionalization approaches to relational data mining," in Relational data mining, S. Džeroski and N. Lavrač, Eds. Springer-Verlag, 2001, ch. 11, pp. 262–286.

[21] M.-A. Krogel and S. Wrobel, "Transformation-based learning using multirelational aggregation," in ILP. Springer, 2001, pp. 142–155.

[22] H. Blockeel, L. De Raedt, and J. Ramon, "Top-Down Induction of Clustering Trees," in *Proceedings of the Fifteenth International Conference on Machine Learning,* 1998, pp. 55–63.

[23] A. Bondu and M. Boullé, "A supervised approach for change detection in data streams," in *Proceedings of International Joint Conference on Neural Networks,* 2011.

[24] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning.* MIT Press, Cambridge, MA, 2006.

[25] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, pp. 1345–1359, 2010.

[26] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* ser. EMNLP '06, 2006, pp. 120–128.

[27] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics,* 2007, pp. 440–447.

[28] H. Daum′ e III, "Frustratingly easy domain adaptation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics,* 2007, pp. 256–263.

[29] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Co-clustering based classification for out-of-domain documents," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ser. KDD '07. ACM, 2007, pp. 210–219.

[30] D. Xing, W. Dai, G.-R. Xue, and Y. Yu, "Bridged Refinement for Transfer Learning," in *Proc. 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD),* ser. PKDD 2007. Springer Berlin / Heidelberg, 2007, pp. 324–335.

[31] Y. LeCun and C. Cortes, "The MNIST database of handwritten digits," 1998, http://yann.lecun.com/exdb/mnist/.

DOI: 10.15439/2015F425
ACSIS, Vol. 5

# Window-Based Feature Extraction Framework for Multi-Sensor Data: A Posture Recognition Case Study

Marek Grzegorowski
Faculty of Mathematics,
Informatics and Mechanics,
University of Warsaw,
Banacha 2, 02-097 Warsaw, Poland
Email: M.Grzegorowski@mimuw.edu.pl

Sebastian Stawicki
Faculty of Mathematics,
Informatics and Mechanics,
University of Warsaw,
Banacha 2, 02-097 Warsaw, Poland
Email: Stawicki@mimuw.edu.pl

*Abstract*—**The article introduces a novel mechanism for automatic extraction of features from streams of numerical data. It was originally designed for the purpose of processing multiple streams of readings generated by sensors in coal mines. The original research was conducted on methane concentration analysis in the DISESOR project. The article demonstrates an application of the elaborated mechanism for the case of tagging short series of readings from sensors that monitor activities and movements of firefighters during the action with labels corresponding to firefighter activities. The purpose of the experiment was to assess how the automatic feature extraction and construction of classifiers (without parameters tuning and without the use of classifier ensembles) can cope with the competition's task in comparison to other participants.**

## I. Introduction

Every day, the surrounding world is being monitored by a still increasing number of sensors. Starting with sensors from our neighborhood as: mobile phones, intelligent home appliances, GPS, automotive sensors, cardio-in watches etc. ending with specialized sensors that support the manufacturing processes deployed in factories, mines or platforms. The velocity of data acquisition makes that the methods of analysis are expected to adapt rapidly to the changes and the emergence of data. On the other hand, the similarity of the nature of the data generated by the sensors appears to allow the construction of generic, reusable mechanisms for data processing and analysis.

The recent emergence of data storage technologies like columnar databases with high level of compression as Infobright [24] and the solutions that can scale up to thousands of machines like MapReduce [8] allow us to store machine generated data that is extremely large. What has to be done at this point, is to develop a generic approach to process data and to introduce a mechanism for automatic (or semi-automatic) knowledge discovery from acquired data in order to support analysts. This aims to reduce the time needed to perform the laborious, manual data analysis.

This article introduces a novel mechanism for automatic extraction of features from streams of numerical data and verifies its effectiveness based on data mining competition results. The elaborated mechanism was originally prepared for the purpose of processing multiple streams of readings generated by sensors in coal mines. The article demonstrates an application of the developed mechanism for the case of the AAIA'15 Data Mining Competition[1]: Tagging Firefighter Activities at a Fire Scene[16] which was the continuation of the previous contest investigating key risk factors for Polish Fire Service [11]. The competition was concerned the process of automatic labels (activities) assignment to a short series of readings from sensors that monitor activities and movements of firefighters during the action. The aim of the competition was to maximize balanced accuracy measure which is defined as an average accuracy within all decision classes while the aim of our research was to assess how the automatic feature extraction and classifiers learning (without parameters tuning) can cope with the competition's task.

Another of our objectives was a requirement that the total effort spent on the data preparation and experiments should be limited, which enables easier management of human resources. The overall time was limited in advance by 2MD (two man days - that is 16 h) which has been recognized as sufficient for researchers to become familiar with the task and to adjust original data representation to a format accepted by the evaluated feature extraction mechanisms. A part of the available time was used for a classifier selection and learning process and was conducted by means of the algorithms available in packages for R programming language[2].

This paper is organised as follows. In Section II the original data set and features extraction mechanisms are presented. In Section III, the assumptions of experiments, an approach to the features selection, the final solution, as well as verified (but finally discarded) approaches to data analysis are shown. In Section IV, the original application of elaborated mechanisms for the extraction of features from multiple streams within the DISESOR project is described. Finally, in Section V a

[1] https://knowledgepit.fedcsis.org/contest/view.php?id=106

[2] See. http://www.r-project.org/

978-83-60810-66-8/$25.00©2015, IEEE

summary of research, conclusions and plans for the nearest future are presented.

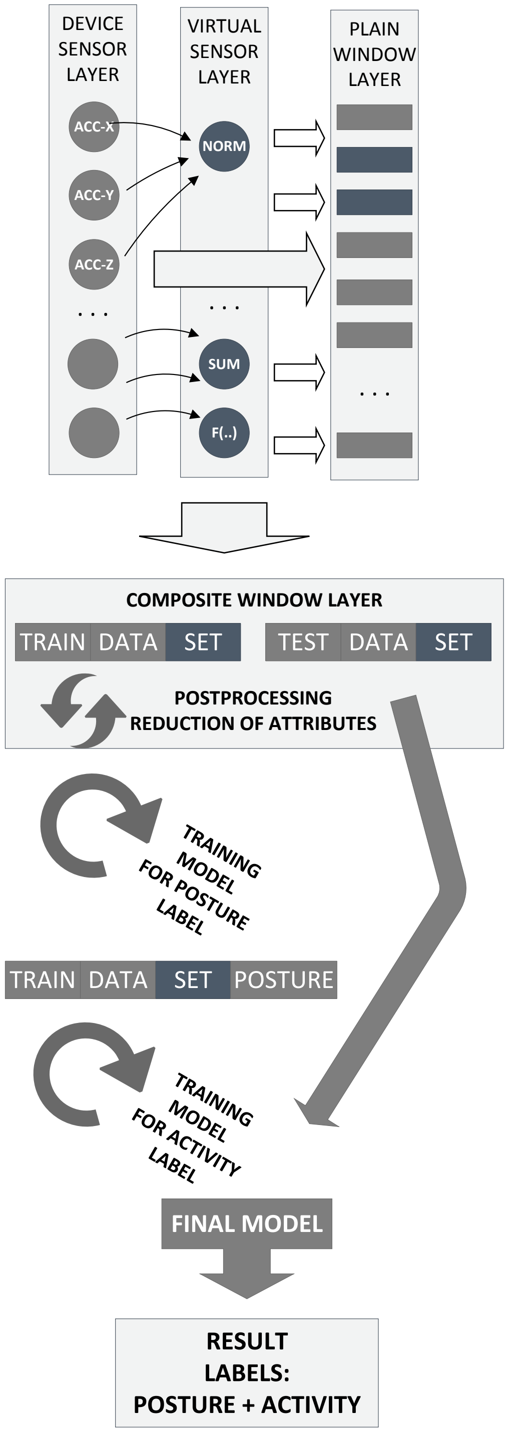


Figure 1. The diagram shows the whole process of feature extraction and model training which was carried out in order to solve the problem of labeling sensor time series with posture and main activity of a firefighter. Model responsible for recognizing a firefighter's posture uses windows constructed on the raw-sensory and virtual-sensory data. The model labeling main activity takes into account both sensory data and a posture label.

## II. Data preprocessing

### A. Original data set

The data provided in the competition were obtained during training exercises conducted by a group of eight firefighters from the Main School of Fire Service. The sensors placed on a chest were registering vital functions, while the sensors placed on torso, hands, arms and legs were registering movements of a firefighter. Along with recording the data from sensors, all training sessions were also filmed. The video recordings firstly synchronized with the sensor readings, were presented to experts who manually labeled them with actions performed during the exercises. The data were provided as CSV files.

The training and test data sets contain 20,000 rows and 17,242 columns each. A given row in a file corresponds to a short time series with length equal to approximately 1.8 s. The first 42 columns contain basic statistics (aggregations like mean, standard deviation, maximum, minimum, etc.) of data from sensors monitoring a firefighter's vital functions over the given fixed time period. The raw readings for the vital functions were recorded using Equivital Single Subject Kit (EQ-02-KIT-SU-4) fitted with two medical-quality ECG units, heart rate and breath rate units, and thermometers for measuring skin temperature. The remaining columns contain readings from a set of kinetic sensors attached to seven places on a body (torso, hands, both arms and both legs) identified as important during the realization of the main ICRA project's objectives. They are divided into 400 chunks that represent consecutive points in time. Each set is composed of readings from an accelerometer (dynamic bandwith: +/- 16G) and a gyroscope (scale up to 2,000 $deg/s$), therefore a total number of kinetic sensors are equal to 14. Each sensor of the both types (an accelerometer or a gyroscope) produces three readings $x, y, z$ corresponding to the tree dimensions, hence we have the total number of reading streams equals to 42. A single chunk of columns, therefore, consists of 43 numeric values, from which the first one is time from the beginning of the series and the following 42 values represent the readings from the accelerometers (measured in $m/s^2$) and gyroscopes (measured in $deg/s$). An average time difference between consecutive sensory readings in the data is 4.5 ms. The task is even more challenging since the training and test data sets consist of recordings from disjoint groups of firefighters.

The above description shows the details of the values arrangement in the provided data. We considered each row as a separate data set containing readings from many sensors. As described above, values from the vital sensors were aggregated externally, but the kinetic ones are provided in the raw form of time series. Let us present a fragment of an example in a visual form of data plots to better illustrate the amount of available data and their internal dependence. The references to the sensor readings are consistent with the naming from metadata provided by the organizers. There are seven places on a body that the sensors were placed on, i.e. left leg, right leg, left hand, right hand, left arm, right arm, and torso. The body areas corresponds to the following name prefixes:

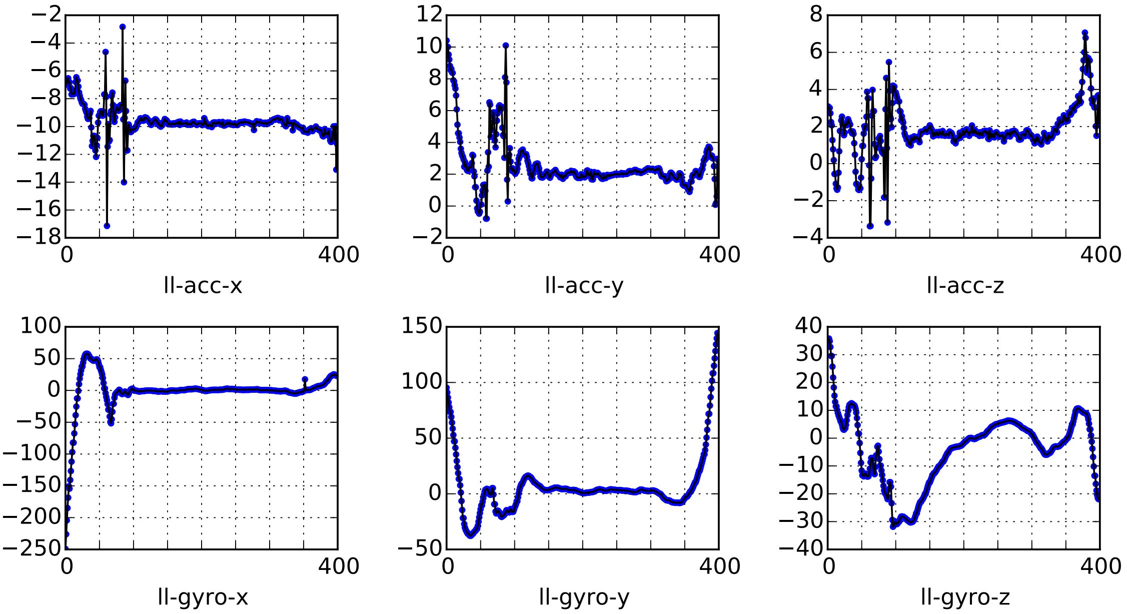


Figure 2. A fragment of an example row

$ll, rl, lh, rh, la, ra, torso$. An name infix *acc* or *gyro* refers to an accelerometer or gyroscope type of sensors. Finally, a suffix $x, y$, or $z$ names the axis from which the readings came from. In Figure 2 we present an example row that was tagged in the data with "standing" and "no_action" labels describing a posture of a firefighter and his current activity. The figure contain six time series (each consisted of 400 values that correspond to approximately 1.8 s) from the set of sensors placed on a left hand of a subject performing an exercise.

### B. Feature extraction

In the process of development of our feature extraction system we decided to follow the sliding window method. In general, for a given set of readings we put a window of a fixed *length* – the size of a window, i.e., a number of readings or a time interval, which travels through the values from the beginning to the end. We can control the amount of processed windows not only by setting the window length but also defining the *offset* for the consecutive windows – the extent to which the consecutive windows overlap to each other. Figure 3 presents four examples of sliding window set-ups. The first example, marked in red, shows the situation when the length of a sliding window is equal to the offset. The green and blue examples show the consecutive positions of a sliding window when the offset is equal to $\frac{1}{2}$ and $\frac{1}{3}$, respectively, of the length. The system is also capable to express the situation when the offset is greater than the length – the example marked in cyan.

For each basic window that is created during the process of moving a sliding window through the time series a defined aggregate function is applied. This step of the process may be adjusted for the actual task by supplying a specific implementation. The following list presents features which are calculated to represent the time series in a window:

- fill – a ratio of correct readings in the window $= \frac{nValid}{n}$,
- firstValue – a value of the first reading in the window,
- lastValue – a value of the last reading,
- max – a maximum value of the readings in the window,
- maxMinDiff – a difference between the max and min,
- mean – a mean value of readings in the window,
- min – a minimum value of the readings in the window,
- n – a total number of readings in the window,
- nValid – a number of valid readings in the window
- percentile25 – a percentile 25% for the readings,
- percentile5 – a percentile 5% for the readings,
- percentile50 – a percentile 50% for the readings (median),
- percentile75 – a percentile 75% for the readings,
- percentile95 – a percentile 95% for the readings,
- percentiles5Diff – a subtraction of the percentiles 95% and 5%,
- sourceFullId – a data source identifier included in the statistics of the window, e.g. ID or a name of the sensor,
- stdDev – a standard deviation of the readings,
- windowEndDate – a window end date
- windowEndMillis – an end timestamp of the window,
- windowMetaInfo – a meta information of the sliding window configuration, encoded in a form of a string, e.g. "o60l60" is equivalent to $offset = 60$ and $length = 60$,
- windowStartDate – a window start date,
- windowStartMillis – a start timestamp of the window,

A sliding window in a fixed position for which the aggregate function was applied and produced the statistics is referred later as a basic (or plain) window. An example of a basic window for an axis $x$ of the sensor placed on a left leg of a firefighter is presented in Table I.

For the purpose of the competition we have processed the data with three layouts of a sliding window. An illustration of our choice is presented in Figure 4. We have decided to

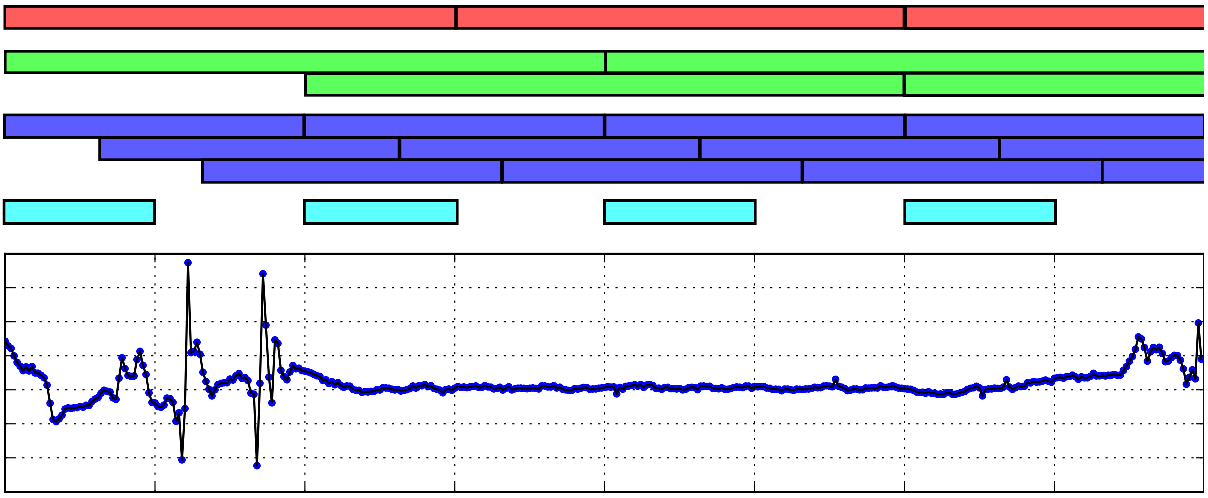


Figure 3. A set of examples showing the possible set-ups. Sliding windows are defined by a *length* and an *offset*. The length determines the size of a window, whether it is a fixed number of readings contained in a window or a fixed time interval. The offset is the extent to which the consecutive windows overlap to each other. The example marked in red shows the situation when the length of a sliding window is equal to the offset. The green and blue examples show the consecutive positions of a sliding window when the offset is equal to $\frac{1}{2}$ and $\frac{1}{3}$ of the length. The example marked in cyan illustrates the situation when the offset is twice as large as the length (or in general just greater) of a sliding window.

calculate statistics for each row by splitting each time series to 1, 2 or 5 consecutive non-overlapping windows.

We have described earlier in the section the capabilities of the feature extraction system to express different layouts of a sliding window in terms of its length and offset. If there is more than one window generated for the time series we can extract additional features in addition to those included in a basic window statistics. We have implemented also inter-window stats extraction, i.e., a set of values that express the changes between a pair of consecutive windows. We have introduced the following inter-window stats:

- firstFill – a ratio of correct readings in the first window,
- firstN – a total number of readings in the first window,
- firstNValid – a number of valid readings in the first window,
- firstWindowDate – a start date in the first window,
- firstWindowMillis – a start timestamp in the first window,
- maxDiff – a difference between *max* statistics in the windows,
- meanDiff – a difference between *mean* statistics in the windows,
- minDiff – a difference between *min* statistics in the windows,
- percentile25Diff – a difference between *percentile25* statistics in the windows,
- percentile5Diff – a difference between *percentile5* statistics in the windows,
- percentile50Diff – a difference between *percentile50* statistics in the windows,
- percentile75Diff – a difference between *percentile75* statistics in the windows,
- percentile95Diff – a difference between *percentile95* statistics in the windows,
- secondFill – a ratio of correct readings in the first window,
- secondN – a total number of readings in the second window,
- secondNValid – a no. of valid readings in the second window,
- secondWindowDate – a start date in the second window,
- secondWindowMillis – a start timestamp in the second window,
- sourceFullId – a data source identifier,
- windowMetaInfo – a meta information of the sliding window,

An example of the inter-window stats for an axis $x$ of the sensor placed on a left leg of a firefighter is presented in Table II. A sliding window configuration used in the example, i.e., the length of the window is equal to its offset, has produced two basic non-overlapping windows that split the time series from a given row into two halves.

| | stat | value |
|---|---|---|
| 1 | fill | 1 |
| 2 | firstValue | -7 |
| 3 | lastValue | -11.2 |
| 4 | max | -2.8 |
| 5 | maxMinDiff | 14.3 |
| 6 | mean | -9.6 |
| 7 | min | -17.1 |
| 8 | n | 400 |
| 9 | nValid | 400 |
| 10 | percentile25 | -9.9 |
| 11 | percentile5 | -10.8 |
| 12 | percentile50 | -9.8 |
| 13 | percentile75 | -9.5 |
| 14 | percentile95 | -7.6 |
| 15 | percentiles5Diff | 3.2 |
| 16 | sourceFullId | ll-acc-x |
| 17 | stdDev | 1.1 |
| 18 | windowEndDate | 2015-05-03 00:06:40 |
| 19 | windowEndMillis | 1430604400000 |
| 20 | windowMetaInfo | o4001400 |
| 21 | windowStartDate | 2015-05-03 00:00:00 |
| 22 | windowStartMillis | 1430604000000 |

Table I
AN EXAMPLE OF AGGREGATION FUNCTION COMPUTATION – A BASIC WINDOW STATS FOR THE FIRST ROW OF THE TRAINING DATA.

### C. Virtual sensors

According to the task description, the kinetic sensors (accelerometers and gyroscopes) used during the exercises have symmetric scales with 0 as their neutral reading. The specificity of the firefighter activities like walking, running, moving up the stairs or ladder, may cause the readings to be more

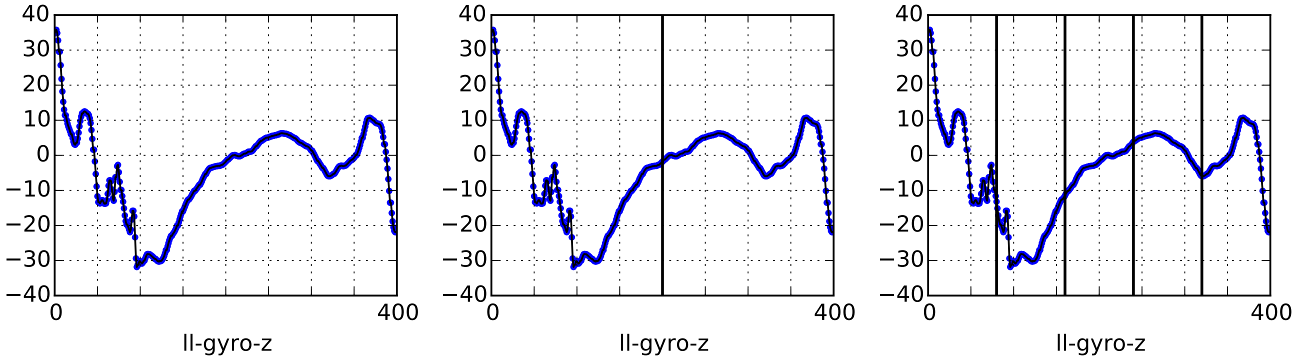


Figure 4. An illustration of the sliding window configurations applied in our solution. We have decided to process the time series with a varied granularity, ranging from the statistics computed for the whole time series, to calculate them for 2 or 5 shorter, non-overlapping windows which divided the time series to the parts of equal length.

| | stat | value |
|---|---|---|
| 1 | firstFill | 1 |
| 2 | firstN | 200 |
| 3 | firstNValid | 200 |
| 4 | firstWindowDate | 2015-05-03 00:00:00 |
| 5 | firstWindowMillis | 1430604000000 |
| 6 | maxDiff | 6.6 |
| 7 | meanDiff | 0.6 |
| 8 | minDiff | -4 |
| 9 | percentile25Diff | 0.3 |
| 10 | percentile50Diff | 0.2 |
| 11 | percentile5Diff | -0.4 |
| 12 | percentile75Diff | 0.8 |
| 13 | percentile95Diff | 2.7 |
| 14 | secondFill | 1 |
| 15 | secondN | 200 |
| 16 | secondNValid | 200 |
| 17 | secondWindowDate | 2015-05-03 00:03:20 |
| 18 | secondWindowMillis | 1430604200000 |
| 19 | sourceFullId | ll-acc-x |
| 20 | windowMetaInfo | o2001200 |

Table II
AN EXAMPLE OF INTER-WINDOW STATISTICS

significant when considered as a group, e.g. a whole tuple $(x,y,z)$ from a given sensor rather than separate readings $x$, $y$, and $z$, to express the intensity of the movement. We have decided to introduce a concept of *virtual sensors*. Besides applying the aggregate functions to the original time series available in the delivered files, we have implemented an idea of creating artificial time series derived from the original ones. The virtual sensors are created on the basis of one or more time series from other sensors (whether original or virtual) after applying a particular function. In our solution, we decided to create virtual sensors for readings from all accelerometers and gyroscopes' axes separately, applying an *abs* (absolute value) function. We created also virtual sensors for readings grouped in tuples $(x,y,z)$ for each kinetic sensor – computing the Manhattan and Euclidean norms for the $(x,y,z)$ vectors. An example that illustrates the concept of virtual sensors that we have used in our solution can be seen in Figure 5.

After all basic windows for original and virtual sensors that comes from a given data row are calculated, they are joined (in the sense of appending all their values) together, forming a row of data that will serve as an input for further steps of data analysis and experiments.

## III. EXPERIMENTS

### A. Evaluation

The submitted solutions were evaluated using the balanced accuracy measure which is defined as an average accuracy within all decision classes. It was computed separately for the labels describing the posture and main activities of firefighters. The final score is a weighted average of balanced accuracies computed for those two sets of labels and is defined as follows:

$$score(s) = \frac{BAC_p(s) + 2 \cdot BAC_a(s)}{3}.$$

Where $BAC_p$ is the balanced accuracy for labels describing the posture and $BAC_a$ for the main activity. Precise definition of balanced accuracy is as follows:

$$BAC(preds, labels) = \frac{\sum_{1<i<l} ACC_i(preds, labels)}{l}$$

$$ACC_i(preds, labels) = \frac{|\; j : preds_j = labels_j = i \;|}{|\; j : labels_j = i \;|}.$$

### B. Constraints

We considered the competition as a good opportunity to verify the developed mechanisms of automation of knowledge discovery process and their usefulness in the production environment. Therefore, working on the solution we have imposed a few additional constraints and requirements. All have been set arbitrarily for the issue of labeling firefighters activity. We consider them to be satisfactory for the task:

1) Overall working time, to be spent on solving of the problem by all members of the research team must not exceed a total of 2MD.

2) The overall time required to train the classifiers must not exceed the total of 10 minutes. In case of classifiers which are mutually independent this is 10 minutes for training each of them, since the process can be run in parallel.
3) The time required to pre-process a single row of data to a format accepted by classifier and assignment of both labels must not exceed one second.

The first of the imposed restrictions is intended to help to verify whether it is possibile to immediately familiarize analysts with both data and the problems. In the simulated case, two analysts were working on adaptation of the data provided in the new format to the already existing mechanisms. Possibility to adapt quickly to new data and to new expectations while maintaining a satisfactory accuracy of the model is very important especially in the threats monitoring.

The second point poses a constraint on the time that is necessary to re-train the model on the new data, in case after a certain time the quality of the assessment has fallen below the a predetermined score level due to, e.g. concept shift/drift [6]. We assumed that the time required to re-training the model should not exceed 10 minutes. Nevertheless, we consider this point to be the least important and in our opinion exceeding proposed limit should not disqualify the approach. However, in the final embodiment, the total time of training classifiers did not exceed 7 minutes, wherein the classifiers are independent and can be trained simultaneously.

The last point we consider to be the most important because it imposes limits on the permissible delay in operation of preprocessor and classifiers when acting in a production environment. According to the assumptions maximum delay between data collection and complete processing and labeling of single row of data should not exceed one second. This is one of the main reasons for excluding from consideration all object based methods as well as heavy classifier ensembles. Generation of all the features, including those for both: raw and virtual-sensors readings, took approximately 450 milliseconds per a single csv file row. The postprocessing and assignment of the labels has been performed in the R - software environment for statistical computing and consisted of: importing data (overall 30 seconds per 20000 rows of test data set), feature selection (overall 10 seconds per 20000 rows of test set) and labeling (classification with SVM took overall of 70 seconds for both labels for 20000 rows).

### C. Post processing

Generated data sets have the following quantity of attributes for each of 20000 objects, depending on configuration:

- 2199 – one sliding window per short time series
- 6315 – two sliding windows per short time series
- 18663 – five sliding windows per short time series

Making a total of 27177 attributes [27] from the conditional- and inter-sliding windows constructed for both raw and virtual sensors. Elements of the automatic feature selection and reducing the number of attributes are in the study phase and still have not been introduced to the data processing mechanisms. Hence, the feature selection was carried out manually.

In a first step, all features exhibiting signs of identifiers and all constants values, that is: fill, n, nValid, sourceFullId, windowEndDate, windowEndMillis, windowMetaInfo, windowStartDate, windowStartMillis have been removed from the prepared data set. We have also removed maximum and minimum of values in windows to limit the influence of outliers on the final result. After applying the model on acquired attributes of training data set we have noticed that the model has been extremely overfitted. As the main reason for this, we find the fact that the training data set was prepared based on the observation of a small number of firefighters, hence data could not contain all possible patterns of motor behavior and vital signs. This observation led us to change our approach and forced to look for features that maintain a quality of prediction for test set.

In the process of feature selection we used a wrapper approach[13]. We have been progressively enlarging the number of utilized features and making periodic evaluations, after each step we have either remained the selected features or we resigned from them, depending on the result of the evaluation. Ultimately, the SVM was run on the 163 attributes for the classifier that labels the objects with posture and with one additional attribute (the computed posture label) for the second SVM model which classifies the data with a main activity. Beyond selected features in sliding windows, described in Section II final set of objects includes additional attributes to exclude the symmetry of right- and left-handed people e.g. the sum of the selected features for the left and right hand as well as sum for the left and right leg. This is very important since the training samples were created basing on the behavior of different people than the test samples. Moreover, training and test set data were acquired during observation of small group of firefighters, hence the training sample could not contain all possible patterns. The situation when training set differs significantly from the test set forced us to make additional step during verification of selected attributes.

### D. Classifier training and labeling

Because of the pre-processing of data that has been provided in the competition, the real problem of monitoring the firefighters activities, which is originally associated with processing of streams of sensor readings [10] that constituting time series [2], has been reduced for the problem of classification [25], more precisely to multi-labeling [26]. Original sensor readings has been pre-processed and subdivided into frames [21], [28] of given length and made available in a csv file. To apply the developed feature extraction mechanism each row of the csv file has been split into short time series of readings from sensors respectively to csv header names: "ll-acc-x", "ll-acc-y", "ll-acc-z", etc. and passed as an input stream to the feature extraction mechanism. Eventually, we obtained a set of elaborated features ready for multi-labeling [14].

During data analysis, not only the conditional variables have been inspected but also posture and activity labels. The preliminary conclusions of labels aggregation allowed to state that there is a huge imbalance [29] in classes defined by

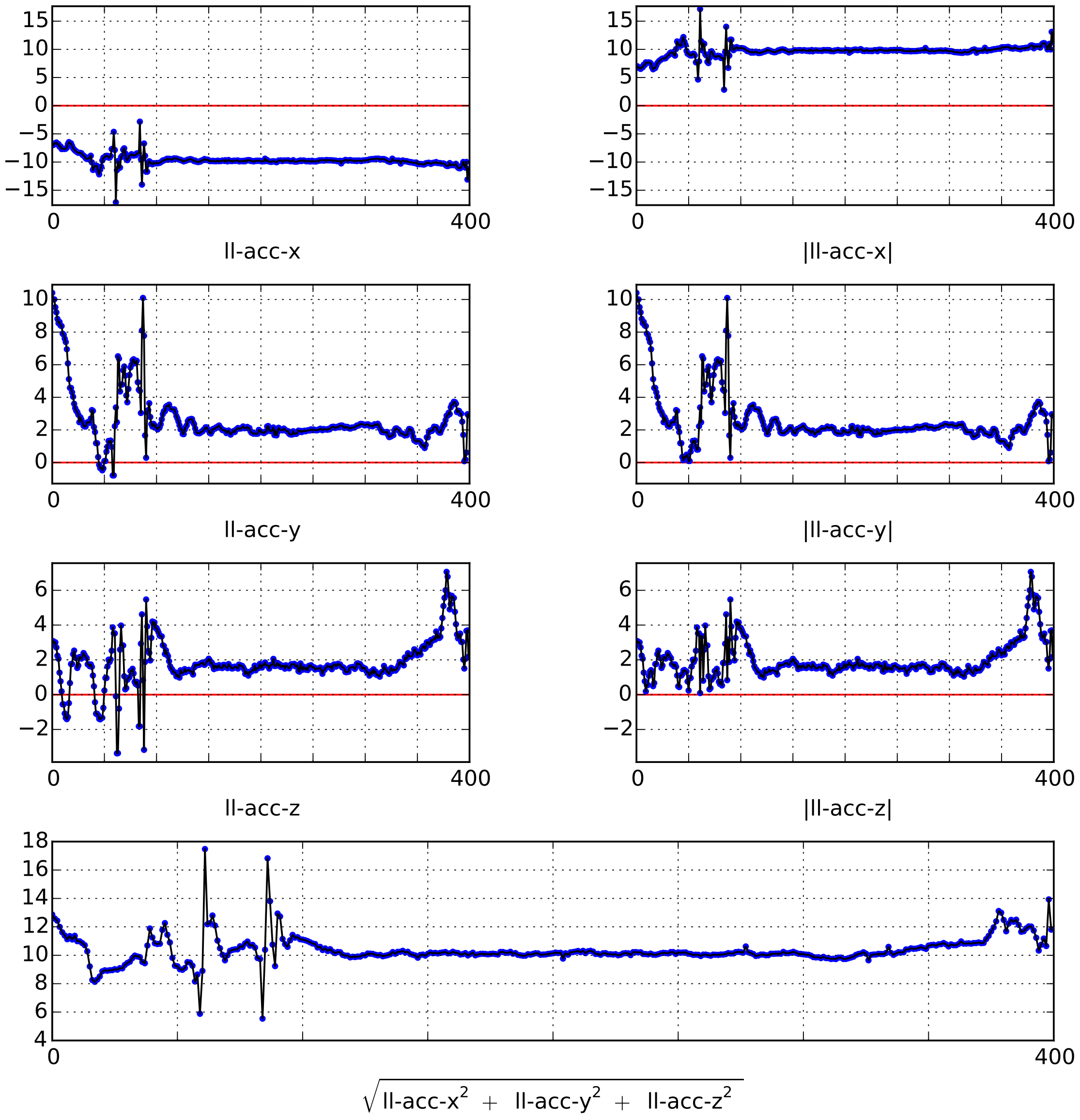


Figure 5. An example of deriving virtual sensors by applying an absolute value function and the Euclidean norm to the original time series.

particular labels and that the labels for firefighters posture and activity are not independent [20] and there is a connection between them [19]. The application of label power-set methods [23], [30] did not provide satisfactory results but classifier chains[22] improved the achieved score significantly. The way in which assessment of solutions was defined, that is uneven importance of labels for posture and activity encouraged to consider various concepts like a multilabel classification with label ranking [9] or a graded multilabel classification [5].

Experiments have been implemented and carried out in the R software environment. We have experimented with the following classification algorithms: rPart (decision trees [4]), rFerns (random forests [3]) and e1071 (support vector machines [1]). The final solution is based on SVM. While learning classifiers we have used the relationship between labels by training two SVM models on slightly different data.

Model 1, which recognizes posture, is SVM with 4356 support vectors. Model has been trained on the basis of the features described above with the default parameters, that is:

- SVM-Type: C-classification
- SVM-Kernel: radial
- Cost: 1
- Gamma: 0.006134969

$$model1 \leftarrow svm(posture \sim .,$$
$$data = trainSet[,c(selectedFeatures, posture)]);$$

Model 2, which recognizes the main activity, is SVM with 5011 support vectors. Model has been trained on data enriched by the posture label with the default parameters, that is:

- SVM-Type: C-classification
- SVM-Kernel: radial
- Cost: 1
- Gamma: 0.005952381

$$model2 \leftarrow svm(activity \sim .,$$
$$data = trainSet[,c(selectedFeatures, posture, activity)])$$

During labeling, data were firstly described with the posture label, and after that with the main activity label:

$$testLabelsForPosture \leftarrow predict(model1,$$
$$newdata = testSet[,selectedFeatures],$$
$$type = "class");$$
$$testSet\$posture \leftarrow testLabelsForPosture;$$
$$testLabelsForActivity \leftarrow predict(model2,$$
$$newdata = testSet[,c(selectedFeatures, posture)],$$
$$type = "class");$$

## IV. DISESOR

The most significant application of the presented solution for automated feature extraction is the ongoing DISESOR project. DISESOR aims to build a decision support system for threats monitoring and early warnings in coal mines.

Nowadays, the coal mining is playing a crucial role on Polish energy market and is employing hundreds of thousands of people. Coal mines are well equipped with monitoring, supervising and dispatch systems connected with machinery, devices and transport facilities. There are a lot of systems that support essentially different aspects of the mine operation, e.g.: ARES, ARAMIS, HESTIA for seismo-acoustic monitoring; RODOS, ALFA for quality control, MAKS, Ergon, Hades for machinery monitoring; SMP, STAR, CTT, UTS, Venturon, Univers for risk control, ZEFIR, THOR, sD2000 - central systems and many, many others. Each of these gathers readings from specific sensors placed in mines, depending on their domain: methane sensor, $CO$ and $CO_2$ sensor, seismic sensor, shearer state sensors etc. Assembly of a variety of data from multiple systems enables performing a wide-ranging analysis.

Monitoring systems are developed by many providers what causes problems with integration and proper interpretation of the data, therefore there is need to deploy a decision support system integrating different aspects of coal mine operations, what is the main task of the DISESOR system. The high level design of DISESOR takes into account the data cleaning process, the process of building data mining models and online predictive reasoning for the latest data readings. The most important use cases of the DISESOR system are:

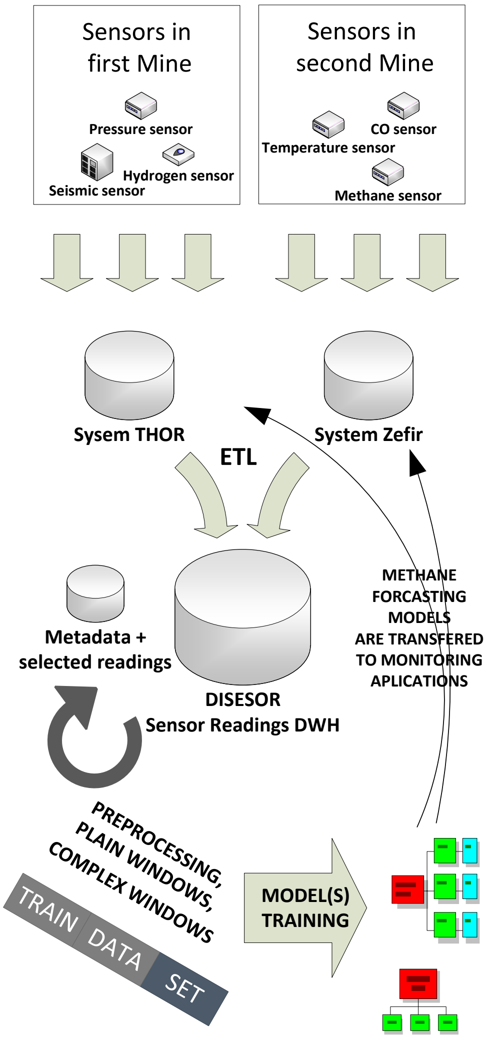


Figure 6. DISESOR ETL process collects sensor readings from mine monitoring systems like THOR or ZEFIR. Raw data is cleand and after preprocesing predictive models are generated.

- The assessment of seismic hazard probabilities in the vicinity of the mine.
- Forecasting dangerous increase of the methane concentration in the mine shafts.
- Detection of endogenous fires and conveyor belts fires.
- Detecting anomalies in the consumption of media.
- Diagnostics of machines: roadheaders and shearers.

## V. Conclusions and Further Research

The developed feature extraction system can be configured to accept a data set consisted of readings from multiple sensors. The algorithm that builds sliding windows divides reading streams into consecutive fragments and then processes each of them separately. This approach allows for effective parallelization of the whole feature extraction process. However, there are some important issues that have not been addressed in a prepared solution or have been taken into account in a very simplified manner, e.g. a quantization of real value attributes [17], [18] or an attribute selection [7], [12] which we recognize as very important elements of a knowledge discovery [15] process. We are going to extend the discussed mechanisms with modules covering those issues in the nearest future.

The conducted experiments showed that the features prepared by the elaborated mechanism are suitable for machine learning algorithms, which in the next step can give very promising results without neither long lasting manual data cleaning nor classifier tuning. The results of experiments turned out to be significantly better than the baseline solution. Therefore, it seems that the elaborated system is prepared to work in production. However, there is still a lot of space for further improvements since results achieved by other participants in case of manual transformation of data and tuning of classifiers turned out to be even better.

## VI. Acknowledgements

This research was partially supported by Polish National Science Centre grant DEC-2012/05/B/ST6/03215 and by Polish National Centre for Research and Development grants PBS2/B9/20/2013 and O ROB/0010/03/001.

## References

[1] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

[2] G. E. P. Box and G. Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[5] W. Cheng, K. Dembczynski, and E. Hüllermeier. Graded multilabel classification: The ordinal case. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning, June 21-24, 2010, Haifa, Israel*, pages 223–230. Omnipress, 2010.

[6] J. Coble and D. J. Cook. Real-time learning when concepts shift. In J. N. Etheredge and B. Z. Manaris, editors, *FLAIRS Conference*, pages 192–196. AAAI Press, 2000.

[7] C. Cornelis, R. Jensen, G. H. Martín, and D. Ślęzak. Attribute selection with fuzzy decision reducts. *Inf. Sci.*, 180(2):209–224, 2010.

[8] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008.

[9] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Mach. Learn.*, 73(2):133–153, Nov. 2008.

[10] M. Grzegorowski. Scaling of complex calculations over big data-sets. In D. Ślęzak, G. Schaefer, S. T. Vuong, and Y. Kim, editors, *Active Media Technology - 10th International Conference, AMT 2014, Warsaw, Poland, August 11-14, 2014. Proceedings*, volume 8610 of *Lecture Notes in Computer Science*, pages 73–84. Springer, 2014.

[11] A. Janusz, A. Krasuski, S. Stawicki, M. Rosiak, D. Ślęzak, and H. S. Nguyen. Key risk factors for polish state fire service: a data mining competition at knowledge pit. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, September 7-10, 2014.*, pages 345–354, 2014.

[12] A. Janusz and D. Ślęzak. Rough set methods for attribute clustering and selection. *Appl. Artif. Intell.*, 28(3):220–242, Mar. 2014.

[13] A. Janusz and S. Stawicki. Applications of approximate reducts to the feature selection problem. In *Rough Sets and Knowledge Technology - 6th International Conference, RSKT 2011, Banff, Canada, October 9-12, 2011. Proceedings*, pages 45–50, 2011.

[14] W. Jiang, Z. W. Ras, and A. Wieczorkowska. Clustering driven cascade classifiers for multi-indexing of polyphonic music by instruments. In Z. W. Ras and A. Wieczorkowska, editors, *Advances in Music Information Retrieval*, volume 274 of *Studies in Computational Intelligence*, pages 19–38. Springer, 2010.

[15] K. Kreński, A. Krasuski, M. Szczuka, and S. Łazowy. Granular knowledge discovery framework for fire and rescue reporting system. *Intelligent Decision Technologies*, pages 1–12, 2014.

[16] M. Meina, A. Janusz, K. Rykaczewski, D. Ślęzak, B. Celmer, and A. Krasuski. Tagging firefighter activities at the emergency scene: Summary of aaia'15 data mining competition at Knowledge Pit. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, 2015. In print September 2015.

[17] H. S. Nguyen. On efficient handling of continuous attributes in large data bases. *Fundam. Inf.*, 48(1):61–81, Oct. 2001.

[18] H. S. Nguyen. On exploring soft discretization of continuous attributes. In S. K. Pal, L. Polkowski, and A. Skowron, editors, *Rough-Neural Computing*, Cognitive Technologies, pages 333–350. Springer Berlin Heidelberg, 2004.

[19] S.-H. Park and J. Fürnkranz. Multi-label classification with contraints. In *Proceedings of the workshop on Preference Learning at ECML PKDD'08*, Antwerp, Belgium, 2008.

[20] S.-H. Park and J. Fürnkranz. Multi-Label Classification with Label Constraints. Technical report, Knowledge Engineering Group, TU Darmstadt, 2008.

[21] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Trans. Knowl. Discov. Data*, 7(3):10:1–10:31, Sept. 2013.

[22] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Mach. Learn.*, 85(3):333–359, Dec. 2011.

[23] J. Read, A. Puurula, and A. Bifet. Multi-label classification with meta-labels. In R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, and X. Wu, editors, *2014 IEEE International Conference on Data Mining, Shenzhen, China, December 14-17, 2014*, pages 941–946. IEEE, 2014.

[24] D. Ślęzak and V. Eastwood. Data warehouse technology by infobright. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 841–846, New York, NY, USA, 2009. ACM.

[25] D. Ślęzak and A. Janusz. Ensembles of bireducts: Towards robust classification and simple representation. In T. Kim, H. Adeli, D. Ślęzak, F. E. Sandnes, X. Song, K. Chung, and K. P. Arnett, editors, *Future Generation Information Technology - Third International Conference, FGIT 2011 in Conjunction with GDC 2011, Jeju Island, Korea, December 8-10, 2011. Proceedings*, volume 7105 of *Lecture Notes in Computer Science*, pages 64–77. Springer, 2011.

[26] D. Ślęzak, A. Janusz, W. Świeboda, H. S. Nguyen, J. G. Bazan, and A. Skowron. Semantic analytics of pubmed content. In *Information Quality in e-Health - 7th Conference of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2011, Graz, Austria, November 25-26, 2011. Proceedings*, pages 63–74, 2011.

[27] M. S. Szczuka and D. Ślęzak. How deep data becomes big data. In *Joint IFSA World Congress and NAFIPS Annual Meeting, IFSA/NAFIPS, Edmonton, Alberta, Canada, June 24-28, 2013*, pages 579–584, 2013.

[28] A. Wieczorkowska, J. Wróblewski, D. Ślęzak, and P. Synak. Problems with automatic classification of musical sounds. In *Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'03 Conference held in Zakopane, Poland, June 2-5, 2003*, pages 423–430, 2003.

[29] E. S. Xioufis, M. Spiliopoulou, G. Tsoumakas, and I. Vlahavas. Dealing with concept drift and class imbalance in multi-label stream classification. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1583–1588. AAAI Press, 2011.

[30] Y. Yang and S. Gopal. Multilabel classification with meta-level features in a learning-to-rank framework. *Machine Learning*, 88(1-2):47–68, 2012.

DOI: 10.15439/2015F419
ACSIS, Vol. 5

# A Versatile Approach to Classification of Multivariate Time Series Data

Adam Zagorecki
Centre for Simulation and Analytics
Cranfield University
Defence Academy of the United Kingdom
Shrivenham, SN6 8LA, United Kingdom
Email: a.zagorecki@cranfield.ac.uk

***Abstract*—During the recent decade we have experienced a rise of popularity of sensors capable of collecting large amounts of data. One of most popular types of data collected by sensors is time series composed of sequences of measurements taken over time. With low cost of individual sensors, multivariate time series data sets are becoming common. Examples can include vehicle or machinery monitoring, sensors from smartphones or sensor suites installed on a human body. This paper describes a generic method that can be applied to arbitrary set of multivariate time series data in order to perform classification or regression tasks. This method was applied to the 2015 AAIA Data Mining Competition concerned with classifying firefighter activities and consecutively led to achieving the second-high score of nearly 80 participant teams.**

## I. Introduction

IN THIS paper I present a generic approach to classification of multivariate time series data. This approach was developed and evaluated in the context of the 2015 AAIA Data Mining Competition, where it led to the second highest score of nearly 80 solutions.

During the recent decade we have experienced a rise of popularity of sensors capable of collecting large amounts of data. One of most popular types of data collected by sensors is time series composed of sequences of measurements taken over time. With low cost of individual sensors, multivariate time series data sets are becoming common. Examples can include vehicle or machinery monitoring, sensors from smartphones or sensor suites installed on human body. The collected measurement data is typically not directly useful to the users, as it consists of typically a large number of data points and is very noisy. It should be processed and transformed into knowledge that can be useful to the user. Because of the sheer volume of the data and typically non-trivial patterns present in data this task is suitable for data-mining approaches. In fact in recent years we observe a significant increase of applications that rely on data mining to interpret sensor data and provide useful and actionable knowledge to the users. One of such areas is human body monitoring that can be valuable for healthcare applications, such as post-surgery patient monitoring, monitoring patients with chronic diseases and general well-being promotion, among others.

In this paper the time series data was generated by a sensor suite worn by firefighters during training sessions. The main focus will be on time series generated by a set of accelerometers and gyroscopes installed on different parts of human body. The data generated by sensors will be used to fully automatically identify activities performed by a subject such as running, climbing a ladder, etc.

The rest of the paper is composed as follows: in the next section the competition task will be introduced with details of the sensors, available data and the evaluation. In the following section I will discuss the proposed approach to classification of multivariate time series data. Consequently each step in of the proposed approach will be discussed in more detail: feature engineering, feature selection, and actual classification. I will finish the paper with a short discussion.

## II. The Competition Task

This paper describes a solution to the AAIA‘15 data mining competition [1] was organized using the Knowledge Pit competition platform [2]. The objective of the competition was to develop efficient methods for automatic labeling of short series of the sensory data in the context of firefighter training activities.

The basic task of the competition was to create a data mining model to predict training activities performed by a firefighter based on data collected from sensor readings installed on the firefighter body. For this purpose a commercial off-the-shelf body sensor suite was used to generate the data.

### A. Data

The data for the competition was generated using *smart jacket* – a wearable set of body sensors for monitoring kinematics and psycho-physical condition of firefighters. For each record the data was divided into two subsets.

The first subset consisted of 42 columns that represented aggregations of data from sensors monitoring firefighter‘s vital functions. Examples of measurements taken are ECG, heart rate, respiration rate, skin temperature, etc. The data for those measurements was pre-processed by the organizers and made available in the form of statistics (mean, standard deviation, skewness, etc.) rather than time series.

The second subset of data consisted of a set of 42 time series, each consisting of 400 data points. The time series was generated by a set of accelerometers and gyroscopes.

978-83-60810-66-8/$25.00©2015, IEEE

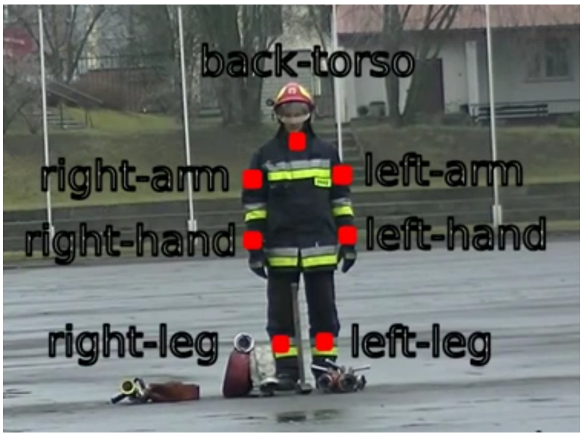


Fig. 1. Distribution of sensors on a firefighter body.

There were 7 pairs of accelerometer-gyroscope installed on firefighter‘s body. The locations of the sensors are shown in Figure 1. Each pair of sensors generated 6 data streams ($x$, $y$, and $z$ axes for an accelerometer and $x$, $y$, and $z$ axes for a gyroscope). The 400 points corresponded to approximately 1.8 second period of continuous measurement. Since the measurements in the time series were not taken in equal intervals (but all of them were taken at the same time for a given set of 42 time series), the the organizers provided a set of 400 time stamps that corresponded to time of measurements.

In total a data record consisted of 17,242 columns, all of them were real numbers.

There were two class attributes associated with data record. They related to activities during the firefighter's training. The first attribute was the body posture that had 5 states: *standing*, *stooping*, *moving*, *crawling* and *crouching*. The second attribute related to the main activity with 16 different activities, such as *no action*, *walking*, *running*, *searching*, *stairs up*, *manipulating*, *throwing hose*, etc.

The data sets consisted of 20,000 training records and 20,000 test cases that were collected during firefighter training. Multiple firefighters participated in data collection.

### B. Evaluation

The evaluation of the model performance was determined using the score $s$ which was defined in the following manner:

$$s(p, y) = \frac{1}{3}\big(BAC_p(p, y) + 2 * BAC_a(p, y)\big),$$

where $BAC_p(p, y)$ and $BAC_a(p, y)$ are *balanced accuracies* for posture and main activity respectively, determined for a set of predictions $p$ given true labels $y$. The balanced accuracy $BAC$ is defined as as the average of accuracies for individual labels. Let $l$ be the number of all possible labels, then the balanced accuracy is defined as follows:

$$BAC(p, y) = \frac{1}{l}\sum_{i=1}^{l} ACC_i(p, y),$$

where $ACC$ is accuracy for a given label $i$ and it is defined as:

$$ACC_i(p, y) = \frac{|j : p_j = y_j = i|}{|j : y_j = i|}.$$

The goal was to propose a model that would generate a set of predictions $p$ for the cases for which are known true labels $y$ in order to maximize the score $s$. The true labels were known only to the organizers, but not to the competitors. The competition platform was used to present provisional evaluation results based on the subset of the actual evaluation set. The final evaluation was made on the remaining test data set.

## III. Solution Overview

In this section I present overview of the solution to the competition task that I developed. The basic steps are presented in Figure 2.

The first, and probably the most critical step was the feature engineering step. At this step the original data set was converted to a secondary data set that consisted of the features generated from the time series data. This step is discussed in detail in the Section IV. It is important to note, that I decided to reject the features related to firefighter's vital measurements and I completely relied on data generated by kinetic sensors. That meant that the data used consisted entirely of a set of 42 time series, each of the same length and all of coupled. The next decision was to ignore time stamp data and to assume that measurements were taken in equal intervals. This decision was dictated entirely by desire to simplify the data processing.

There was another important decision related to data preprocessing I made: I decided to collapse two class attributes into one. Initially, I approached the two class classification problem as two independent classification problems – building two two models one for body posture and the other for activity, with no information shared between the two models. However, I noticed that even though theoretically there were $5 \cdot 16 = 80$ possible states of combined class attributes, in practice only 24 were present in the training data set, which was only slightly higher than the number of states for the second class attribute. Using one class attribute led to dramatic increase of classification performance.

As the competition progressed and the number of features increased, it has become clear that feature selection step would provide benefit. Toward the end of competition a typical features data set would consist of 4,000 to 8,000 attributes. By experimentation it has become clear that reducing the number of attributes to the number between 200 and 600 would clearly improve classification performance. I used a feature selection algorithm to reduce the number of features. It turned out that selecting different number of features from the same feature data set can have quite profound effect on the classification performance. Feature selection led to generation of the reduced feature data set that was used for actual classification task.

As the basic classifier I used combination of the Random Forest classifier with the Multi-Class classifier that converted multiple class problem into set of forests each corresponding

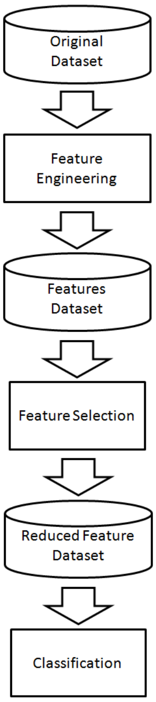


Fig. 2. The outline of the basic tasks used during the competition.

to a set of binary problem Random Forests. I experimented with other classification algorithms available in Weka such as Neural Networks, Logistic Regression, Naive Bayes, Decision Trees, Support Vector Machines, but all other algorithms seemed to perform significantly worse or were taking too long to finish. I did not attempt to compensate for imbalanced distribution of attribute classes.

For feature selection and classification I used Weka software [3]. The feature engineering step was performed using my own code written especially for the purpose of the competition.

For the sake of competition I decided to ignore checking my results for over-fitting. This decision was made strictly for pragmatic reasons – initial attempts to cross-validation did not seem to be representative to the results obtained on the leaderboard. The models that were achieving 100% accuracy on the test set seemed to perform better on the leaderboard than those with lower accuracies on the test set. Obviously, getting the right prediction of the accuracy error would likely lead to improved results, but because of the limited time I wanted to spend on the competition and the fact that it was possible to test results using the submission system I did not focus on getting proper handling of over-fitting.

## IV. Feature Engineering

The first step in data pre-processing was transformation time series data into a set of numerical values that would summarize different aspects of the time series data. This step is commonly referred as *feature engineering*. The features can be derived from individual time series (e.g. mean, standard deviation) or from some form of a function that can take more than one data series (such example can be a correlation coefficient between two time series).

### A. Original signals

For the feature engineering I used the original 42 time series (generated by accelerometers and gyroscopes). I ignored the time stamps provided and assumed that the measurements are taken in equal intervals.

### B. Derived Signals

For the feature generation I decided to use additional time series data that were derived from the original time series. In particular, I combined $x$, $y$, and $z$ coordinates using Eucleadian norm for each of the accelerometers and gyroscopes, which led to additional 14 derived time series.

### C. Extracted Features

For each of the time series (either original or derived) the following features were extracted:

- the mean value
- the maximal value
- the minimal value
- the range (difference between the maximal and minimal values)
- the sum of squared values (mean power)
- the logarithm of the sum of squared values (log mean power)
- the standard deviation
- skewness
- kurtosis
- the 5th central moment
- the maximal difference between two consecutive measurements
- autocorrelation taken at $t$=1,2,5,20, and 50
- power for the bin with the maximal value (power) Fast Fourier Transform (excluding the zeroth frequency)
- maximal value of frequency (in the form of an index) for the bin with maximal value for Fast Transform (excluding the zeroth frequency)
- slope and intercept for the linear regression
- mean square error for the linear regression
- parameters for polynomial fitting with $n = 2$ $(a_0, a_1, a_2)$

Each of the above features generated a single number that was used as an individual feature for further analysis. This produced 1400 features.

### D. Correlations

Finally, I decided to add correlation coefficients between time series. I did it for the original and derived signals separately, that led to 861 features and 91 features, respectively.

## V. Feature Selection

The feature selection has quickly become a necessity as the number of features in the feature set increased. I tested various feature selection algorithms available in Weka. The best results were achieved with the `CfsSubsetEval` algorithm.

The algorithm determines the worth of attribute' subsets by considering the individual predictive ability of each attribute along with the degree of redundancy between attributes in the subset. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred. The Weka's default best first search method was used with default parameters. I used 10-cross validation.

One of the challenges was to decide on the actual number of features to be used. For the winning solution, the feature data set had 2352 features. The 10 runs of cross-validation for feature selection resulted with 541 features that were selected by the feature selection in at least 1 fold. However by trying only those features that were selected at least in 8 out of 10 folds, turned out to result with better prediction score. This resulted in the reduced set having 394 attributes for the best score I could achieve. I did not have chance to explore the effect of the number of features further, but clearly it may have been an important factor.

## VI. Classification

I used Random Forest [5] as the basic classifier. An interesting twist was applying a multi-class meta classifier which resulted with a classifier that had multiple Random Forests, one for each class. This approach was effectively comparing the class records vs. remaining records for each class. This step, although not strictly required, resulted in improved classification score.

One of the challenges with applying Random Forest effectively is selection of optimal number of features used for each tree. In the case of competitions it is typically done by trial and error approach. That was the case in this case – I experimented with different numbers of features per tree and for the particular feature set the numbers between 40 and 80 features seemed to work well. For the best score I could achieve, each Random Forest had 1000 trees. The number of features for each tree was limited to 40.

## VII. Conclusions

In this paper I presented an approach to classification of multivariate time series. The approach was developed for the data mining competition and this approach led to scoring the second high result from nearly 80 solution.

I believe that the approach presented in this paper can be easily generalized to similar problems for which multiple measurements in form of time series are available. It should be expected that different features may turn out to be more predictive or even different classifier may prove to be more suitable. Actually, the author used this approach to another competition where the method allowed to achieve the highest score of nearly 50 submitted solutions.

## Acknowledgment

The author would like to thank to Krasin Georgiev for inspiring discussions.

## References

[1] Meina, M., Janusz, A., Rykaczewski, K., Ślęzak, D., Celmer, B., and Krasuski, A., "Tagging Firefighter Activities at the Emergency Scene: Summary of AAIAâĂŹ15 Data Mining Competition at Knowledge Pit", Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, 2015.

[2] Janusz, A., Krasuski, A., Stawicki, S., Rosiak, M., Slezak, D., and Hung Son Nguyen, "Key risk factors for Polish State Fire Service: A Data Mining Competition at Knowledge Pit," Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on pp.345–354, 7-10 Sept. 2014, doi: 10.15439/2014F507.

[3] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1. 2009.

[4] Hall, M. A., "Correlation-based Feature Subset Selection for Machine Learning". Hamilton, New Zealand. 1998.

[5] Breiman, L., "Random Forests", Machine Learning, Volume 45, Issue 1, pp. 5-32. October 2001.

DOI: 10.15439/2015F424
ACSIS, Vol. 5

# Clustering Approach to the Problem of Human Activity Recognition using Motion Data

Szymon Wawrzyniak
Nicolaus Copernicus
University
Toruń, Poland
Email: wawrzyniaksz@wp.pl

Wojciech Niemiro
Nicolaus Copernicus
University and
Warsaw University
Warsaw, Poland
Email: wniemiro@gmail.com

***Abstract*—This paper describes authors' solution to the task set in *AAIA'15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene* (https://knowledgepit.fedcsis.org/contest/view.php?id=106). Method involves LDA classification on a preprocessed time series data with a unique label transformation technique using K-Means clustering. Data were collected from accelerometer and gyroscope readings.**

## I. Introduction

ACTIVITY recognition of a person using motion data aims to label actions and activities. The task has applications in medicine, sports and surveillance, depending on the technology used. One of the most interesting approaches is sensor-based human activity recognition using data collected from accelerometer and gyroscope readings. By applying machine learning methods to the time series data and labels prepared by the experts, it becomes possible to classify a single person's motion or more general set of motions that can be assigned to a specific group, for instance: firefighters. With the use of a tracking system we can monitor their position and infer potential threats during the action. Such classification task was introduced to the contestants of *AAIA'15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene* [1].

The organizers provided contestants with data generated by ”smart jacket”—a wearable set of body sensors for monitoring kinematics and psychophysical condition of firefighters. The sensors were registering firefighter's vital functions (i.e. ECG, heart rate, respiration rate, skin temperature) and movement (i.e. seven sets of accelerometers and gyroscopes placed on torso, hands, arms and legs).

Each row represented a window data. The consecutive devices' data were placed one after another, creating 17242 elements long row. First 42 elements described vital functions and last 17200 elements (43 chunks of 400 readings) - the movement. One chunk stood for the time between a reading and the start of time window. The window data consisted of triaxial readings (from every accelerometer/gyroscope) that were collected during the 1.8s period (every 4.5ms, on average). Figure 1 presents some window data from three accelerometers. The importance of Z axis (for the classification of posture) can be easily noticed (torso). We can observe long right leg motion that started the whole body move, and during it, a short and firm motion of the left leg appeared. It is already hard to link it to any of the postures, but it can be interpreted as three-step move (right leg → left leg → right leg). Torso Z axis (blue) reading points out a motion toward bottom that was disrupted when left leg slowed down. It is probably crouching. Similar analysis can be performed for the activity label, but it seems to be a much more complicated task.

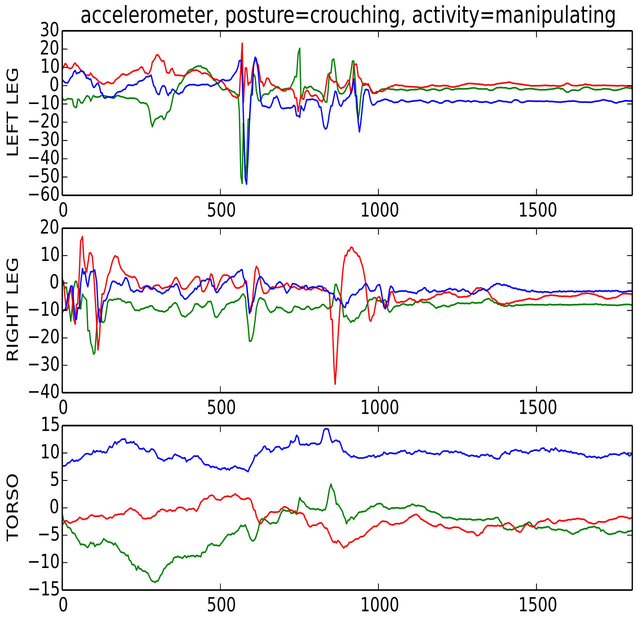


Fig. 1. Triaxial accelerometer readings from different body parts.

There were three sets of data: one training (20000 rows) and two testing datasets (respectively 2000 and 18000 rows). The first testing dataset was a random selection of observations from a bigger set (20000 rows) and obviously the second one was only the remaining part. All the data were labeled by experts. There were two labels (posture and main activities). The task was to find a classification method that maximizes the *score* value. The contestants had to compute (a solution $s$ that consisted of) two vectors of label predictions (a matrix with two $preds$ columns). Each one was compared to the proper

978-83-60810-66-8/$25.00©2015, IEEE

vector of label values (from a matrix with two $labels$ columns) to calculate the final score in the following manner: for $l$ - total amount of label values, $i \in \{1, \ldots, l\}$ (if we replace names with numbers) and $j \in \{1, \ldots, 18000\}$ we have

$$ACC_i(preds, labels) = \frac{|\{j : preds_j = labels_j = i\}|}{|\{j : labels_j = i\}|}$$

$$BAC(preds, labels) = \left( \sum_{i=1}^{l} ACC_i(preds, labels) \right) / l$$

and we denote by:

$BAC_p$ – balanced accuracy for labels describing the posture,
$BAC_a$ – bal. acc. for labels describing the main activity,

then the final score in the competition for a solution s will be computed as:

$$score(s) = (BAC_p(s) + 2 * BAC_a(s)) / 3.$$

To gain satisfactory performance and reduce the amount of data needed to perform the task, authors proposed the following steps:

1) preprocessing - data columns were processed to produce new and less numerous set of columns;
2) label transformation - new data, original labels and K-Means clustering method were used to generate new and more numerous set of label values;
3) classification - new data and new label(s) provided input for learning. The resulting classifier allowed authors to predict new label values for testing set (testing data had to be transformed in the same way as the training data in the preprocessing stage). In the end these new label values were translated to original label values.

It's important to notice that authors' method requires every new dataset to pass the preprocessing stage, because the classification is performed on the preprocessing's output. Clustering is perfomed (only once) for the training dataset.

## II. Related work

There are various approaches to the problem of activity recognition, for instance: feature extraction ($FE$) using moving window over time series and subsequent matching. The raw data are converted differently and treated as input for the main processing. Many works describe sensor-based systems, e.g. [2], Their authors create usually specific set of devices (including accelerometers) to collect data for a model training. The data can be collected from one or more subjects. Using more than one subject can significantly affect the overall accuracy. In paper [4] authors contribute a linear-time method for extracting features from acceleration sensor signals in order to identify human activities, that is based on Support Vector Machines ($SVM$) classifier with a linear kernel. For three types of activities they gained high accuracy ($> 92\%$ for each one), when one person's activity was the data source. When the number of subjects was increased by 4, they gained about $10\%$ loss. The amount of 20 was critical for the efficiency in the case of one of these activities (the accuracy decreased to $16.67\%$). The loss is serious, but it must be considered that if we produce realtime systems, it is expected to prefer simpler but robust systems.

Feature extraction is used to produce new attributes and is usually combined with machine learning algorithms. It is supplied with the data in a form of windows that were collected arbitrarily from the raw dataset. Every window includes readings from devices and is converted to one row usually by setting readings one after another.

Authors of [6] present the whole procedure from receiving the acceleration signal to the final classification. The simplest method is to transform every window to a vector of statistics. More sophisticated methods include Principal Component Analysis ($PCA$), Linear Discriminant Analysis ($LDA$), Fourier Transform or Autoregressive Model. There are various techniques to enrich context awareness by adding enviromental variables and analyzing vital signs, but there are no effective techniques to automatically select windows or choose proper windows length. The size of a window is strongly related to the activity and type of extracted features. Authors mentioned windows from 0.08 up to 30 seconds long which they found during their research. It is also needed to reject attributes that contain redundant or irrelevant information and perform Feature Selection ($FS$). The common method is Minimum Redundancy Maximum Relevance [7], which minimizes average mutual information between selected features and maximizes mutual information between classes and features. There are many different algorithms applied to the learning stage, including decision trees, Naive Bayes, Bayesian Networks, $SVM$, Hidden Markov Models, regression methods and $k$ Nearest Neighbours ($kNN$). The most standard evaluation metric is the $accuracy$ measure, which is the total fraction of correctly predicted label values.

In subsequent matching we introduce time series set of patterns and use some technique to measure similarity like distance measure (Euclidean, cosine etc.). There are effective ways to handle shifts by using Dynamic Time Warping (as in [2]). Since it is not a statistical approach, the pattern data should be carefully chosen. The choice is much more difficult, when the data are noisy and variations within classes are large.

Some authors include (to the moving window approach) special representation like bag-of-features (BoF) representation [5], which produces local features vectors out of windows smaller than the activity vectors itself, that allow them to create a "motion" vocabulary. The authors even compare their framework to another obtained using subsequent matching approach. Across all vocabulary sizes, the average misclassification error for subsequent matching approach ranges from $37\%$ to $54\%$. Although the BoF framework performs better for almost every size value, it is more important that it becomes stable in the sense of the misclassification error (for the vocabulary sizes of 50 and more, the variation decreases significantly).

## III. Clustering approach to the competition task

The basic strategy of finding the best solution was to follow the *score* value changes and apply such method modifications that brought even small improvement. In general, it was indeed a trial and error approach. It should be emphasized that the training dataset and test dataset were obtained from recordings of different groups of firefighters and that fact could lead to overfitting issues making crossvalidation useless (such situation is mentioned at the end of (III-A)).

After simple preprocessing stage (that is described in next paragraphs) the calculations were performed for different classifiers and three most efficient were left for further research. Those three classifiers used $LDA$, $SVM$ and $kNN$ algorithm. When authors decided to include label transformation, they left only $LDA$ and $SVM$ classifiers (and examined the performance of $SVM$ classifier with One vs Rest strategy). When there was no $score$ improvement some simple modifications of the preprocessing stage were included (by adding more statistical values to the set).

The three-step method is based on the feature extraction approach. Authors had no previous experience and wanted to introduce themselves to this kind of methodology. Preprocessing stage consists of generating statistical values and scaling. Each row is taken and reshaped to a matrix that represents a window data. We count $10$ statistics and scale columns separately (by substracting mean and dividing by standard deviation). We obtain new data and cluster observations to create new labels. The algorithm for creating new labels requires two kinds of data: label values and labels from the clustering. K-Means clustering method was chosen. The algorithm produces vector of new label values. The new training data and the new label values are an input for machine learning algorithms. Before the testing stage it is necessary to preprocess the testing data the same way as described above. When we have trained the classifier and produced new testing data we can perform prediction. The label values that we obtain are the ones that need to be translated to the original values.

### A. Preprocessing

Authors decided to choose only movement data for the learning process. Due to time constraints and other obligations there was no opportunity to focus on vital data, which seemed to be much more complicated to analyze.

Each row in dataset represeted one observation. There were a total of $42$ devices (accelerometers and gyroscopes), which generated simultaneous data for a short period (every row was reshaped to a window that had $42$ time series columns). Each time series column consisted of $400$ readings.

$$(D_1, \ldots, D_{16800}) \to$$
$$\to \left| D_{(i \bmod 400+1)(i \operatorname{div} 400+1)} \right|_{i=1,\ldots,16800}$$

The starting point was to keep all the $42$ sources of data, but it had to be preprocessed to lower its dimension. For each device and axis data there were created $10$ new attributes (minimum, maximum, mean, standard deviation, first quartile, median, third quartile, interquartile range, skewness and kurtosis). If we introduce some simplification and write that $D^{jk}$ stands for a column of $j$th device's $k$th statistic value (for every observation), then we can easily write new dataset as

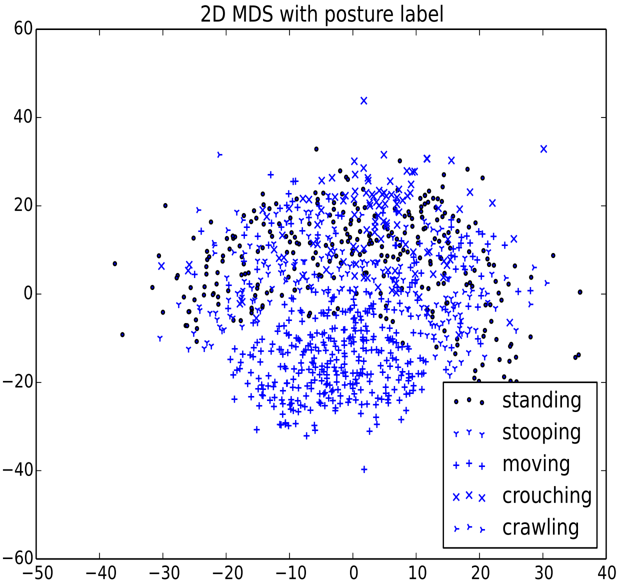


Fig. 2. The result of 2D MDS with posture label.

$$D = (D^{jk}),$$

for $j = 1, \ldots, 42$
and $k \in \{max, min, std, q25, med, q75, IQR, skew, kurt\}$.

$D$ consisted of $420$ columns, where every $D^{jk}$ was standarized.

When a crossvalidation was performed (using randomly selected halves of the unstandarized data), the crossvalidation $score$ values were always greater than $0.9$. Using the same classifiers to generate solutions ended up with receiving competition $score$s around $0.2$.

When the final classifier set was chosen, a new (ineffective and finally abandoned) strategy was introduced to improve the $score$ by adding new quantiles ($0.1$ and $0.9$ quantiles and/or $0.4$ and $0.6$ quantiles).

### B. A premise to label modification

It was an original authors' idea to improve $score$ by transforming the given set of label values. Furthermore, a new set of label values had to be generated according to the new data $D$. The foundation of this procedure lies in assumption that some of label values form too vast data aggregates due to physical differences between firefighters (there was no information about the number). The task would be much more complicated if these aggregates were intermingled. Multi-Dimensional Scaling (MDS) gives some insight into data structure. Figure 2 shows two-dimensional MDS performed on $D$ ($1000$ observations) plotted with distinction on the first label values.

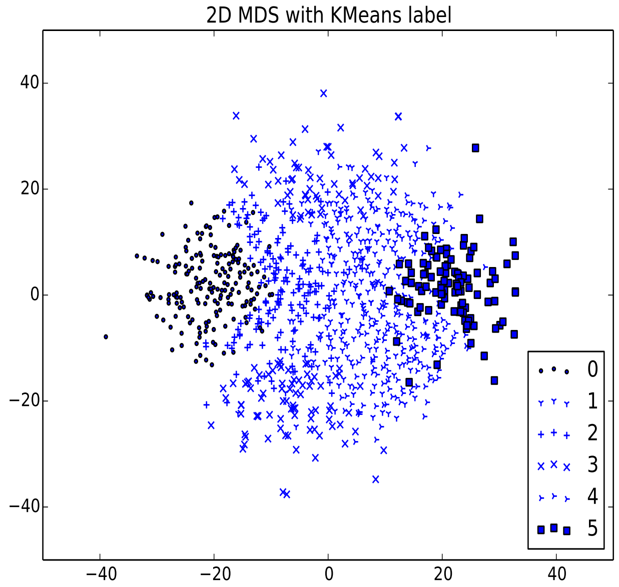


Fig. 3. The result of 2D MDS with K-Means label.

Most of the elements labeled as 'moving' form a structure that is easily separable. On the other hand, if we consider 'standing' posture there is no compact subset that could be separated from other points. It is important to note, that correspondingly larger weights are assigned to smaller sets (for the purposes of *score* evaluation), then any dispersion should be treated as equally serious. Therefore there is strong reason to think that this kind of posture could be difficult to classify. If we label previous data with K-Means (for $k = 6$, Figure 3) we get another graph that strongly indicates existence of apparent groups. It should be understood that the example is only an outline to the idea and does not imply any general statement. Similar proceedings can be made for the main movement label. There would be some clarity issues with plotting (16 values to be marked), so authors decided to leave it. Obviously, the actual calculations for the second label were performed and the resulting graph points to the same conclusions.

### C. Label transformation

K-Means clustering was performed on $D$ and it tagged observations with $CLV$ labels. To generate the label for further training we needed original label values $LV$. Both $CLV$ and $LV$ were vectors of equal length of 20000. We got two mappings

$$D \xrightarrow{K-Means} CLV,$$

$$(LV, CLV) \xrightarrow{algorithm} NLV.$$

### D. Algorithm

K-Means algorithm (python's $sklearn.cluster.KMeans$ implementation which uses $kmeans{+}{+}$ algorithm for choosing initial seeds) allowed us to label observations. The second mapping was carried out as shown in Algorithm 1.

**Algorithm 1** The algorithm for creating a new label

```
Require: $LV$ (names replaced with numbers), $CLV$
Ensure: $NLV$
  Find the length $lenLV$ of $LV$ vector
  Create $NLV$ vector with the length equal to $lenLV$
  for each unique $i \in LV$ do
      Create $l\_num$ vector of every observation number
      that has its $LV$ equal to $i$
      Create $l\_list$ vector by finding all unique $CLV$s
      within the observations with numbers from $l\_num$
      Create a dictionary $dict$ with keys from $l\_list$, such that
      for an element $l\_list[j]$ assign a value of $(lenLV \cdot j + i)$
      for each consecutive $k$ from $l\_num$ do
          $NLV[k] \leftarrow dict(CLV[k])$
      end for
  end for
```

To find the best $k$ value authors proposed to seek for a local maximum within a tiny range of small $k$ values (from 2 to 8). There were only 13 test computations at all, because prediction of values can be done separately for every label. We set one $k$ value (for one label and other $k$ was fixed for the remaining label), performed clustering, learning (and classification) and then checked the *score* of our prediction. The maximum *score* (one of seven) indicated $k$ value, which we treated as optimal value and from this moment this $k$ value was fixed. If we found more than one maximum *score*, then we would chose the one with the smallest $k$ value. The *score* maximum for the second label indicated a pair of $k$ values (the one fixed and the new found) that was the optimal choice. Finally, $k = 6$ was selected for the first label and $k = 5$ was selected for the second label.

When the $k$ value was chosen, it was necessary to create and preserve a dictionary to be able to translate every new label value to the original label value. The other way was to propose a method that did employ the clustering to the testing stage, but authors wanted to avoid such requirement. If our machines provided better computing performance we would examine larger range of $k$ values. Then, if the optimum was very high, the computing time would be significantly extended (and the clustering would have to be performed twice, if we included it into testing stage). Due to possible practical applications such effect is not expected.

The probability distribution of the new label seems to be quite obvious to count. If we assume that $lv_i$ stands for $i$th possible label value and $clv_j$ stands for $j$th possible cluster level value (derived from K-Means clustering), and we assume that the probability distributions of appropiate labels are known, then the probability of the occurence of the new label value ($nlv_{i,j}$) produced using those two values can be counted as below

$$P(nlv_{i,j}) = P(clv_j|lv_i) = \frac{P(clv_j, lv_i)}{P(lv_i)}.$$

Furthermore, a big score difference (about $0.2$) was observed between the approach that implied use of label transformations as described above and the one that did not (in the favor of the first one).

Another approach is to use only one label with all possible (96) combinations of values. When it was applied to the training labels, there could be found only 24 values. However, the $scores$ were so low that in authors' opinion there was no premise to examine it more precisely (for instance by applying K-Means).

### E. Classification

The classifier training phase was conducted using newly generated data and vectors of new label values. Then, the trained classifier was used to predict label values for the preprocessed testing dataset. Its output consisted of values that had to be translated to original label values. It had to be an unambiguous transformation, so a new set of label values had to be more numerous than the original one (an equinumerous case is trivial). This assumption was provided if we examine the specifics of the label transformation algorithm. Let $T$ be the test data preprocessed in the same way as $D$, $LV_{pred}$ be the vector of predicted labels (which is part of the submitted solution) and respectively $NLV_{pred}$ be the vector of predicted new label values. Generally, we can enclose this stage in three steps

$$(D, NLV) \xrightarrow{learning} classifier,$$

$$T \xrightarrow{classifier} NLV_{pred},$$

$$NLV_{pred} \xrightarrow{dictionary} LV_{pred}.$$

### F. Results

Authors constructed several classifiers (pairs of classifiers, one for every label):

- $LDA$ with clustering approach, $k \in \{2, \dots, 8\}$, 13 classifiers, dataset $D$;
- $LDA$ with clustering approach, $k \in \{2, \dots, 6\}$, 9 classifiers for a dataset obtained in the same manner as $D$, but with different set of statistics
$$(min1, max1, std1, q25, med1, q75, kurt1, skew);$$
- $LDA$ with clustering approach, $k \in \{2, \dots, 6\}$, 9 classifiers for a dataset obtained in the same manner as $D$, but with different set of statistics
$$(min1, max1, std1, med1, kurt1, skew1);$$
- $LDA$ with clustering approach, fixed $k$s, a dataset obtained in the same manner as $D$, but with different set of statistics
$$(max, min, std, q10, q25, med, q75, q90, IQR, skew, kurt);$$
- $LDA$ with clustering approach, fixed $k$s, a dataset obtained in the same manner as $D$, but with different set of statistics
$$(max, min, std, q25, q40, med, q60, q75, IQR, skew, kurt);$$
- $LDA$ with clustering approach, fixed $k$s, a dataset obtained in the same manner as $D$, but with different set of statistics
$$(max, min, std, q10, q25, q40, med, q60, q75, q90, IQR, skew, kurt);$$
- $LDA$ without clustering approach, dataset $D$;
- $SVM$ (kernels: $rbf$, $linear$) for the first label, $LDA$ for the second label, both with clustering approach, $k \in \{3, 4, 5\}$, $k$ for the second label was fixed, 6 classifiers, dataset $D$;
- $SVM$ (kernels: $rbf$, $linear$) with One vs Rest strategy for the first label, $LDA$ for the second label, both with clustering approach, $k \in \{3, 4, 5\}$, $k$ for the second label was fixed, 6 classifiers, dataset $D$;
- $kNN$ without clustering approach, (number of neighbours) $k \in \{5, 7, 10\}$, 5 classifiers, dataset $D$;

To sum up the overall classification performance, $LDA$ classifier that uses approach described in part (III-D) with $k = 6$ for the first label and $k = 5$ for the second label performed best among other classifiers.

The *score* for the less numerous testing set was $0.8067$ and for the remaining set of observations was $0.77288289$.

## IV. Conclusion

When the $SVM$ classifier with One vs Rest strategy (for the first label) and $LDA$ classifier (for the second label) were combined, they produced only slightly worse *score*s (about $0.79$) than the best one ($0.8067$), assuming that predictions for the second label remained unchanged. An interesting observation was made when the vector of predictions for the first label was compared to the corresponding output of the final solution. Only about half of label value predictions were the same, which probably indicates a problem with a proper classification of the first label.

Taking into account simplicity and significant improvements in evaluation of the model due to application of K-Means approach and small difference between first and second stage scores, the authors' method can be treated as suitable for the activity recognition task. Furthermore, the model behaved stable (for different $k$s) and performed very efficiently in terms of learning time.

The authors are also aware that the results and performance can be dependant on implementation issues, so it should be noticed that everything was prepared in Python with the use of $sklearn$ library (standarization, clustering, classifiers) [3].

## References

[1] Michał Meina and Andrzej Janusz and Krzysztof Rykaczewski and Dominik Ślęzak and Bartosz Celmer and Adam Krasuski, *Tagging Firefighter Activities at the Emergency Scene: Summary of AAIA'15 Data Mining Competition at Knowledge Pit*, Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, editors: Maria Ganzha and Leszek A. Maciaszek and Marcin Paprzycki, In print September 2015

[2] Michał Meina and Bartosz Celmer and Krzysztof Rykaczewski, *Towards Robust Framework for On-line Human Activity Reporting Using Accelerometer Readings*, Proceedings of the Active Media Technology - 10th International Conference, Warsaw, 2014, pp. 347-358, http://dx.doi.org/10.1007/978-3-319-09912-5_29

[3] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, No. 12, 2011, pp. 2825-2830

[4] M. Khan, S. I. Ahamed, M. Rahman, and R. O. Smith, *A feature extraction method for real time human activity recognition on cell phones*, isQoLT, 2011. http://www.mridulkhan.com/pdf/khan-QoL10.pdf

[5] Mi Zhang and Alexander A. Sawchuk. *Motion primitive-based human activity recognition using a bag-of-features approach*, Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI '12), ACM, New York, pp. 631-640. http://doi.acm.org/10.1145/2110363.2110433

[6] Oscar D. Lara, Miguel A. Labrador, *A Survey on Human Activity Recognition using Wearable Sensors*, Communications Surveys & Tutorials, IEEE, vol. 15, No. 3, 2013, pp. 1192-1209, doi:10.1109/surv.2012.110112.00192

[7] Hanchuan Peng (Member, IEEE), Fuhui Long, and Chris Ding, *Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, No. 8, 2005, http://doi.ieeecomputersociety.org/10.1109/TPAMI.2005.159

# 8th Workshop on Computational Optimization

MANY real world problems arising in engineering, economics, medicine and other domains can be formulated as optimization tasks. These problems are frequently characterized by non-convex, non-differentiable, discontinuous, noisy or dynamic objective functions and constraints which ask for adequate computational methods.

The aim of this workshop is to stimulate the communication between researchers working on different fields of optimization and practitioners who need reliable and efficient computational optimization methods.

We invite original contributions related to both theoretical and practical aspects of optimization methods.

## Topics

The list of topics includes, but is not limited to:

- unconstrained and constrained optimization
- combinatorial optimization
- continues optimization
- global optimization
- multiobjective optimization
- optimization in dynamic and/or noisy environments
- large scale optimization
- parallel and distributed approaches in optimization
- random search algorithms, simulated annealing, tabu search and other derivative free optimization methods
- nature inspired optimization methods (evolutionary algorithms, ant colony optimization, particle swarm optimization, immune artificial systems etc)
- hybrid optimization algorithms involving natural computing techniques and other global and local optimization methods
- computational biology and optimization
- distance geometry and applications
- optimization methods for learning processes and data mining
- application of optimization methods on real life and industrial problems
- computational optimization methods in statistics, econometrics, finance, physics, chemistry, biology, medicine, engineering etc

## Event Chairs

**Fidanova, Stefka,** Bulgarian Academy of Sciences, Bulgaria

**Mucherino, Antonio,** INRIA, France

**Zaharie, Daniela,** West University of Timisoara, Romania

## Program Committee

**Bartl, David,** University of Ostrava, Czech Republic

**Bonates, Tibérius,** Universidade Federal do Ceará, Brazil

**Breaban, Mihaela**

**Chira, Camelia**

**Fidanova, Stefka,** Bulgarian Academy of Science

**Gonçalves, Douglas,** Universidade Federal de Santa Catarina, Brazil

**Gualandi, Stefano**

**Hosobe, Hiroshi,** Hosei University, Japan

**Iiduka, Hideaki,** Kyushu Institute of Technology, Japan

**Krislock, Nathan,** Northern Illinois University, United States

**Lavor, Carlile,** IMECC-UNICAMP, Brazil

**Marinov, Pencho,** Bulgarian Academy of Science, Bulgaria

**Mihalas, Stelian,** West University of Timisoara

**Muscalagiu, Ionel,** Politehnica University Timisoara, Romania

**Nannicini , Giacomo**

**Ninin, Jordan,** ENSTA-Bretagne, France

**Parsopoulos, Konstantinos,** University of Patras

**Pintea, Camelia,** Tehnical University Cluj-Napoca, Romania

**Pop, Petrica**

**Roeva, Olympia,** Institute of Biophysics and Biomedical Engineering, Bulgaria

**Siarry, Patrick,** Universite Paris XII Val de Marne, France

**Slezak, Dominik,** University of Warsaw & Infobright Inc., Poland

**Stefanov, Stefan,** South-West University ""Neofit Rilski, Bulgaria

**Stuetzle, Thomas,** Université Libre de Bruxelles (ULB), Belgium

**Suganthan, Ponnuthurai Nagaratnam,** Nanyang Technological University, Singapore

**Tamir, Tami,** The Interdisciplinary Center (IDC), Israel

**Tvrdik, Josef,** University of Ostrava, Czech Republic

**Voller, Zach**

**Vrahatis, Michael,** University of Patras, Greece

**Zilinskas, Antanas,** Vilnius University, Lithuania

DOI: 10.15439/2015F178
ACSIS, Vol. 5

# InterCriteria Analysis of Crossover and Mutation Rates Relations in Simple Genetic Algorithm

Maria Angelova, Olympia Roeva, Tania Pencheva
Institute of Biophysics and Biomedical Engineering
Bulgarian Academy of Sciences
105 Acad. G. Bonchev Str., Sofia 1113, Bulgaria
Email:{maria.angelova, olympia, tania.pencheva}@biomed.bas.bg

***Abstract*—In this investigation recently developed InterCriteria Analysis (ICA) is applied to examine the influences of two main genetic algorithms parameters – crossover and mutation rates during the model parameter identification of *S. cerevisiae* and *E. coli* fermentation processes. The apparatuses of index matrices and intuitionistic fuzzy sets, which are the core of ICA, are used to establish the relations between investigated genetic algorithms parameters, from one hand, and fermentation process model parameters, from the other hand. The obtained results after ICA application are analysed towards convergence time and model accuracy and some conclusions about derived interactions are reported.**

## I. Introduction

INTERCRITERIA Analysis (ICA), given in details in [3], is a contemporary approach for multi-criteria decision making. ICA implements the apparatuses of index matrices (IM) and intuitionistic fuzzy sets (IFS) in order to compare some criteria reflecting the behaviour of considered objects. Recently ICA has been successfully applied for EU Member States competitiveness analysis [8], thus provoking the search for further ICA applications. The idea to implement ICA in the field of tuning of the optimization techniques parameters has intuitively appeared.

Fermentation processes (FP) are objects of increased research interest because of their widespread use in different branches of industry. The FP modeling and optimization are a real challenge for the investigators due to the fact that FP models have complex structures based on systems of non-linear differential equations with several specific growth rates [9]. The choice of appropriate model parameter identification procedure is the most important problem for FP adequate modeling. Among others biologically inspired optimization techniques, genetic algorithms (GA) [12] has been proved as a global search method [10] for solving different engineering and optimization problems [18], among that for parameter identification of FP [1], [16], [17], [19]. GA efficiency strongly depends on the tuning of different operators, functions, and parameters. These settings are specifically implemented to different problems. Current investigation is focused on the examining the impact of two of the main GA parameters, namely crossover ($xovr$) and mutation ($mutr$) rates. Simple GA (SGA) is applied for the purposes of model parameter identification of two fed-batch FP – *S. cerevisiae* and *E. coli*. Both yeast and bacteria have numerous applications in food and pharmaceutical industries. Also both microorganisms are widely used as model organisms in genetic engineering and cell biology due to their well known metabolic pathways [11], [14].

In this investigation the obtained results from SGA parameter identification of considered here FP models are used to determine some dependencies between some criteria preliminary defined as of significant importance. The establishment of the influences and relations between criteria – model parameters, from one hand, and GA parameters crossover and mutation rates, from the other hand, is performed by the ICA implementation. This is expected to lead to additional exploring of the models or the relation between models and optimization algorithm outcomes, which will be valuable especially in the case of modelling of living systems, such as FP.

## II. Problem Formulation

### *A. Mathematical models of fermentation processes*

Two Case studies are going to be presented here – for the fermentation processes of *S. cerevisiae* (Case study 1) and of *E. coli* (Case study 2).

**Case study 1. *S. cerevisiae* fed-batch fermentation model**

The mathematical model of *S. cerevisiae* fed-batch process is presented by the following non-linear differential equations system [16]:

$$\frac{dX}{dt} = (\mu_{2S}\frac{S}{S+k_S} + \mu_{2E}\frac{E}{E+k_E})X - \frac{F_{in}}{V}X \qquad (1)$$

$$\frac{dS}{dt} = -\frac{\mu_{2S}}{Y_{S/X}}\frac{S}{S+k_S}X + \frac{F_{in}}{V}(S_{in} - S) \qquad (2)$$

$$\frac{dV}{dt} = F_{in} \qquad (3)$$

where $X$ is the biomass concentration, [g/l]; $S$ – substrate concentration, [g/l]; $E$ – ethanol concentration, [g/l]; $F_{in}$ – feeding rate, [l/h]; $V$ – bioreactor volume, [l]; $S_{in}$ – substrate concentration in the feeding solution, [g/l]; $\mu_{2S}$, $\mu_{2E}$ – the maximum values of the specific growth rates, [1/h]; $k_S$, $k_E$ – saturation constants, [g/l]; $Y_{S/X}$ – yield coefficient, [-].

For the considered here model (Eqs. (1-3)), the vector of parameters to be identified is as follows:

$$p_1 = [\mu_{2S}\ \mu_{2E}\ k_S\ k_E\ Y_{S/X}].$$

978-83-60810-66-8/$25.00©2015, IEEE

**Case study 2. *E. coli* fed-batch fermentation model**

The mathematical model of *E. coli* fed-batch process is presented by the following non-linear differential equations system [9], [16]:

$$\frac{dX}{dt} = \mu_{max} \frac{S}{k_S + S} X - \frac{F_{in}}{V} X \tag{4}$$

$$\frac{dS}{dt} = -\frac{\mu_{max}}{Y_{S/X}} \frac{S}{S + k_S} X + \frac{F_{in}}{V} (S_{in} - S) \tag{5}$$

$$\frac{dV}{dt} = F_{in} \tag{6}$$

where all notations keep their meaning as described above, and, additionally, $\mu_{max}$ is the maximum value of the specific growth rate, [1/h].

For the considered here model (Eqs. (4-6)), the vector of parameters to be identified is as follows:

$$p_2 = [\mu_{max}\ k_S\ Y_{S/X}].$$

Model parameters identification of both fed-batch FP is performed based on experimental data for biomass, glucose and ethanol concentrations. The detailed description of the process conditions and experimental data can be found in [16].

### B. Optimization criterion

The objective function is designed aiming at identification of parameter vectors $p_1$ and $p_2$ in order to obtain the best fit to a data set and is defined as:

$$J = \sum_{i=1}^{m} (X_{\text{exp}}(i) - X_{\text{mod}}(i))^2 + \sum_{i=1}^{n} (S_{\text{exp}}(i) - S_{\text{mod}}(i))^2 \rightarrow \min \tag{7}$$

where $m$ and $n$ are the experimental data dimensions; $X_{\text{exp}}$ and $S_{\text{exp}}$ – available experimental data for biomass and substrate; $X_{\text{mod}}$ and $S_{\text{mod}}$ – model predictions for biomass and substrate with a given model parameter vector.

### C. Simple genetic algorithms for parameter identification

Simple genetic algorithm, initially presented in Goldberg [12], searches a global optimal solution using three main genetic operators in a sequence selection, crossover and mutation. SGA starts with a creation of a randomly generated initial population. Each solution is then evaluated and assigned a fitness value. According to the fitness function, the most suitable solutions are selected. After that, crossover proceeds to form a new offspring. Mutation is then applied with determinate probability aiming to prevent falling of all solutions into a local optimum. The execution of GA has been repeated until the termination criterion (i.e. reached number of populations, or found solution with a specified tolerance, etc.) is satisfied.

Crossover and mutation are among of the most important operators that can increase the efficiency of GA. The crossover operator is used to generate offspring by exchanging bits in a pair of parents chromosomes chosen from the population. Crossover occurs with a crossover probability (crossover rate, $xovr$), that indicates a ratio of how many couples will be picked for mating. The mutation operator changes some elements in selected chromosomes with a mutation probability (mutation rate, $mutr$). As such, the operator introduces genetic diversity and helps GA to escape the local optimum. It is well known that optimal crossover and mutation rates vary for different problems and the success of GA depends on their choice [13]. Usually, determining what rates of crossover and mutation should be used is doing on the trial-and-error basis. In the literature there exist a number of guidelines how crossover and mutation rates to be tuned [12], [13], [15]. Recommended values of crossover rate are high, usually in the range 0.5-1.0 [13], [15]. On the other hand, low mutation rate values for preventing search process to be turn into a simple random search are commonly adopted in GA. Typical values of mutation rate are in the range 0.001-0.1 [13], [15].

In this investigation the impact of crossover and mutation rates is going to be examined choosing different values of the both GA parameters. In Case study 1 SGA is applied with the following values of crossover rate: $xovr$ = {0.65; 0.75; 0.85; 0.95}, while in Case study 2 – with $xovr$ = {0.5; 0.6; 0.7; 0.8; 0.9; 1}. Due to the specific peculiarities of two fed-batch FP, again different strategies were applied for mutation rates in both Case studies. In Case study 1 SGA is applied with the following values of mutation rate: $mutr$ = {0.02; 0.04; 0.06; 0.08; 0.1}, while in Case study 2 – with $mutr$ = {0.001; 0.01; 0.1; 0.5; 1}. The selected values of $xovr$ and $mutr$ are chosen based on the following prerequisites: i) concerning the recommended by the literature values and trying to comprise different values in the ranges for both Case studies [12], [13], [15]; ii) concerning the previous authors' experience of modelling of FP using GA [1], [16], [17], [18], [19]. All other GA operators and parameters are tuned as presented in [1], [19].

## III. InterCriteria Analysis

InterCriteria analysis, based on the apparatuses of index matrices and intuitionistic fuzzy sets, is given in details in [3]. Here, for a completeness, the proposed idea is briefly presented.

An intuitionistic fuzzy pair (IFP) [4] is an ordered pair of real non-negative numbers $\langle a, b\rangle$, where $a, b \in [0, 1]$ and $a + b \leq 1$, that is used as an evaluation of some object or process. According to [4], the components ($a$ and $b$) of IFP might be interpreted as degrees of "membership" and "non-membership" to a given set, degrees of "agreement" and "disagreement", degrees of "validity" and "non-validity", degrees of "correctness" and "non-correctness", etc.

The apparatus of index matrices (IM) is presented initially in [5] and discussed in more details in [6], [7]. For the purposes of ICA application, the initial index set consists of the criteria (for rows) and objects (for columns) with the IM elements assumed to be real numbers. Further, an IM with index sets consisting of the criteria (for rows and for columns) with IFP elements determining the degrees of correspondence between the respective criteria is constructed, as it is doing to be briefly presented below.

Let the initial IM is presented in the form of Eq. (8), where, for every $p, q,\ (1 \leq p \leq m, 1 \leq q \leq n)$, $C_p$ is a criterion, taking part in the evaluation; $O_q$ – an object to be evaluated; $a_{C_p,O_q}$ – a real number or another object, that is comparable about relation $R$ with the other $a$-objects, so that for each $i, j, k$: $R(a_{C_k,O_i}, a_{C_k,O_j})$ is defined. Let $\overline{R}$ be the dual relation of $R$ in the sense that if $R$ is satisfied, then $\overline{R}$ is not satisfied, and vice versa. For example, if "$R$" is the relation "$<$", then $\overline{R}$ is the relation "$>$", and vice versa. If $S^{\mu}_{k,l}$ is the number of cases in which $R(a_{C_k,O_i}, a_{C_k,O_j})$ and $R(a_{C_l,O_i}, a_{C_l,O_j})$ are simultaneously satisfied, while $S^{\nu}_{k,l}$ is the number of cases is which $R(a_{C_k,O_i}, a_{C_k,O_j})$ and $\overline{R}(a_{C_l,O_i}, a_{C_l,O_j})$ are simultaneously satisfied, it is obvious, that

$$S^{\mu}_{k,l} + S^{\nu}_{k,l} \leq \frac{n(n-1)}{2}.$$

Further, for every $k, l$, satisfying $1 \leq k < l \leq m$, and for $n \geq 2$,

$$\mu_{C_k,C_l} = 2\frac{S^{\mu}_{k,l}}{n(n-1)}, \quad \nu_{C_k,C_l} = 2\frac{S^{\nu}_{k,l}}{n(n-1)} \tag{9}$$

are defined. Therefore, $\langle \mu_{C_k,C_l}, \nu_{C_k,C_l} \rangle$ is an IFP. Next, the following IM is constructed:

$$\begin{array}{c|ccc} & C_1 & \dots & C_m \\ \hline C_1 & \langle \mu_{C_1,C_1}, \nu_{C_1,C_1} \rangle & \dots & \langle \mu_{C_1,C_m}, \nu_{C_1,C_m} \rangle \\ \vdots & \vdots & \ddots & \vdots \\ C_m & \langle \mu_{C_m,C_1}, \nu_{C_m,C_1} \rangle & \dots & \langle \mu_{C_m,C_m}, \nu_{C_m,C_m} \rangle \end{array},$$

that determines the degrees of correspondence between criteria $C_1, ..., C_m$.

In the most of the obtained pairs $\langle \mu_{C_k,C_l}, \nu_{C_k,C_l} \rangle$, the sum $\mu_{C_k,C_l} + \nu_{C_k,C_l}$ is equal to 1. However, there may be some pairs, for which this sum is less than 1. The difference

$$\pi_{C_k,C_l} = 1 - \mu_{C_k,C_l} - \nu_{C_k,C_l} \tag{10}$$

is considered as a degree of "uncertainty".

## IV. Numerical Results and Discussion

In order to obtain reliable results for convergence time, optimization criterion and model parameters estimations, thirty independent runs of SGA have been performed for each value of crossover and mutation rates for both examined here Case studies. Obtained results have been averaged and two IMs are constructed for each Case study, involving values for crossover or mutation rates, respectively. In other words, altogether four IMs are constructed: IMs $A_{1(xovr)}$ (Eq. (11)) and $A_{1(mutr)}$ (Eq. (12)) for the Case study 1 and IMs $A_{2(xovr)}$ (Eq. (13)) and $A_{2(mutr)}$ (Eq. (14)) for the Case study 2.

IM $A_{1(xovr)}$ presents average estimates of the model parameters $\mu_{2S}, \mu_{2E}, k_S, k_E$, and $Y_{S/X}$, as well as the resulting convergence time $T$ and objective function value $J$, respectively for $xovr = \{0.65; 0.75; 0.85; 0.95\}$, denoted as $\text{GA}^{xovr}_{1,1} \div \text{GA}^{xovr}_{1,4}$. In the same way, IM $A_{1(mutr)}$ presents the results for $\mu_{2S}, \mu_{2E}, k_S, k_E$, and $Y_{S/X}, T, J$ and $mutr$, respectively for $mutr = \{0.02; 0.04; 0.06; 0.08; 0.1\}$, denoted as $\text{GA}^{mutr}_{1,1} \div \text{GA}^{mutr}_{1,5}$.

IMs $A_{2(xovr)}$ and $A_{2(mutr)}$ for the Case study 2 have been created by analogy with the Case study 1.

Based on Eq. (9), ICA algorithm calculates the IFP $\langle \mu, \nu \rangle$ for every two pairs of considered criteria based on the obtained IMs $A_{1(xovr)}$, $A_{1(mutr)}$, $A_{2(xovr)}$ and $A_{2(mutr)}$. Values of $\pi$ (Eq. (10)) are calculated too. Obtained results are grouped in Table 1 for both Case studies, considering dependences between crossover and mutation rates, optimization criterion, convergence time and model parameters themselves.

Applied here non-linear models for two Case studies (respectively Eqs. (1)-(3) and Eqs. (4)-(6)) are a prerequisite some closer relations between observed criteria to be expected after ICA application. On the other hand, some differences in the parameters relations might appear caused by the different specific growth rates in *S. cerevisiae* and *E. coli* FP.

As it could be seen from Table 1, there is a strong relation between $T \leftrightarrow xovr/mutr$ for the Case study 1, while in the Case study 2 a weak relation is observed. The similar discrepancy is identified in the correlation between $Y_{S/X} \leftrightarrow xovr/mutr$: in the Case study 2 there is a strong relation for GA parameter $xovr$, while in the Case study 1 – a weak. These discrepancies might be explained by the stochastic nature of GA. Crossover rate strongly influences evaluation of model parameter $\mu_{2E}$ in Case study 1. In the Case study 2, there is a significant indication for high correlation between $J \leftrightarrow mutr$. For the rest of model parameters the observed correlations are weak – there are no significant dependencies between $T$ and these parameters.

Going further in investigation of relations between algorithm accuracy $J$ and model parameters, higher $\mu$-values is observed between $Y_{S/X} \leftrightarrow J$ in Case study 1 and for GA parameter $mutr$. Less stronger correlations are identified in the Case study 1 for GA parameter $mutr$ between $\mu_{2S} \leftrightarrow J$, as well as in Case study 2 for GA parameter $mutr$ between $Y_{S/X} \leftrightarrow J$. These similarities are caused by the physical meaning of considered model parameters. For the rest of parameters the observed correlations are weak – there are no significant dependencies between these parameters and $J$.

When considering the influence of convergence time $T$ over the model parameters, higher $\mu$ is observed in pairs $\mu_{2E} \leftrightarrow T$ in Case study 1 for GA parameter $xovr$. In the Case study 2, higher $\mu$-values are observed between $\mu_{max} \leftrightarrow T$ and $k_S \leftrightarrow T$ for $mutr$ GA parameter. Observed $\mu$-values for the rest of pairs of model parameters and $T$ show that there are no significant correlations between them.

The last group of examined correlations is between model parameters themselves in both considered Case studies. Different model structures in both FP complicate the extraction of some common correlations. Although that fact, there are some coincidences for both Case studies. In the Case study 1 for GA parameter $xovr$, the strongest correlations are found respectively for $\mu_{2S} \leftrightarrow k_S$ and $\mu_{2S} \leftrightarrow Y_{S/X}$, while less stronger correlations are identified for the pairs $k_S \leftrightarrow k_E$,

$$A = \begin{array}{c|ccccccc} & O_1 & \ldots & O_k & \ldots & O_l & \ldots & O_n \\ \hline C_1 & a_{C_1,O_1} & \cdots & a_{C_1,O_k} & \cdots & a_{C_1,O_l} & \cdots & a_{C_1,O_n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_i & a_{C_i,O_1} & \cdots & a_{C_i,O_k} & \cdots & a_{C_i,O_l} & \cdots & a_{C_i,O_n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_j & a_{C_j,O_1} & \cdots & a_{C_j,O_k} & \cdots & a_{C_j,O_l} & \cdots & a_{C_j,O_n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_m & a_{C_m,O_1} & \cdots & a_{C_m,O_k} & \cdots & a_{C_m,O_l} & \cdots & a_{C_m,O_n} \end{array}, \tag{8}$$

**Case study 1**, IM $A_{1(xovr)}$:

$$A_{1(xovr)} = \begin{array}{r|cccc} & \text{GA}_{1,1}^{xovr} & \text{GA}_{1,2}^{xovr} & \text{GA}_{1,3}^{xovr} & \text{GA}_{1,4}^{xovr} \\ \hline J & 0.0222 & 0.0222 & 0.0222 & 0.0221 \\ T & 69.140600 & 70.212400 & 69.475000 & 71.359200 \\ xovr & 0.65 & 0.75 & 0.85 & 0.95 \\ \mu_{2S} & 0.962120 & 0.949840 & 0.974790 & 0.923920 \\ \mu_{2E} & 0.103840 & 0.107940 & 0.115320 & 0.129580 \\ k_S & 0.124640 & 0.119580 & 0.128700 & 0.119780 \\ k_E & 0.799020 & 0.798700 & 0.798860 & 0.798960 \\ Y_{S/X} & 0.417885 & 0.413705 & 0.413850 & 0.409500 \end{array} \tag{11}$$

**Case study 1**, IM $A_{1(mutr)}$:

$$A_{1(mutr)} = \begin{array}{r|ccccc} & \text{GA}_{1,1}^{mutr} & \text{GA}_{1,2}^{mutr} & \text{GA}_{1,3}^{mutr} & \text{GA}_{1,4}^{mutr} & \text{GA}_{1,5}^{mutr} \\ \hline J & 0.022200 & 0.022167 & 0.022133 & 0.022300 & 0.022100 \\ T & 71.677000 & 76.104333 & 90.479000 & 101.400667 & 98.161667 \\ mutr & 0.02 & 0.04 & 0.06 & 0.08 & 0.1 \\ \mu_{2S} & 0.963433 & 0.987333 & 0.943333 & 0.960033 & 0.914933 \\ \mu_{2E} & 0.113100 & 0.111900 & 0.129733 & 0.094967 & 0.146100 \\ k_S & 0.124000 & 0.123333 & 0.128167 & 0.117033 & 0.121300 \\ k_E & 0.799867 & 0.799500 & 0.799600 & 0.792433 & 0.797833 \\ Y_{S/X} & 0.410841 & 0.411348 & 0.407914 & 0.421965 & 0.398290 \end{array} \tag{12}$$

**Case study 2**, IM $A_{2(xovr)}$:

$$A_{2(xovr)} = \begin{array}{r|cccccc} & \text{GA}_{2,1}^{xovr} & \text{GA}_{2,2}^{xovr} & \text{GA}_{2,3}^{xovr} & \text{GA}_{2,4}^{xovr} & \text{GA}_{2,5}^{xovr} & \text{GA}_{2,6}^{xovr} \\ \hline J & 0.010700 & 0.000310 & 0.000320 & 0.000170 & 0.000450 & 0.000310 \\ T & 143.156 & 77.782 & 218.234 & 104.719 & 158.078 & 86.953 \\ xovr & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1 \\ \mu_{max} & 0.553000 & 0.549000 & 0.550000 & 0.551000 & 0.549000 & 0.548000 \\ k_S & 0.011700 & 0.009800 & 0.010100 & 0.010000 & 0.009800 & 0.009900 \\ Y_{S/X} & 0.500275 & 0.499975 & 0.499950 & 0.500000 & 0.500250 & 0.500500 \end{array} \tag{13}$$

$Y_{S/X} \leftrightarrow k_S$ and $Y_{S/X} \leftrightarrow k_E$. Considering GA parameter $mutr$, the strongest correlations are between $k_S \leftrightarrow k_E$ and $\mu_{2S} \leftrightarrow Y_{S/X}$. Comparing to Case study 2 and taking into account the simpler specific growth rate model structure, the similar result for the pair $\mu_{max} \leftrightarrow k_S$ is observed. The highest correlation is observed for both GA parameters $xovr$ and $mutr$. These strong parameter dependencies are again caused by the physical meaning of FP models parameters. For the rest correlations between model parameters themselves, the $\mu$-values are low – there are no significant dependencies.

It is also interesting to be noted that during the investigation of $xovr$ influence, there are some pairs of considered criteria with reported degree of uncertainty $\pi$. For the Case study 1, all observed appearances of degrees of uncertainty are in pairs with optimization criterion value, while in Case study 2 – in pairs of optimization criterion value or specific growth rate. All

**Case study 2**, IM $A_{2(mutr)}$:

$$A_{2(mutr)} = \begin{array}{c|ccccc} & \text{GA}_{2,1}^{mutr} & \text{GA}_{2,2}^{mutr} & \text{GA}_{2,3}^{mutr} & \text{GA}_{2,4}^{mutr} & \text{GA}_{2,5}^{mutr} \\ \hline J & 0.019000 & 0.000360 & 0.007300 & 4.130700 & 25.622800 \\ T & 53.250000 & 116.594000 & 193.641000 & 70.937000 & 39.234000 \\ mutr & 0.001 & 0.01 & 0.1 & 0.5 & 1 \\ \mu_{max} & 0.546000 & 0.550000 & 0.554000 & 0.599000 & 0.432000 \\ k_S & 0.007800 & 0.010200 & 0.011000 & 0.044400 & 0.002300 \\ Y_{S/X} & 0.499500 & 0.500250 & 0.500501 & 0.500325 & 0.518403 \end{array} \tag{14}$$

these facts have an obvious explanation – as it can be seen from IM $A_{1(xovr)}$ for Case study 1, there are equal values for optimization criterion value. In analogy, as seen from IM $A_{2(xovr)}$, there are equal evaluations of optimization criterion value and specific growth rate in Case study 2. Observed equal values logically cause an uncertainty and makes difficult the process of decision making.

As a summary of ICA implementation, the following main results might be outlined:

- Considered GA parameters $xovr$ and $mutr$ show a high correlation with $T$ in both Case studies. In Case study 2, parameter $mutr$ is in a high correlation with $J$ and model parameter $Y_{S/X}$. The values of $xovr$ and $mutr$ reflect on $T$ because of the more complex model used in Case study 1 [1], [12], [15]. In opposite, the more simple model structure in Case study 2 allows the relations between $mutr$ and $J$ and one of the most sensitive model parameter $Y_{S/X}$ [17] to be outlined.
- When looking at $T$ and $J$ relations, strong connections are observed for $J \leftrightarrow Y_{S/X}$, especially in Case study 1; between specific growth rates (respectively $\mu_{2E}$ and $\mu_{max}$) and $T$ in both Case studies, as well as for $k_S \leftrightarrow T$ in Case study 2. The stochastic nature of GA is a preposition of a relatively small number of observed strong relations [10], [12], [15].
- In the last group of examined correlations between model parameters themselves, higher dependencies are obtained between specific growth rates $\mu_{2S}$ and $\mu_{max}$ and model parameter $k_S$ in both Case studies, especially in Case study 2 at GA parameter $mutr$. Considering Case study 1, strong correlations are observed for $k_S \leftrightarrow k_E$ and for $Y_{S/X} \leftrightarrow \mu_{2S}$. The ascertained results are caused by the physical meaning of FP models parameters, as well as by the strong non-linearity of FP model structures [9], [11], [14], [16].

## V. Conclusion

In this paper the recently proposed InterCriteria Analysis is applied to establish the relations and dependencies between two GAs parameters – crossover and mutation rates, on one hand, and convergence time, model accuracy and FP model parameters, on the other hand. Simple GA with different values of crossover and mutation rates is used for parameter identification of two FP models – of yeast *S. cerevisiae* and bacteria *E. coli*.

The obtained results from ICA show some existing relations and dependencies that result from the physical meaning of the model parameters, on one hand, and from stochastic nature of the considered meta-heuristic, on the other hand. Moreover, derived additional knowledge for ascertained correlations will be useful in further identification procedures of FP models and, in general, for more accurate SGA application.

## Acknowledgment

The work is supported by the Bulgarian National Scientific Fund under the grant DFNI-I-02-5 “InterCriteria Analysis – A New Approach to Decision Making”.

## References

[1] M. Angelova, *Modified Genetic Algorithms and Intuitionistic Fuzzy Logic for Parameter Identification of Fed-batch Cultivation Model*, PhD Thesis, Sofia, 2014. (in Bulgarian)

[2] K. Atanassov, *On Intuitionistic Fuzzy Sets Theory*, Springer, Berlin, 2012, DOI 10.1007/978-3-642-29127-2.

[3] K. Atanassov, D. Mavrov and V. Atanassova, “Intercriteria Decision Making: A New Approach for Multicriteria Decision Making, Based on Index Matrices and Intuitionistic Fuzzy Sets”, *Issues in on Intuitionistic Fuzzy Sets and Generalized Nets*, vol. 11, 2014, pp. 1–8.

[4] K. Atanassov, E. Szmidt and J. Kacprzyk, “On Intuitionistic Fuzzy Pairs”, *Notes on Intuitionistic Fuzzy Sets*, vol. 19, No. 3, 2013, pp. 1–13.

[5] K. Atanassov, “Generalized Index Matrices”, *Compt. rend. Acad. Bulg. Sci.*, vol. 40, No. 11, 1987, pp. 15–18.

[6] K. Atanassov, “On Index Matrices, Part 1: Standard Cases”, *Advanced Studies in Contemporary Mathematics*, vol. 20, No. 2, 2010, pp. 291–302.

[7] K. Atanassov, “On Index Matrices, Part 2: Intuitionistic Fuzzy Case”, *Proceedings of the Jangjeon Mathematical Society*, vol. 13, No. 2, 2010, pp. 121–126.

[8] V. Atanassova, L. Doukovska, K. Atanassov and D. Mavrov, “Intercriteria Decision Making Approach to EU Member States Competitiveness Analysis”, *in International Symposium on Business Modeling and Software Design*, 2014, pp. 289–294, DOI 10.5220/0005427302890294.

[9] G. Bastin and D. Dochain, *On-line Estimation and Adaptive Control of Bioreactors*, Elsevier Scientific Publications, 1991.

[10] I. Boussaid, J. Lepagnot and P. Siarry, “A Survey on Optimization Metaheuristics”, *Information Sciences*, vol. 237, 2013, pp. 82–117, DOI 10.1016/j.ins.2013.02.041.

[11] R. J. Dickinson and M. Schweizer, *Metabolism and Molecular Physiology of Saccharomyces cerevisiae*, 2nd Edition, CRC Press, 2004.

[12] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley Longman, London, 2006.

[13] W. Lin, W. Lee and T. Hong, “Adapting Crossover and Mutation Rates in Genetic Algorithms”, *Journal of Information Science and Engineering*, vol. 19, 2003, pp. 889–903.

[14] Y. Matsuoka and K. Shimizu, “Importance of Understanding the Main Metabolic Regulation in Response to the Specific Pathway Mutation for Metabolic Engineering of *Escherichia coli*”, *Comput Struct Biotechnol Journal*, vol. 3, No. 4, 2012, e201210018, DOI 10.5936/csbj.201210018.

TABLE I
RESULTS FROM THE ICA OF *S. serevisiae* AND *E. coli* FED-BATCH FERMENTATION PROCESSES

| **Correlation** | ***S. serevisiae* fed-batch fermentation process** | | | | ***E. coli* fed-batch fermentation process** | | | |
|---|---|---|---|---|---|---|---|---|
| | ***xovr*** | | ***mutr*** | | ***xovr*** | | ***mutr*** | |
| | $\langle\mu,\nu\rangle$ | $\pi$ | $\langle\mu,\nu\rangle$ | $\pi$ | $\langle\mu,\nu\rangle$ | $\pi$ | $\langle\mu,\nu\rangle$ | $\pi$ |
| $T \leftrightarrow xovr/mutr$ | 0.8, 0.2 | 0 | 0.9, 0.1 | 0 | 0.5, 0.5 | 0 | 0.4, 0.6 | 0 |
| $J \leftrightarrow xovr/mutr$ | 0, 0.5 | 0.5 | 0.3, 0.7 | 0 | 0.3, 0.6 | 0.1 | 0.8, 0.2 | 0 |
| $\mu_{2S} \leftrightarrow xovr/mutr$ | 0.3, 0.7 | 0 | 0.2, 0.8 | 0 | | | | |
| $\mu_{2E} \leftrightarrow xovr/mutr$ | 1, 0 | 0 | 0.6, 0,4 | 0 | | | | |
| $\mu_{max} \leftrightarrow xovr/mutr$ | | | | | 0.2, 0.7 | 0.1 | 0.6, 0.4 | 0 |
| $Y_{S/X} \leftrightarrow xovr/mutr$ | 0.2, 0.8 | 0 | 0.4, 0.6 | 0 | 0.7, 0.3 | 0 | 0.9, 0.1 | 0 |
| $k_S \leftrightarrow xovr/mutr$ | 0.5, 0.5 | 0 | 0.3, 0.7 | 0 | 0.3, 0.7 | 0 | 0.6, 0.4 | 0 |
| $k_E \leftrightarrow xovr/mutr$ | 0.5, 0.5 | 0 | 0.2, 0.8 | 0 | | | | |
| $T \leftrightarrow J$ | 0, 0.5 | 0.5 | 0.4, 0.6 | 0 | 0.6, 0.3 | 0.1 | 0.2, 0.8 | 0 |
| $\mu_{2S} \leftrightarrow J$ | 0.5, 0 | 0.5 | 0.7, 0.3 | 0 | | | | |
| $\mu_{2E} \leftrightarrow J$ | 0, 0.5 | 0.5 | 0.1, 0.9 | 0 | | | | |
| $\mu_{max} \leftrightarrow J$ | | | | | 0.5, 0.3 | 0.2 | 0.4, 0.6 | 0 |
| $Y_{S/X} \leftrightarrow J$ | 0.5, 0 | 0 | 0.9, 0.1 | 0 | 0.5, 0.4 | 0.1 | 0.7, 0.3 | 0 |
| $k_S \leftrightarrow J$ | 0.3, 0.2 | 0.5 | 0.4, 0.6 | 0 | 0.5, 0.3 | 0.2 | 0.4, 0.6 | 0 |
| $k_E \leftrightarrow J$ | 0.2, 0.3 | 0.5 | 0.5, 0.5 | 0 | | | | |
| $\mu_{2S} \leftrightarrow T$ | 0.2, 0.8 | 0 | 0.3, 0.7 | 0 | | | | |
| $\mu_{2E} \leftrightarrow T$ | 0.8, 0.2 | 0 | 0.5, 0.5 | 0 | | | | |
| $\mu_{max} \leftrightarrow T$ | | | | | 0.6, 0.3 | 0.1 | 0.8, 0.2 | 0 |
| $Y_{S/X} \leftrightarrow T$ | 0, 1 | 0 | 0.5, 0.5 | 0 | 0.4, 0.6 | 0 | 0.5, 0.5 | 0 |
| $k_S \leftrightarrow T$ | 0.3, 0.7 | 0 | 0.2, 0.8 | 0 | 0.7, 0.3 | 0 | 0.8, 0.2 | 0 |
| $k_S \leftrightarrow T$ | 0.3, 0.7 | 0 | 0.1, 0.9 | 0 | | | | |
| $\mu_{2S} \leftrightarrow \mu_{2E}$ | 0.3, 0.7 | 0 | 0.2, 0.8 | 0 | | | | |
| $\mu_{2S} \leftrightarrow k_S$ | 0.8, 0.2 | 0 | 0.5, 0.5 | 0 | | | | |
| $\mu_{2E} \leftrightarrow k_S$ | 0.5, 0.5 | 0 | 0.7, 0.3 | 0 | | | | |
| $\mu_{max} \leftrightarrow k_S$ | | | | | 0.8, 0.2 | 0 | 1, 0 | 0 |
| $\mu_{2S} \leftrightarrow k_E$ | 0.5, 0.5 | 0 | 0.6, 0.4 | 0 | | | | |
| $\mu_{2E} \leftrightarrow k_E$ | 0.5, 0.5 | 0 | 0.6, 0.4 | 0 | | | | |
| $k_S \leftrightarrow k_E$ | 0.7, 0.3 | 0 | 0.9, 0.1 | 0 | | | | |
| $Y_{S/X} \leftrightarrow \mu_{2S}$ | 0.8, 0.2 | 0 | 0.8, 0.2 | 0 | | | | |
| $Y_{S/X} \leftrightarrow \mu_{2E}$ | 0.2, 0.8 | 0 | 0, 1 | 0 | | | | |
| $Y_{S/X} \leftrightarrow \mu_{max}$ | | | | | 0.4, 0.5 | 0.1 | 0.5, 0.5 | 0 |
| $Y_{S/X} \leftrightarrow k_S$ | 0.7, 0.3 | 0 | 0.3, 0.7 | 0 | 0.5, 0.5 | 0 | 0.5, 0.5 | 0 |
| $Y_{S/X} \leftrightarrow k_E$ | 0.7, 0.3 | 0 | 0.4, 0.6 | 0 | | | | |

[15] M. Obitko, Genetic Algorithms, available at http://www.obitko.com/tutorials/genetic-algorithms/

[16] T. Pencheva, O. Roeva and I. Hristozov, *Functional State Approach to Fermentation Processes Modelling*, Prof. Marin Drinov Academic Publishing House, Sofia, 2006.

[17] O. Roeva, T. Pencheva, B. Hitzmann and St. Tzonkov, "A Genetic Algorithms Based Approach for Identification of *Escherichia coli* Fed-batch Fermentation", *International Journal Bioautomation*, vol. 1, 2004, pp. 30–41.

[18] O. Roeva (Ed.), *Real-world Application of Genetic Algorithms*, InTech, 2012, DOI 10.5772/2674.

[19] O. Roeva, "Genetic Algorithm and Firefly Algorithm Hybrid Schemes for Cultivation Processes Modelling", *Transactions on Computational Collective Intelligence XVII*, R. Kowalczyk, A. Fred and F. Joaquim (Eds.), vol. 8790, 2014, pp. 196–211, DOI 10.1007/978-3-662-44994-3_10

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 425–434

DOI: 10.15439/2015F137
ACSIS, Vol. 5

# Correlation clustering by contraction

László ASZALÓS, Tamás MIHÁLYDEÁK
University of Debrecen
Faculty of Informatics
26 Kassai str., H4028 Debrecen, Hungary
Email: {aszalos.laszlo, mihalydeak.tamas}@inf.unideb.hu

***Abstract*—We suggest an effective method for solving the problem of correlation clustering. This method is based on an extension of a partial tolerance relation to clusters. We present several implementation of this method using different data structures, and we show a method to speed up the execution by a quasi-parallelism.**

## I. Introduction

THE principle of minimum total potential energy (MTPE) is a fundamental concept used in physics, chemistry, biology and engineering. It asserts that a structure or body shall deform or displace to a position that minimizes the total potential energy. This concept could be used at other fields, too. In this paper, we show its application in clustering. The clustering is an important tool of unsupervised learning. Its task is to group the objects in such a way, that the objects in one group (cluster) are similar, and the objects from different groups are dissimilar, so it generates an equivalence relation: the objects being in the same cluster. If we want to apply the principle MTPE, then we can say that the objects aim to achive a situation in which they are in a cluster containing minimal number of dissimilar, and maximal number of similar objects. In the last fifty years, many different clustering methods were invented based on different demands.

Correlation clustering is a new method, Bansal at al. published a paper in 2004, proving several of its properties, and gave a fast, but not quite optimal algorithm to solve the problem [1]. Naturally, correlation clustering has a predecessor. Zahn proposed this problem in 1965, but using a very different approach [2]. The main question is the following: which equivalence relation is the closest to a given tolerance (reflexive and symmetric) relation? Bansal et al. have shown, that this is an NP-hard problem [1]. The number of equivalence relations of $n$ objects, i.e. the number of partitions of a set containing $n$ elements is given by Bell numbers $B_n$, where $B_1 = 1$, $B_n = \sum_{i=1}^{n-1} \binom{n-1}{k} B_k$. It can be easily checked that the Bell numbers grow exponentially. Therefore if $n > 15$, in a general case we cannot achieve the optimal partition by exhaustive search, thus we need to use some optimization methods, which do not give optimal solutions, but help us achieve a near-optimal one.

This kind of clustering has many applications: image segmentation [3], identification of biologically relevant groups of genes [4], examination of social coalitions [5], improvement of recommendation systems [6] reduction of energy consumption [7], modelling physical processes [8], (soft) classification [9], [10], etc.

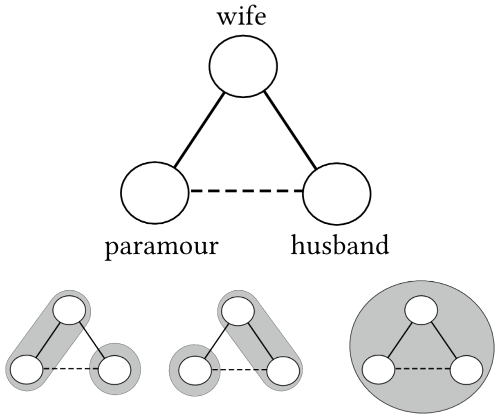


Fig. 1. Minimal frustrated graph and its optimal partitions

At correlation clustering, those cases where two dissimilar objects are in the same cluster, or two similar objects are in different clusters are treated as a conflict. Thus the main question could be rewritten as: which partition generates the minimal number of conflicts? It can be shown, that in the general case—where the transitivity does not hold for the tolerance relation—the number of conflicts is a positive number for all partitions of a given set of objects. Let us take a graph on Figure 1, where the similarity is denoted by solid, and dissimilarity by dashed lines. In case of persons, the similarity and dissimilarity are treated as liking or disliking each other, respectively. Mathematically, the similarity is described as a weighted graph, where 1 denotes similarity, and -1 denotes dissimilarity. As the absolute value of these weights are the same, thus it is enough to only use the signs of the weights. Hence the graph is called a *signed graph* in the literature. The lower part of this figure shows the three optimal partitions from the five, each of them containing only one conflict. As all partitions of the graph on Figure 1 have at least one conflict, it is called a *frustrated graph*, and this one is the smallest such graph.

If the correlation clustering is expressed as an optimization problem, the traditional optimization methods (hill-climbing, genetic algorithm, simulated annealing, etc.) could be used in order to solve it. We have implemented and compared the results in [11]. With these methods the authors were able to determine a near optimal partition of signed graphs with 500 nodes.

978-83-60810-66-8/$25.00©2015, IEEE

In this paper we introduce a new method which was invented directly to solve the problem of correlation clustering, and can use the specialities of the problem. The main idea is extremely simple, but it needs several witty concepts, to get a fast and effective algorithm.

Before going into details, let us see the hierarchical clustering—where the solution is usually received as a result of a series of contractions. The user needs to choose a dissimilarity metric of clusters. These metrics are based on the distance of the members (objects) of the clusters. Next, the hierarchical clustering constructs a hierarchy, which starts with singleton clusters, and each higher level is earned by joining two clusters of the previous level. The last level consists of one cluster which contains all the objects. This clustering is an automatic process, where the user selects one level of this hierarchy (and the equivalence relation generated by it), and uses accordingly.

In case of correlation clustering the user has limited options. If the hierarchy, according to the tolerance relation is generated, the cost function (i.e. the number of conflicts) can be calculated for each level (for each partition) of the hierarchy, and we take the minimal. This does not mean, that we reach the optimal partition for each tolerance relation. This hierarchy has $n$ levels ($n$ partition) from the $B_n$ possible ones, e.g. if $n$ is 500, then we have 500 levels, and $1.6 \cdot 10^{844}$ different partitions. Therefore we need to choose carefully which partitions to add to the hierarchy.

There are two ways to construct the hierarchy: bottom-up and top-down. In the latter, we split one cluster into two. If the original cluster contains $k$ objects, there are $2^k - 2$ ways to split it, thus the full search is not feasible. There are other methods to find a good split, but the hierarchy construction by splitting is not so common. The bottom-up way is based on joining clusters. There are $k(k-1)/2$ ways of joining clusters, and after the contraction of two clusters we update the pre-calculated distances of clusters accordingly, as if we had $k$ clusters before contraction.

In this paper we will use Python programs to present the algorithms. The Python programming language is perfect tool for prototyping:

- Python code listings are shorter than pseudo codes, because we can use program-libraries without long explanations or repetition of well-known algorithms.
- Python has high level data structures (list of set, set of lists, etc.) which simplify the programs.
- Although Python programs are slow, we are not interested in exact running times, but in the time rates: the effect of replacing an algorithm with a better one?

We tested programs on one core of a 2.3 GHz double core processor under Linux and Python 3.4.0. The times are given in seconds.

The first implementation uses the most trivial data structure to store a graph: the adjacency matrix. The adjacency matrix of a tolerance relation contains elements $-1$ and $1$ only, so its graph is total. The asymptotic behaviour of the correlation clustering of tolerance relation is known from [8]. If we allow zero values in the adjacency matrix, i.e. we have a partial tolerance relation, the behaviour of the clustering changes. The more zeros are in the adjacency matrix, the bigger the difference in behaviour. These behaviours are not yet described or explained mathematically, computer experiments for many objects are needed to discover the exact tendencies. Hence for us the most interesting graphs are the sparse graphs. In general, it is easy to implement the generation of Erdős–Rényi type random graphs [12], but it is hard to ensure that it generates a connected graph for small probability parameter $p$. Therefore in our experiments we used Barabási-Albert type random graphs [12] (we refer to them as BA graphs in the following) where the connectivity follows from the algorithm of the generation.

The structure of the paper is the following: Section 2 explains the correlation clustering and shows the result of joining two clusters. In Section 3 we present a naive contraction method. Next we specialize the method for sparse graphs. In Section 5 we give the quasi-parallel version of the specialized algorithms. Finally we discuss our plans and conclude the results.

## II. Correlation clustering

In the paper we use the following notations: $V$ denotes the set of the objects, and $T \subset V \times V$ the tolerance relation defined on $V$. We handle a partition as a function $p : V \to \{1, \ldots, n\}$. The objects $x$ and $y$ are in a common cluster, if $p(x) = p(y)$. We say that objects $x$ and $y$ are in conflict at given tolerance relation $T$ and partition $p$ iff value of $c_T^p(x,y) = 1$ in (1), i.e. if they are similar and are in different clusters, or if they are dissimilar and in the same cluster.

$$c_T^p(x,y) \leftarrow \begin{cases} 1 & \text{if } (x,y) \in T \text{ and } p(x) \neq p(y) \\ 1 & \text{if } (x,y) \notin T \text{ and } p(x) = p(y) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We are ready to define the cost function of relation $T$ according to partition $p$:

$$c_T(p) \leftarrow \frac{1}{2} \sum c_T^p(x,y) = \sum_{x<y} c_T^p(x,y) \tag{2}$$

As relation $T$ is symmetric, we sum $c_T^p$ twice for each pair, or restrict the summing in order to use each pair only once. The objects are usually represented with numbers, hence we can calculate the cost function as the last part of (2) shows.

Our task is to determine the value of $\min_p c_T(p)$, and one partition $p$ for which $c_T(p)$ is minimal. Unfortunately this exact value cannot be determined in practical cases, except for some very special tolerance relations. Hence we can get only approximative, near optimal solutions.

The statistical software R has 6 different distance functions for determining distance of objects. In our case the tolerance relation replaces these metrics. The same software has 8 different cluster distance functions. None of them is suitable for our needs.

Let us see, what is the result of joining two clusters according to the cost function. Let $A$ and $B$ be these clusters, moreover let us denote the partition before and after joining by $p$ and $q$, respectively.

1) If $\{x, y\} \cap (A \cup B) = \emptyset$, then $c_T^p(x, y) = c_T^q(x, y)$ holds.
2) If $x \in (A \cup B)$ and $y \notin (A \cup B)$ (or in reverse), then $c_T^p(x, y) = c_T^q(x, y)$ holds.
3) If $x \in A$ and $y \in B$, then $c_T^p(x, y) = 1 - c_T^q(x, y)$ holds. As objects $x$ and $y$ are from different clusters, $p(x) \neq p(y)$. So if $c_T^p(x, y) = 1$, then $(x, y) \in T$, but after the joining $q(x) = q(y)$, hence $c_T^q(x, y) = 0$. Similarly if $c_T^p(x, y) = 0$, then $(x, y) \notin T$, so $c_T^q(x, y) = 1$.
4) Finally, if $x, y \in A$ (or $x, y \in B$), then $c_T^p(x, y) = c_T^q(x, y)$.

To determine $c_T(q)$ if we know $c_T(p)$, it is enough to calculate the difference $c_T(q) - c_T(p)$. From the previous list it is obvious that we only need to take into account the third item. Before the contraction between clusters $A$ and $B$ there were $\#\{(x, y) | x \in A, y \in B, (x, y) \in T\}$ conflict, and after the contraction these conflicts disappear. But $\#\{(x, y) | x \in A, y \in B, (x, y) \notin T\}$ new conflicts are created by the contraction. The change is the difference of these numbers.

If we treat the conflict of two objects as a distance, then the aggregating function is the difference of two sums. At hierarchical clustering the closest clusters are joined. For us, the contraction method is a greedy algorithm, and we promote the joining of those possible ones which produce maximal profit, which mostly decreases the cost function. These two ideas are confluent, because the difference mentioned before becomes negative, and we join those clusters where the absolute value of this difference is maximal, so where the difference is minimal. Unfortunately, according to its negativity we cannot call the difference as *distance*.

We note that our programs use notations of Bansal, and if the tolerance relation holds for two objects, the number denoting it is positive. From this, it naturally follows that the program uses the differences $c_T(p) - c_T(q)$, i.e. the profits of joining. We only execute the contraction if this difference is positive.

The upper part of the Figure 2 shows a relation. For the sake of simplicity of the picture, this relation is a partial tolerance relation. The middle part of the figure displays the "distances" of the singleton clusters earned from the relation. We formally define it in the next chapter. Finally the lower part of the figure shows the updated distances when some clusters are contracted. These distances are the superposition of the predecessor clusters.

## III. Contraction method

We can now define the Contraction method for correlation clustering. Algorithm 1 is implemented in Python. In the following formulae the brackets refer to the lines of the code. The implementations use the routines of https://www.ics.uci.edu/~eppstein/PADS/UnionFind.py for handling disjoint sets: it contracts two clusters in line 8. We do not present the preprocessing of neither the tolerance relation, nor of the result, but the clustering phase is emphasized.

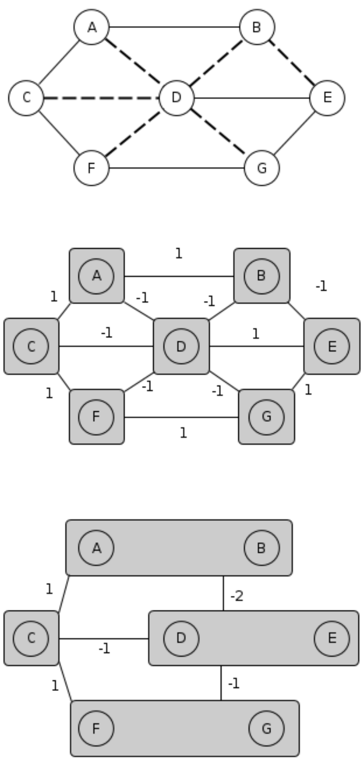


Fig. 2. The partial tolerance relation, the generated "distances" at the beginning and after several contraction steps.

1) Construct a "distance" matrix $D$ based on a tolerance relation $T$:

$$d(i, j) \leftarrow \begin{cases} 1 & \text{if } (i, j) \in T, \\ -1 & \text{if } (i, j) \notin T, \end{cases} \tag{3}$$

Although $d(i, i)$ would be 1 by definition, we set up 0, in order to avoid contracting any clusters by themselves. We denote the preprocessing $T$ and setup of $D$ with dots. [line $3 - 4$]
2) Now, each object—as a singleton cluster—in one-one bijection with rows and columns of $D$. Select one maximal element of matrix $D$, and then its row and column coordinates will refer to the clusters to join. Let these be $x$ and $y$ [line 6].
3) If the maximal element is not positive, the algorithm ends [line 7].
4) Otherwise add to column $x$ the column $y$ (contraction), next delete column $y$, i.e. fill it with zero [lines $9 - 12$]. Then repeat this for the rows.
5) Continue from Step 2. [line 13]

The statistics describing the running time of Algorithm 1 is given in Table I. Here the columns denote graphs with different sizes, and the rows tagged by the value of $q$. This parameter gives the ratio of fulfilment of tolerance relation $T$ among objects. If this number is small, only a few clusters can be joined, so the Contraction method finishes soon and

**Algorithm 1** Naive contraction with matrix

```
import numpy; import UnionFind
N=2000; q=0.4
d = numpy.zeros( (N,N), dtype = numpy.int)
...
uf = UnionFind.UnionFind()
x,y = numpy.unravel_index(d.argmax(),d.shape)
while d[x,y] > 0:
    uf.union(x,y)
    for z in range(N):
        d[x,z] += d[y,z] ; d[z,x] += d[z,y]
    d[...,y] = 0; d[y,...] = 0
    d[x,x] = 0
    x,y = numpy.unravel_index(d.argmax(),
                 d.shape)
```

TABLE I
RUNNING TIMES FOR COMPLETE SIGNED GRAPHS.

| | 100 | 500 | 1000 | 2000 | 5000 |
|---|---|---|---|---|---|
| $q = 0.1$ | 0.017 | 0.598 | 3.015 | 16.080 | 174.511 |
| $q = 0.5$ | 0.027 | 0.773 | 3.736 | 19.152 | 198.269 |
| $q = 0.9$ | 0.029 | 0.785 | 3.789 | 19.226 | 197.325 |

the solution contains many but quite small clusters. But if this ratio is large, there are lot of possibilities to join clusters, and the Contraction method runs for a long time and gives only a few, but big clusters. It can even occur, that we get only one cluster, containing all the elements.

In our previous article we examined the behaviour of contraction of tolerance relation on $n$ objects [13]. For this, we took 101 different values for $q$ from 0 to 1, and for each $q$ we tested the clustering for 10 different relations (graphs), and used their average. By using this implementation to calculate everything for 2000 objects, took more than 5 hours. These are independent calculations, so they can be run in parallel. We note that this program uses the low-level routines of Numpy extensions, which speeds up the execution according to the conventional Python implementation.

This 8-line-long naive implementation does not take into account such details, which are obvious for the reader. As we contract rows and columns, and delete one row and one column (fill up with zeros), in the future the program will not put any non-zero number into these rows and columns, thus it is unnecessary to include these elements of the matrix in the calculations. We could tag them, and later skip them in the cycles. It could be a better solution, however, to change the deletion of a row and a column by swapping them with the last row and column, respectively. Next, we can logically reduce the size of the matrix $D$, and set the upper limit in cycles to the size of the matrix.

Up to here, for any pair of objects the tolerance relation either holds or does not, i.e. the objects are either similar or dissimilar. However, in some cases we have no knowledge about either, thus the relation is partial. (A well-known partial relation is the ordering of $n$-tuples, where the relation holds only if one tuple Pareto dominates the other.) We can use (1) for any partial tolerance relation. By definition, if two objects are neither similar nor dissimilar, there is no conflict between them. The partial tolerance relations are visualized by signed graphs: if two objects are comparable then there is an edge between them, otherwise there is not, like in Figure 2. If the relation holds, the weight of the edge is 1, if not, its weight is -1. If the relation is not defined between the objects, the weight of the non-existent edge could be defined as 0. From these weights we can construct the “distance matrix” $D$ for any partial relation.

TABLE II
RUNNING TIME OF ALGORITHM 1 ON RANDOM ER GRAPHS WITH PARAMETERS $(2000, p_i)$, WHERE THE RATE OF THE POSITIVE EDGES IS $q$

| | $q = 0.1$ | $q = 0.5$ | $q = 0.9$ |
|---|---|---|---|
| $p_1 = 0.9$ | 16.870 | 19.294 | 19.418 |
| $p_2 = 0.5$ | 18.040 | 19.240 | 19.372 |
| $p_3 = 0.1$ | 18.148 | 19.112 | 19.371 |

It is not suprising that Algorithm 1 could be used for partial tolerance relations without any modification. We generated several partial relations with 2000 objects, to test our algorithm. The graphs of the partial tolerance relations were Erdős–Rényi type random graphs, where any two nodes are connected with probability $p_i$. Next, the weight of this edge (if exists) would be 1 with probability $q$ and $-1$ with probability $1 - q$.

The algorithm is the same, it takes the same steps, therefore we can assume, that we get very similar results for clustering. But the data in Table II refutes this assumption. If many edges are missing from the complete graph ($p_i$ is small) and most of the edges are negative ($q$ is small), then the algorithm does not stops after a few steps, as did it for a bigger $p_i$. Similarities force the contractions. The more 1s in $D$, the more contractions are executed by the algorithm, which takes time. The more -1s in $D$—the contractions are obstructed—the less contractions are executed by the algorithm, and the process stops early. The zeros (i.e. missing edges) here decrease the effects of 1 and -1: the number of contractions fall between the two extreme cases (when all zeros are replaced with 1 and -1, respectively).

As we wrote before, Néda at al. described the asymptotic behaviour of correlation clustering for complete graphs [8], thus the computer simulations in this case are not challenging. In case of any signed graph (including the partial tolerance relations) we have conjectures only about the asymptotic behaviour. Néda at al. simulated correlation clustering of BA graphs with 140 nodes in 2009, which we extended to 500 nodes in 2014 [13], and even though the results are impressive, they are not quite sufficient to see the tendencies.

The BA graphs are sparse graphs, they have $O(|V|)$ edges. It is superfluous to reserve $O(|V|^2)$ memory cells in a matrix to store these edges. In the next section we show a memory efficient algorithm to solve the correlation clustering problem for sparse graphs.

## IV. Refining the method

The previous Python implementation shows the crucial questions of the method:

Q1 For constructing a greedy algorithm we need the most profitable contraction, i.e. from the stored and updated (recalculated) values we need to select the biggest one, or one of the biggest ones.

Q2 We need to be able to reach the element $d_{x,y}$ of the matrix $D$, to read it, to update it, or to delete it.

As the previous implementation stores matrix $D$ as a two-dimensional array (matrix), the solution of Q2 is trivial, however to solve Q1, we need a full search in $D$.

In BA graph with ten thousands of nodes, a node is only part of a few edges. This means that in its row of "distance" matrix $D$ there are only a few non-zero elements (which are interesting at calculations). Therefore it is worth to represent the matrix $D$ as a sparse matrix. There are several ways to store a sparse matrix in the memory. The simplest way is to store the list of $(row,\ column,\ value)$ triplets. This storage type is called a *coordinate list*.

In case of using a coordinate list, the solution to Q1 does not change substantially, because a full search is needed in triplets. (Now we have no element without valuable information.) Yet, the solution of Q2 becomes more complicated. Until now, $d_{x,y}$ was reachable in constant time (with pointer-arithmetic). Now the triplets of $D$ are in a list, so in a worst case scenario, a full-search is needed. If this list is ordered, a binary search is enough, but it is very hard to preserve this ordering during contractions:

- If $d_{x,z} = -d_{y,z}$, then after contraction $d_{x,z}$ becomes 0, but we do not want to store this element, so we need to delete it.
- If $d_{x,z} = 0$ and $d_{y,z} \neq 0$, then after the contraction, one new item is created and needs to inserted into the list, while an old one is to deleted.

The hash table could help us solve Q2 effectively, because its speed is near to $O(1)$ at insertion. The hash table is a possible (and fast) implementation of the data structure *associative arrays* (dictionary in Python). Algorithm 2 is the reimplementation of the Contraction method with an associative array.

**Algorithm 2** Contraction by associative array

```
 1  import UnionFind
 2  def contraction(d, s):
 3      d2 = {}
 4      for pair, value in d.items():
 5          x,y = pair
 6          if x in s:
 7              x = s[x]
 8          if y in s:
 9              y = s[y]
10          if x == y:
11              continue
12          if x > y:
13              x,y = y,x
14          if (x,y) in d2:
15              d2[(x,y)] += value
16              if d2[(x,y)] == 0:
17                  del d2[(x,y)]
18          else:
19              d2[(x,y)] = value
20      return d2
21
22  N=1000; q=0.2
23  d = ...
24  uf = UnionFind.UnionFind()
25  if len(d) < 2:
26      return
27  max_d = max(d.values())
28  while max_d > 0:
29      pairs = [pair for pair,value in
30          d.items() if value ==max_d]
31      i,j = pairs[0]; s = {j:i}; uf.union(i,j)
32      d2 = contraction(d,s)
33      if len(d2) < 2:
34          break
35      d = d2.copy()
36      max_d = max(d.values())
```

As before, we only indicate the phase of preprocessing [line 23]. This code is much longer than the previous one, we need to care about more details. While in the previous program, we retrieved the indices of the biggest element of $D$ with one complicated instruction, we needed to take it into pieces here: from the values of the associative array the algorithm selects the biggest one [line 27], and collects the keys belonging to this maximal value [lines 29–30]. From these keys it chooses a pair. We refer the members of this pair as $i$ and $j$. We wish to reuse the part of this code, which describes the contraction, therefore we implemented the contraction step as a function [lines 2–20]. For similar reasons, the pair $(i, j)$ is stored in an associative array, called $s$. Our construction guaranties, that $i < j$, and at contraction we keep the smaller one. This means, that upon meeting a key (which is a coordinate-pair), we need to check whether one of it is equal to $j$ (to the key in the associative array $s$) or not. If it is, then we need to replace it with $i$ (with the value according the key) [lines 6–9]. In the diagonal of $D$, only zero values are allowed, but we do not store them, so if something gets into this diagonal it is ignored [lines 10–11]. We store the indices by ordering [lines 12–13]. If a known pair occurs, then we update its value, whereas unknown pairs are stored [lines 14–19]. Sometimes one cluster is connected with two different clusters, and their effects are opposite. Remember on Figure 1 the husband who loves her wife (+1) but hates her paramour (-1). If the wife left him with her paramour, these numbers add up, and we get 0. So the best if he forget about this new couple. We can speed up our method by not storing zero values, so we omit this one, too [lines 16–17].

Unfortunately the associative arrays do not allow us to pick items by specific needs, hence, to get all the keys in form $(j, \cdot)$ and $(\cdot, j)$, we need traverse the whole associative array, and check whether the actual item is concerned in the contraction or not. In the BA random graphs, the nodes $i$ and $j$ have only a few nearby node (connected to it by an edge), hence this traverse needs extra effort (in comparison to direct access of the edges of a node).

The statistic demonstrating Algorithm 2 is in the first data column—denoted with $D$—in Table IV. The numbers in this table are not seconds or number of conflicts, but rates of benchmarks. Figure 3 and Table III shows the real running times of the bases of the benchmarks. The real running times of Algorithm 2 are much smaller than the running times of Algorithm 1. But if we think about it, we can realize, that these numbers are incomparable. The naive algorithm used total graphs, and a total graph with two thousands nodes has about two millions edges, but a BA random graph has only four thousands. To find the one with the maximal value (Q1) is not the same job, and even the solution of Q2 is easier in the case of associative arrays. We have about five-hundredths many edge, and the running time only one third. To tell the truth Algorithm 2 only uses the conventional Python here—does not use any extension compiled to fast machine code—i.e. overall the execution of the code is much slower.

This programming language enables us to combine different data structures. Hence we can construct one associative array about profits related to a given cluster, and organize these associative arrays into an associative array. This *two level dictionary* is very common in the Python literature. To speed up the solution of Q2, we store one "distance" at two places: in each node's associative array of the according edge. This helps us collect all the edges belonging to the clusters needing to join. This double storage complicates our programs. Moreover, the empty associative arrays generate errors, so we need to check whether the associative arrays become empty, or all the objects are in one cluster.

**Algorithm 3** Contraction by associative array of associative arrays.

```
import UnionFind                                                1
N=1000; q=0.2                                                   2
d = ...                                                         3
uf = UnionFind.UnionFind()                                      4
max_d = max([max(di.values()) for di in d])                     5
while max_d > 0:                                                6
    pairs = [(i,j) for i,di in d.items()                        7
        for j,value in di.items()                               8
        if value == max_d]                                      9
    i,j = pairs[0]                                             10
    del d[i][j]; del d[j][i]                                   11
    di_k = list(d[i].keys())                                   12
    uf.union(i,j)                                              13
    dj = list(d[j].items())                                    14
    for z,value in dj:                                         15
        if z in di_k:                                          16
            d[i][z] += value; d[z][i] += value                 17
            if d[i][z] == 0:                                   18
                del d[i][z]; del d[z][i]                       19
        else:                                                  20
            d[i][z] = value; d[z][i] = value                   21
        del d[z][j]                                            22
        if d[z] == {}:                                         23
            del d[z]                                           24
    del d[j]                                                   25
    if d[i] == {}:                                             26
        del d[i]                                               27
    if len(d) == 1:                                            28
        break                                                  29
    max_d = max([max(di.values())                              30
             for i,di in d.items()])                           31
```

TABLE III
RUNNING TIMES FOR $D^2_r$ IN SECONDS

| N | 0.1 | 0.5 | 0.9 |
|---|---|---|---|
| 100 | 0.004–0.009 | 0.004–0.012 | 0.008–0.015 |
| 500 | 0.076–0.227 | 0.075–0.227 | 0.165–0.327 |
| 1000 | 0.483–1.126 | 0.488–1.267 | 0.863–1.581 |
| 2000 | 2.370–5.867 | 2.397–5.927 | 3.985–8.506 |
| 5000 | 0.230–3.310 | 3.576–25.155 | 15.386–36.270 |
| 10000 | 6.205–23.424 | 15.298–104.581 | 122.131–265.172 |

Here the values of the profits of contractions are distributed among the associative arrays. Hence to find the maximal value we need to search for the maximal value in each associative array, and to select the maximal one among these maximal values (line 5). Lines 7–9 is a list comprehension, in which we traverse all associative arrays, and check, whether there is a key, for which this maximal value is assigned. If yes, we record the pair of the key $i$ of the associative array, and the key $j$ of the maximal value. If all the maximal values have been found, we select one (more precisely the first) key from the keys of the maximal values. This key is a pair, and we refer to its elements by $i$ and $j$ in the following (line 10). We delete the edge belonging to the selected pair (line 11), and we store the keys in the associative array belonging to $i$ (line 12). It is an important step, Python does not allow to modify a data structure while traversing it. (Do not cut off the tree you are standing on!) Similarly, we docket the associative array belonging to $j$ (line 14), and by traversing this associative array we update the weights of the edges in contact to the contracted cluster (lines 16–19). If it is not possible (there is no corresponding $(i, z)$ edge to the edge $(j, z)$) then we construct the missing edge. When all edges $(j, \cdot)$ are processed, we can delete the associative array belonging to $j$ (line 25), and the all references to this cluster (i.e. $(\cdot, j)$ type edges), too (line 22). If the contracted cluster becomes empty (any strange, sometimes it can happen), we need to delete it (lines 26–27). If only one cluster remains, we can stop the method (lines 28–29).

The statistics about Algorithm 3 can be found in the third column—denoted with $D^2$— of Table IV. From this table it is obvious, that the two level dictionary is more effective than the ordinal dictionary.

In Python the associative arrays are implemented as hash tables. It is well known, that here the deletion is a costly operation, and thus we delete the whole structure in small steps. By examining the handling of the hash table in Pythonic way, we rewrote the code in Algorithm 3 in a such way, that at contraction we do not delete directly from the joined cluster, but we create a new associative array for the joined cluster, and fill it with its predecessors. As deletion from dictionary is solved by replacing data with dummy items, the recreation of

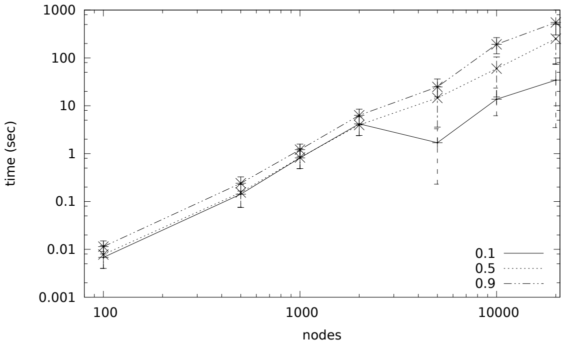


Fig. 3. Running time of implementation $\overline{D_r^2}$ for graphs with different $N$ and $q$

the joined cluster can free up memory and speed up problem Q2. The deletion from the nearby clusters remains the same. It would be very time consuming to recreate them, too. By the fifth column—denoted with $D_r^2$— of Table IV this extra effort has no evident profit: the accuracy is almost the same, but the running time is slightly longer.

## V. Quasi-parallel variant

The hierarchical clustering sometimes has very natural interpretations. For a given distance function $d : V \times V \to \mathbb{R}$, where $d(x, x) = 0$ for any $x \in V$ Sibson defined a *dendogram* function $c : [0, \infty) \to E(V)$ [14], where $E(V)$ is the set of equivalence relations on $V$. This function $c$ fulfils the following criteria:

- If $h \leq h'$ then $c(h) \subseteq c(h')$,
- The final value of $c(h)$ [i.e. $c(\infty)$] is $V \times V$ and
- $c(h + \delta) = c(h)$ for any small $\delta > 0$.

In case of a minimal distance (the distance of clusters defined by the minimum of distances of their elements), let $E = \{(x, y)|x, y \in V, d(x, y) < h\}$, i.e. add to the set of nodes $V$ all the edges shorter than $h$. This is a symmetric relation. Next, take the transitive closure of this relation, which becomes an equivalence relation. Taking these equivalence relations for all $h \geq 0$, we get the dendogram $c$.

Can we use this idea? Unfortunately not. If we have a chain of objects connected by edges, then their optimal clustering produces pairs and maybe one singleton cluster [15]. In the case of a star topology we got a pair and singletons. Therefore the transitive closure of a (partial) tolerance relation is not suitable for us. But we can try to contract independent pairs. As they are independent, the contraction can be done in parallel! This means that we lose the purely greediness of the contraction method. Moreover the dynamics of the method is changing. If we started by contracting two singleton clusters $i$ and $j$, for which there was a third cluster $k$ such that $d_{i,k} > 0$ and $d_{j,k} > 0$, then the next contraction used clusters $i \cup j$ and $k$ at algorithms before. Moreover if there was a fourth cluster $l$, for which $d_{i,l} > 0$, $d_{j,l} > 0$ and $d_{k,l} > 0$, then the following contraction used clusters $i \cup j \cup k$ and $l$, and so on. In other words the initial combined cluster grows in each steps until it is surrounded with clusters only, for which the profit of the contraction with this giant cluster is negative. If we execute the contractions in parallel, then we have more centres (not so giant clusters).

**Algorithm 4** Determining the independent edges

```
def independent(pairs):
    s = {}
    random.shuffle(pairs)
    for i,j in pairs:
        if len({i, j} &
            (s.keys() | s.values())) == 0:
             s[j] = i
    return s
```

We do not need to rewrite the whole Algorithm 2 as most of the code is reusable. For example, to get the independent pairs we have the list of best pairs. Previously we used the first pair from the list, now we will use more. Algorithm 4 selects the independent ones. This algorithm gets all best pairs, and constructs an associative array from the independent edges. The algorithm traverses the best edges, and an edge is independent from the stored ones, if its nodes do not occur in the associative array, neither as a key, nor as a value. The original ordering limits the set of independent edges, so we shuffle the starting list, in order for the subsequent runs to give a different results, therefore the clustering becomes indeterministic, and by repeating the whole process it should be possible to choose the best of them.

The Algorithm 5 is a slight modification of Algorithm 2 so we only provide the difference: you need to replace lines 25–36 in Algorithm 2 with the listing of Algorithm 5. The main difference is that Algorithm 2 which allowed only one pair in the associative array, now allows any number of pairs, which need to be independent.

In this case, the question arises: it is worth to complicate things? Based on the first idea, it could speed up the process, because we can omit the superfluous steps, as we do not need to traverse the associative array several times, it is enough only once, to get the same number of contractions. The second

**Algorithm 5** Quasi-parallel contraction, associative array

```
if len(d) < 2:
    return
max_d = max(d.values())
while max_d > 0:
    pairs = [pair for pair,value
        in d.items() if value ==max_d]
    s = independent(pairs)
    for i,j in s.items():
        uf.union(i,j)
    d2 = contraction(d, s)
    if len(d2) < 2:
        break
    d = d2.copy()
    max_d = max(d.values())
```

TABLE IV
COMPARISON OF RUNNING TIMES AND ACCURACY ON 3/2 BA GRAPHS

| | $D$ | $\overline{D}$ | $D^2$ | $\overline{D^2}$ | $D^2_r$ | $\overline{D^2_r}$ |
|---|---|---|---|---|---|---|
| | | | q=0.1 | | | |
| 100 | 2.44/1.01 | 0.88/1.02 | 1.41/1.04 | 1.03/0.99 | 1.47/1.04 | 1.00/1.00 |
| 500 | 3.68/0.98 | 1.22/0.99 | 1.68/1.01 | 0.96/1.00 | 1.74/1.01 | 1.00/1.00 |
| 1000 | 4.12/0.98 | 1.64/1.00 | 1.50/1.00 | 0.99/1.00 | 1.54/1.00 | 1.00/1.00 |
| 2000 | 4.48/0.98 | 1.67/1.00 | 1.59/1.00 | 1.02/1.00 | 1.58/1.00 | 1.00/1.00 |
| 5000 | 61.79/1.08 | 0.96/1.00 | 16.84/1.32 | 1.03/0.99 | 17.43/1.32 | 1.00/1.00 |
| 10000 | 56.66/1.03 | 0.79/1.00 | 13.93/1.16 | 1.07/1.00 | 14.02/1.16 | 1.00/1.00 |
| | | | q=0.5 | | | |
| 100 | 2.46/0.95 | 0.79/0.98 | 1.28/1.02 | 0.97/0.99 | 1.36/1.03 | 1.00/1.00 |
| 500 | 3.77/0.98 | 1.30/1.00 | 1.63/1.01 | 1.01/1.00 | 1.66/1.01 | 1.00/1.00 |
| 1000 | 4.37/0.97 | 1.61/1.00 | 1.58/1.00 | 1.01/1.00 | 1.70/1.00 | 1.00/1.00 |
| 2000 | 4.26/0.98 | 1.67/1.00 | 1.59/1.00 | 1.02/1.00 | 1.61/1.00 | 1.00/1.00 |
| 5000 | 5.43/0.98 | 1.67/1.00 | 2.07/1.02 | 1.04/1.00 | 2.18/1.02 | 1.00/1.00 |
| 10000 | 4.68/0.99 | 1.49/1.01 | 1.87/1.03 | 1.05/1.01 | 1.99/1.03 | 1.00/1.00 |
| | | | q=0.9 | | | |
| 100 | 1.35/1.05 | 0.88/1.04 | 0.84/1.00 | 1.00/1.00 | 0.88/1.00 | 1.00/1.00 |
| 500 | 1.61/1.00 | 0.95/1.01 | 0.86/1.00 | 0.99/1.00 | 0.92/1.00 | 1.00/1.00 |
| 1000 | 1.85/1.00 | 1.12/1.00 | 0.79/1.00 | 1.01/1.00 | 0.82/1.00 | 1.00/1.00 |
| 2000 | 1.66/1.00 | 1.05/1.00 | 0.69/1.00 | 1.00/1.00 | 0.71/1.00 | 1.00/1.00 |
| 5000 | 2.29/1.00 | 1.18/1.00 | 1.01/1.00 | 0.97/1.00 | 1.08/1.00 | 1.00/1.00 |
| 10000 | 1.53/1.00 | 1.05/1.00 | 0.66/1.00 | 1.09/1.00 | 0.68/1.00 | 1.00/1.00 |

column—denoted by $\overline{D}$—of Table IV shows the numbers of the quasi-parallel variant. It is obvious that this parallel variant is better than the original according to the running time. But if we learned that there is no free lunch, maybe the accuracy of method were worse. Let us see the numbers. By Table IV Algorithm 2 produces less conflict, thus gets closer to the optimum. We measured the whole process at three different values of $q$. We wish to repeat this investigation in more details. We executed the original (Algorithm 2) and the parallel (Algorithm 5) version of the Contraction method on the same signed graphs. The original graph was 3/2 BA graph with thousands of nodes. The weight of the edges were given randomly according the value of $q$ for 101 different $q$. After the contraction method we calculated the cost-functions. Figure 4 shows the result. Moreover we summed the values of the cost functions (calculated the "integral" of the curves), and at the parallel version the sum was 94.8–96.0 percent of the original version (these numbers denote conflicts, so the smaller is better here). Five experiments show similar results, so we believe that this is the tendency. This is not a big difference, but disproves our hypothesis. This fact undermines our beliefs in greedy algorithm, so a careful examination is needed. We think, that the giant cluster is the result of an early decision at non-parallel algorithms. This decision could be perfect or bad. Maybe the latter occurs more often. The parallel version postpones the decision, it executes several contractions in parallel, gets smaller clusters, and maybe causes less vital mistakes.

By examining the Table IV, we can see, that the parallel version really is faster than the original. Surprisingly, at different values of $q$ the speed rate is different. For small $q$ we see big differences, while at big $q$ the parallel version will not be twice as fast as the original.

Let us compare this parallel version ($\overline{D}$) with the variant using an associative array of associative arrays ($D^2$), because until now this was the fastest implementation. If $q$ is small, the parallel version is remarkably faster for big $N$. If $q$ is big, the variant with double dictionary is the clear winner.

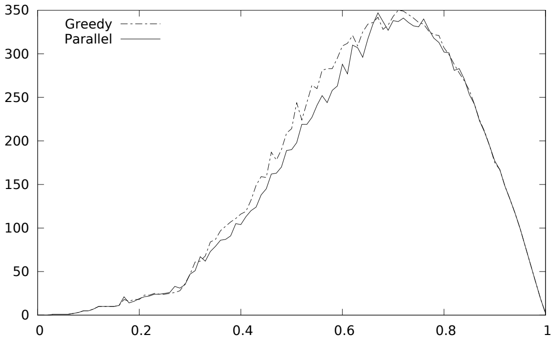


Fig. 4. Cost functions earned by the original and parallel algorithms. $x$-coordinate denotes the rate $q$ of positive edges and $y$-coordinate denotes the value of the cost function.

Of course, the following question arises: It is worth to construct the parallel version of Algorithm 3, or not? The modification is minimal:

- We need to replace line 10 with a cycle for the independent pairs from `pairs`.
- We need to indent lines 11–29, to treat this lines as the core of the cycle in line 10.

We leave these modifications to the reader, and do not present this as a next code listing. Table IV shows the statistics of this parallel variant—denoted by $\overline{D^2}$. If $q$ is small—we only need a small number of contractions to get the near optimal solution—then the second parallel variant ($\overline{D^2}$) is much faster then its origin ($D^2$), but for big $N$s it is slower than the first parallel implementation ($\overline{D}$). Surprisingly if we need a lot of contraction (e.g. $q$ is big), then the parallel version ($\overline{D^2}$) is much slower than its origin ($D$). By observing the accuracy of the last implementation, the tendency is more evident: 84.3–87.7 percent of conflicts of the non-parallel version.

We could continue the comparison of greedy and parallel variants, but we compared the accuracy of the different parallel implementations. The rates of the integrals were 99.9–101.0 percent.

The running times of the three parallel versions are comparable by Table IV. Hence we run the different "parallel" implementations on BA graphs with 20,000 nodes to find the differences. Each implementation solved 15 problems in one hour and a half. The implementation $\overline{D}$ was the slowest, about 8 percent slower than the others. But its accuracy was the best: the difference was less than half percent.

We tried the skip-list data structure for Contraction, where the complexity of the insertion, deletion and search is $O(\log n)$, and not constant. By our statistics this implementation was five times slower than the implementation with double dictionary. The results based on these prototypes suggest for us that the industrial implementation will based on a (double) dictionary, too.

## VI. Future plans

There are tools to find the weakness of codes presented in this paper, and their execution could be optimized. Moreover by choosing a different programming language it gives (maybe several magnitudes) faster implementation. We refer to a quote of D. E. Knuth: *We should forget about small efficiencies, say about 97% of the time: premature optimization is the root of all evil. Yet we should not pass up our opportunities in that critical 3%.* We would like to invent new algorithms and not to patch this ones.

Our main aim was to introduce a simple idea, and its little improvements. We have fulfilled our plans and broke through our former limits of 500 nodes (with significantly higher results), although previously we used faster languages. To solve the correlation clustering problem for bigger and bigger set of objects is challenging for us, so we will continue this path. We believe to go futher we need to use other kind of parallelism: divide and empire.

The zero "distances" of clusters were left along the whole article. We did not examine whether they have any effect to contract clusters where the corresponding distance is zero. This contraction does not decrease and does not increase the number of conflicts *immediately*. It is worth to examine whether this kind of contractions has any future effects, or not. We do not know any examples that could assist this question from neither the natural nor the social sciences.

We can imagine three clusters $i$, $j$ and $k$, where $d_{i,j} = 0$ $d_{i,k} = c$ and $d_{j,k} = -c$. It is obvious, that contracting $i$ and $j$ the value of $d_{i\cup j,k}$ becomes zero, so the number of conflicts does not lessen. While by contracting $i$ and $k$ the number of conflicts decreases by $c$, if $c > 0$. But it is not clear when we have thousands of clusters, the situation is the same, or not. If we have BA random graphs, the zero "distance" clusters, i.e. independent clusters are very common.

## VII. Conclusion

We introduced a correlation clustering problem, and extended the (partial) tolerance relation to clusters. Using this concept we have shown the Contraction method, and its several implementation in Python. Despite of the weakness of this programming language these implementations gave fast results for big sets, although the problem is NP-hard. By our knowledge these are the state of the art algorithms in correlation clustering. We found a near-optimal solution for a problem where the upper bound of the number of possible partitions is $10^{64,079}$ [16].

Our previous measurements show, that the accuracy of this method is among the best optimization methods [11]. These two properties enable the usage of this method in real-life applications.

## References

[1] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning*, vol. 56, no. 1-3, pp. 89–113, 2004. doi: 10.1023/B:MACH.0000033116.57574.95. [Online]. Available: http://dx.doi.org/10.1023/B:MACH.0000033116.57574.95

[2] C. Zahn, Jr, "Approximating symmetric relations by equivalence relations," *Journal of the Society for Industrial & Applied Mathematics*, vol. 12, no. 4, pp. 840–847, 1964. doi: 10.1137/0112071. [Online]. Available: http://dx.doi.org/10.1137/0112071

[3] S. Kim, S. Nowozin, P. Kohli, and C. D. Yoo, "Higher-order correlation clustering for image segmentation," in *Advances in Neural Information Processing Systems*, 2011. doi: 10.1.1.229.4144 pp. 1530–1538.

[4] A. Bhattacharya and R. K. De, "Divisive correlation clustering algorithm (dcca) for grouping of genes: detecting varying patterns in expression profiles," *bioinformatics*, vol. 24, no. 11, pp. 1359–1366, 2008. doi: 10.1093/bioinformatics/btn133. [Online]. Available: dx.doi.org/10.1093/bioinformatics/btn133

[5] B. Yang, W. K. Cheung, and J. Liu, "Community mining from signed social networks," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 10, pp. 1333–1348, 2007.

[6] T. DuBois, J. Golbeck, J. Kleint, and A. Srinivasan, "Improving recommendation accuracy by clustering social networks with trust," *Recommender Systems & the Social Web*, vol. 532, pp. 1–8, 2009. doi: 10.1145/2661829.2662085. [Online]. Available: http://dx.doi.org/10.1145/2661829.2662085

[7] Z. Chen, S. Yang, L. Li, and Z. Xie, "A clustering approximation mechanism based on data spatial correlation in wireless sensor networks," in *Wireless Telecommunications Symposium (WTS), 2010*. IEEE, 2010. doi: 10.1109/WTS.2010.5479626 pp. 1–7. [Online]. Available: http://dx.doi.org/10.1109/WTS.2010.5479626

[8] Z. Néda, R. Florian, M. Ravasz, A. Libál, and G. Györgyi, "Phase transition in an optimal clusterization model," *Physica A: Statistical Mechanics and its Applications*, vol. 362, no. 2, pp. 357–368, 2006. doi: 10.1016/j.physa.2005.08.008. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2005.08.008

[9] L. Aszalós and T. Mihálydeák, "Rough clustering generated by correlation clustering," in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Springer Berlin Heidelberg, 2013, pp. 315–324. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2007.1061

[10] ——, "Rough classification based on correlation clustering," in *Rough Sets and Knowledge Technology*. Springer, 2014, pp. 399–410. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11740-9_37

[11] L. Aszalós and M. Bakó, "Advanced search methods (in Hungarian)," http://morse.inf.unideb.hu/~aszalos/diak/fka, 2012.

[12] R. Durrett, R. Durrett, and R. Durrett, *Random graph dynamics*. Cambridge university press Cambridge, 2007, vol. 200, no. 7.

[13] L. Aszalós, J. Kormos, and D. Nagy, "Conjectures on phase transition at correlation clustering of random graphs," *Annales Univ. Sci. Budapest., Sect. Comp*, no. 42, pp. 37–54, 2014.

[14] R. Sibson, "Slink: an optimally efficient algorithm for the single-link cluster method," *The Computer Journal*, vol. 16, no. 1, pp. 30–34, 1973. doi: 10.1093/comjnl/16.1.30. [Online]. Available: http://dx.doi.org/10.1093/comjnl/16.1.30

[15] D. Nagy, “Correlation clustering of trees,” Master’s thesis, University of Debrecen, Faculty of Informatics, Hungary, 2015. Available: http://hdl.handle.net/2437/211878

[16] D. Berend and T. Tassa, “Improved bounds on bell numbers and on moments of sums of random variables,” *Probability and Mathematical Statistics*, vol. 30, no. 2, pp. 185–205, 2010.

DOI: 10.15439/2015F139
ACSIS, Vol. 5

# Performance Research and Optimization on CPython's Interpreter

Huaxiong Cao, Naijie Gu[1], Kaixin Ren, and Yi Li
1) Department of Computer Science and Technology, University of Science and Technology of China
2) Anhui Province Key Laboratory of Computing and Communication Software
3) Institute of Advanced Technology, University of Science and Technology of China
Hefei, China, 230027
Email: chx319@mail.ustc.edu.cn, gunj@ustc.edu.cn

*Abstract*—**In this paper, the performance research on CPython's latest interpreter is presented, concluding that bytecode dispatching takes about 25 percent of total execution time on average. Based on this observation, a novel bytecode dispatching mechanism is proposed to reduce the time spent on this phase to a minimum. With this mechanism, the blocks associated with each kind of bytecodes are rewritten in hand-tuned assembly, their opcodes are renumbered, and their memory spaces are rescheduled. With these preparations, this new bytecode dispatching mechanism replaces the time-consuming memory reading operations with rapid operations on registers.**

**This mechanism is implemented in CPython-3.3.0. Experiments on lots of benchmarks demonstrate its correctness and efficiency. The comparison between original CPython and optimized CPython shows that this new mechanism achieves about 8.5 percent performance improvement on average. For some particular benchmarks, the maximum improvement is up to 18 percentages.**

## I. Introduction

The past decade has witnessed the widespread use of Python, which is a typical dynamic language designed to execute on virtual machines. Several software engineering advantages over statically compiled binaries, including portable program representations, thread manage -ment, some safety guarantees, built-in automatic memory, and dynamic program composition through dynamic class loading, are provided by Python. These advanced features enhance the user programming model and drive the success of Python language. However, these features also require the dynamic compilers to do quite a lot of extra operations, such as type checking, wrapping/unwrapping of boxed values, virtual method dispatching, bytecode dispatching, etc. These extra operations usually make Python programs several or dozens of times slower than static programs achieving the same functionality. Moreover, traditional static program optimiza -tion technologies are frustrated, introducing new challenges for achieving high performance.

In response, more and more researchers have paid their attention to optimize dynamic language compilers. The technologies proposed aim to improve performance by monitoring programs' behavior and using this information to drive optimization decisions [1]. The dominant concepts that have influenced effective optimization technologies in today's virtual machines include JIT compilers, interpreters, and their integrations.

JIT (just-in-time) techniques exploit the well-known fact that large scale programs usually spend the majority of time on a small fraction of the code [2]. During the execution of interpreters, they record the bytecode blocks which have been executed more than a specified number of times, and cache the binary code associated to these blocks. The next time these bytecode blocks are executed, it has no need to interpreter these bytecodes once again, just jumps to the cached binary code, and continues the execution. By this way, much interpreting work is omitted, and performance improvement can be achieved. However, since the threshold is quite large in general, the interpreting stage still accounts for a large proportion of total execution time. JIT strategies work well for *for/while* blocks which likely exist in science compute field. For the programs, whose purposes are remote configuration, warning, tracing, statistics, communication, etc, it is very hard to find hot blocks. Though not large, these programs are usually executed quite frequently.

Interpreters have a series of advantages which make them attractive [3]. Firstly, optimizing interpreters reduces the uptime of all Python programs with or without hot blocks. Secondly, they are quite simpler to construct than JIT compilers, making them quicker, more reliable to construct and easier to maintain. Thirdly, interpreters require less memory than JIT compilers, for both the interpreted virtual machine code and the interpreter itself. Interpreters for dynamic language have two main structures [1]: switch-based structure and threaded-based structure, whose details are given in section 2 part B. The latter is the latest and most efficient mechanism. Recently, many researchers have applied dynamic techniques to improve the performance of threaded-based structure. Piumarta and Riccardi [4] describe their techniques to dynamically generate threaded codes for purpose of eliminating a central dispatch site and inlining common bytecode sequences. Ertl and Gregg [5] extend Piumarta and Riccardi's work by duplicating bytecode sequences, researching various interpreter generation heuristics,

[1] Corresponding author. Email: gunj@ustc.edu.cn

978-83-60810-66-8/$25.00©2015, IEEE

and concentrating on improving branch prediction accuracy. Gagnon and Hendren [6] adapt Piumarta and Riccardi's research to work in the context of multithreading and dynamic class loading. Sullivan et al. [7] describe a combination between an interpreter implementation and a dynamic binary optimizer, which enhances the efficacy of the underlying dynamic binary optimizer during the execution of interpreter.

Despite the enhancements above, interpreters are still worse than static compilers from the runtime perspective. The purpose of our research is to explore new ways to achieve further performance improvement. In this paper, the performance research on CPython's latest interpreter is presented, finding that bytecode dispatching takes about 25 percent of total execution time on average. Then, a novel bytecode dispatching mechanism, which aims at reducing memory reading operations during bytecode dispatching and reducing the time spent on this phase to a minimum, is proposed. This mechanism is implemented inside CPython's interpreter. Experiments on lots of benchmarks demonstrate the correctness and efficiency of our new mechanism. The comparisons between original CPython and optimized CPython show that our new mechanism can achieve about 8.5 percent performance improvement on average. For some particular benchmarks, the maximum improvement is up to 18 percentages.

The remainder of this paper is structured as follows: In the next section, we make comparisons among different compilers of Python and present both switch-based and threaded-based mechanisms. Section 3 reports our performance research on CPython's interpreter. Section 4 discusses the main techniques we adopted to construct a new bytecode dispatching mechanism. Benchmarks and Experiments are given in section 5. Section 6 discusses the related work. We conclude this paper in the last section.

## II. Background

### *A. CPython VS Other Python compilers*

A series of dynamic compilers are designed to run Python programs, such as CPython, Jython [8], IronPython [9], Pyston [10], PyPy [11], etc. The comparisons among them are shown on Table I, and these benchmarks are provided officially by Pyston/minibenchmarks and Pyston/microbenchmarks. Among these compilers, CPython is the official and standard compiler for Python language, it can support all the grammars and extensions. Others are developed for special applications. They focus on particular scenarios and take in-depth optimization passes. In this context, these compilers show their excellent performances for some benchmarks, but for other benchmarks, they may be several times slower than CPython. As shown on Table I, taking fid.py for example, it takes CPython 3.696 seconds to execute this script, while IronPython, Pyston and PyPy spend less than 2 seconds on the same script. However, it takes CPython 1.082 seconds to execute emwomding.py, while JPython, IronPython, Pyston and PyPy spend more than 3.700 seconds on the execution of emwomding.py.

What's more, some benchmarks, like pydigits.py, empth_lo -op.py, vecf_add.py, nq.py, raytrace.py, cannot be executed successfully by some of these compilers. Based on these comparisons, our research focuses on CPython, and has not only practical application value, but also instruction meaning for the improvement of other interpreters.

### *B. Switch-based Mechanism VS Thread-based Mechanism*

The performance of interpreters depends heavily on their bytecode dispatching mechanisms. CPython's interpreter provides two main bytecode dispatching mechanisms: switch-based mechanism and thread-based mechanism.

The inner loop of switch-based interpreters is quite simple: jump to the dispatch point, fetch the next bytecode and dispatch to its implementation through a switch statement. Its typical framework is shown in Fig. 1.

As shown in Fig. 1, the interpreter is an infinite loop with a big switch block to dispatch bytecodes successively. Each bytecode are implemented by a particular case in the body of this switch block. At the end of each case, control are passed back to the beginning of the infinite loop by breaking out of this switch block. Traditional C compilers like GCC translate this switch block into a series of comparison statements. Considering a particular bytecode whose opcode is BINARY_SUBTRACT, five indispensable comparisons should be executed before reaching its associated block. Assuming that all bytecodes have the same probability of occurrence, it takes N/2 (N is the total kinds of bytecodes) comparisons on average to find the corresponding block. As to

TABLE I.
COMPARISONS AMONG CPYTHON, JYTHON, IRONPYTHON, PYSTON AND PYPY

| Benchmarks | CPython-3.3.0 | Jython | IronPython | Pyston | PyPy |
|---|---|---|---|---|---|
| empty_loop.py | 3.532s | 5.491s | **Failed** | 19.248s | 0.248s |
| pydigits.py | 0.034s | 2.126s | 1.431s | **Failed** | 0.039s |
| fid.py | 3.696s | 4.776s | 1.527s | 0.636s | 0.864s |
| vecf_add.py | 9.890s | 12.970s | 25.150s | **Failed** | 0.059s |
| allgroup.py | 0.836s | 4.428s | 3.052s | **Failed** | **18.804s** |
| chaso.py | 26.268s | 33.091s | 51.366s | **Failed** | 1.392s |
| go.py | 53.787s | 56.746s | 123.404s | **Failed** | 33.638s |
| nbody.py | 12.677s | 19.535s | 17.057s | 25.540s | 1.470s |
| nq.py | 29.879s | **Failed** | 35.673s | **Failed** | **44.418s** |
| raytrace.py | 11.608s | 16.418s | 26.724s | **Failed** | 1.228s |
| polymorphism.py | 4.358s | 7.228s | 7.959s | 4.390s | **14.260s** |
| unwinding.py | 1.082s | 3.757s | 2.802s | 93.180s | 4.481s |

CPython, N is 101. So, performing bytecode dispatching wastes the majority of execution time and it is very inefficient.

Threaded-based mechanism is the latest and most efficient mechanism for interpreters, popularized by the Forth programming language [12]. There are many kinds of threaded-based interpreters, and direct threading is regarded as the most efficient one. Direct threading mechanism improves performance by eliminating redundant comparisons. In addition, rather than returning to a central dispatch point, the implementation of each direct threading opcode ends with the particular code required to dispatch the next opcode. This optimization eliminates the centralized dispatch, removing lots of jump instructions. The framework of direct threading mechanism is shown in Fig. 2.

As shown in Fig. 2, execution starts with fetching the address of the very first bytecode's implementation and then jumping to that address. For each bytecode, it performs its own work at first, then increases the instruction pointer, thirdly fetches the address of next bytecode's implementation from memory, and jumps to the target address to handle successive bytecode. The native instructions associated with bytecode dispatching is shown in Fig. 3, with one stack reading, one memory reading and one jump. It can be seen from Fig. 3 that the bytecode dispatching overheads associated with direct threading mechanism are quite lower than those associated with switch-based mechanism.

## III. Motivation: Performance Research

So far, it can be seen that bytecode dispatching plays a very important role in the performance of CPython. To make this concept intuitional, a series of meticulous experiments are conducted and the experimental results are shown in table II.

To make the results credible and accurate, this work utilizes hardware performance counters provided by Intel, and calculates the number of ticks which are consumed by the procedure to find next bytecode. The results are shown in Table II column 2. The ticks, which are consumed by the total program execution are also recorded and listed in Table II column 3. In Table II column 4, the ratios between them are calculated and listed. According to this table, it can be sure that the time consumed by bytecode dispatching is 25 percent of the total execution time on average. So if this stage could be further optimized, the total performance of CPython will be improved obviously.

As shown in Fig. 3, the bytecode dispatching in direct threading mechanism is composed of one stack reading, one jump and one memory reading. Inside the memory reading, there are an addition and a multiplication. Since operations on stack are quite frequent, the first stack reading instruction will also hit the L1 cache. Table III [13] lists the time needed to read data from register, L1 cache, L2 cache, memory and disc. According to this table, the first instruction takes about 2ns. The second instruction contains an addition operation, a multiplication operation and a load operation. The first two operations can be finished within several ticks. However, since L1 cache can only contain eight pages (32k L1 cache, page size 4k), reading the address of successive bytecode's implementation leads to lots of L1 cache misses. In this context, it has to get the right value from L2 cache, or even memory, which may take dozens of nanoseconds. Assuming that reading from these three memory structures has the same probability of occurrence, it takes about 30ns to finish this memory reading. So, it can be seen that the second instruction takes the majority of the time spent on bytecode dispatching, and optimizing this instruction will contribute a lot to bytecode dispatching.

```
Compiled code:
    Unsigned char code[] = { …
        LOAD_FAST,
        LOAD_CONST,
        BINARY_ADD,
        BINARY_SUBTRACT, … };
Bytecode implementations:
For( ; ; ) {
    insn = get_next_insn(insn);
    opcode = get_opcode(insn);
    switch(opcode)  {
        case NOP: …; break;
        case LOAD_FAST: …; break;
        case LOAD_CONST: …; break;
        case BINARY_ADD: …; break;
        case BINARY_SUBTRACT: …;
break;
        ….
    }
}
```

Fig. 1 The framework of switch structure.

```
compiled code:
void * table[] = { …
    ##LOAD_FAST,
    ##LOAD_CONST,
    ##BINARY_ADD,
    ##BINARY_SUBTRACT, … };
bytecode implementations:
….;
LOAD_FAST:
    …;
    cpcode = get_opcode(insn_next);
    goto *table[opcode];
LOAD_CONST:
    …;
    opcode = get_opcode(insn_next);
    goto *table[opcode];
BINARY_ADD:
    …;
    opcode = get_opcode(insn_next);
    goto *table[opcode];
BINARY_SUBTRACT:
    …;
    opcode = get_opcode(insn_next);
    goto *table[opcode];
….
```

Fig. 2 The framework of direct threading mechanism.

```
//read stack and get the opcode of next
//bytecode
mov    -0x1c4(%ebp),%ebx
//read memory and get the address
//related to next bytecode
mov    0x8220000(,%ebx,4),%eax
//jump to deal with next bytecode
jmp    *%eax
```

Fig. 3 The native instructions associated with opcode dispatch.

TABLE II
TIME SPENT ON BYTECODE DISPATCHING

| benchmarks | Dispatching ticks | Total ticks | Ratio |
|---|---|---|---|
| Queen.py | 124356740795 | 560761937239 | 22.176388% |
| test_pow.py | 66430602 | 605265570 | 10.975447% |
| diff.py | 11349007012 | 42445898103 | 26.737583% |
| test_sqrt.py | 209325165518 | 823381752732 | 25.422613% |
| test_image.py | 121776483 | 633664878 | 19.217801% |
| iterator.py | 218799781767 | 847798974988 | 25.807980% |
| generator.py | 987262878658 | 3744968182830 | 26.362383% |
| range.py | 1238431745418 | 4568662416948 | 27.107098% |
| while.py | 1810252155978 | 6454065479106 | 28.048246% |
| average | 511107298026 | 1.8937026e+12 | 26.989839% |

TABLE III
THE TIME TOKEN TO READ DATA FROM DIFFERENT STORAGES

| | Reg<br>-ister | L1<br>cache | L2<br>cache | Mem<br>-ory | disc |
|---|---|---|---|---|---|
| Time<br>(ns) | 0.5 | 2 | 10~20 | 50~100 | 25~50 |

## IV. THE NEW DISPATCHING MECHANISM

According to the above section, bytecode dispatching spends most of the time on memory reading. To optimize bytecode dispatching procedure, a new bytecode dispatching mechanism is proposed and the framework of CPython's interpreter is reconstructed, drawing the inspiration from [14]. Our new techniques proceed in three phases and their functionalities are described below.

Phase 1: Rewriting and statistics. This phase rewrites the statements associated with each kind of bytecodes in hand-tuned assembly, and calculates the length of the final binary code of each case. The results are shown at Table IV. According to this table, there are 19 kinds of bytecodes which own less than 64 byte binary code, 51 kinds of bytecodes which own less than 128 byte but more than 64 byte binary code, 22 kinds of bytecodes which own more than 128 byte but less than 256 byte binary code, and 9 kinds of bytecodes which own more than 256 byte binary code.

TABLE IV
STATISTICS ON LENGTH OF BYTECODES' FINAL BINARY CODE

| length | [0,64] | (65,128] | (128,256] | (256,512] |
|---|---|---|---|---|
| number | 19 | 51 | 22 | 9 |

Phase 2: calculation. This phase calculates the proper size of each memory unit, named BSIZE. Inside the optimized interpreters, the memory allocated to each kind of bytecodes is an integral multiple of BSIZE bytes. Mark *binary[i]* as the length of binary code associated with *i*th kind of bytecodes, the BSIZE should be the minimum positive number which conforms to condition (1) and condition (2).

$$\exists i : 2^i = BSIZE \tag{1}$$

$$\sum_{i=0}^{100} \left\lceil \frac{binary[i]}{BSIZE} \right\rceil \leq 256 \tag{2}$$

The first constraint assures the calculation of next bytecode's address is simplified to a quick left shift. The second constraint assures that the maximum opcode of bytecodes isn't bigger than the threshold value (256) defined by CPython. If the BSIZE is too big, there will be a lot of NOP instructions and the executable file will be quite big, causing damage to the performance. That's why BSIZE should be set as small as possible. According to Table IV, the proper value of BSIZE is 128.

Phase 3: opcode redefinition. This phase redefines the opcodes of bytecodes, with their order stay the same. Let *opcode[i]* be the new opcode of *i*th kind of bytecode, and the algorithm used here is shown as follows:

$$\begin{aligned} &opcode[0] = 0 \\ &opcode[i] = opcode[i-1] + \left\lceil \frac{binary[i-1]}{BSIZE} \right\rceil, 1 \leq i \leq 100 \end{aligned} \tag{3}$$

Taking the first three bytecode (POP_TOP, ROT_TWO and ROT_THREE) as an example, their original opcodes are 1, 2 and 3, respectively. Assuming the first bytecode has 200 byte binary code, the second bytecode has 350 byte binary code, and the third bytecode has 100 byte binary code, their final opcodes will be 1, 3 and 6. The memory allocation of the interpreter with new dispatching mechanism is shown in Fig. 4.

As shown in Fig. 4, POP_TOP's binary code is stored in the first grid region (from top to bottom), ROT_TWO's binary code is stored in the second grid region, while the ROT_THREE's binary code is stored in the third grid region. There are gaps between neighboring kinds of bytecodes. In addition, the framework of the new interpreter of CPython is shown in Fig. 5. Inside this new interpreter, the binary code of all kinds of bytecodes is arranged in numerical order and each of them occupies several BSIZE memory spaces. So, every time CPython jumps to the implementation associated to next bytecode, it just need to execute this simple statement: "goto (base + BSIZE*opcode)". Fig. 6 shows the native instructions associated with this statement, including one stack reading, one left shift, one addition and one jump. The middle two instructions are operations on registers and can be finish within

2 ticks (0.8ns). Hence, it just takes 2.8ns to get the address of next bytecode's implementation, instead of 30ns. Since NEXTOP operation is executed so many times, this new structure will bring a lot of performance promotion.

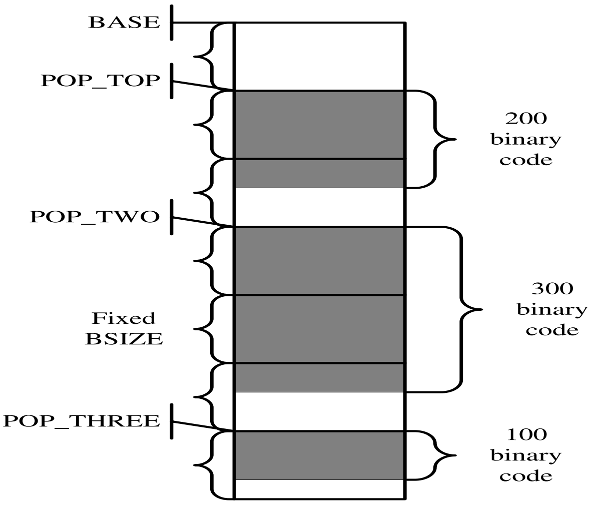


Fig. 4 The memory allocation of post-construction interpreter.

```
bytecode implementations:
BASE:
….;
asm(“.balign BSIZE \n\t”);
LOAD_FAST:
    …;
    opcode = get_opcode(insn_next);
    goto (BASE+opcode* BSIZE);
asm(“.balign BSIZE \n\t”);
LOAD_CONST:
    …;
    opcode = get_opcode(insn_next);
    goto (BASE+opcode* BSIZE);
asm(“.balign BSIZE \n\t”);
BINARY_ADD:
    …;
    opcode = get_opcode(insn_next);
    goto (BASE+opcode* BSIZE);
asm(“.balign BSIZE \n\t”);
BINARY_SUBTRACT:
    …;
    opcode = get_opcode(insn_next);
    goto (BASE+opcode* BSIZE);
….
```

Fig. 5 The framework of interpreter with new bytecode dispatching.

```
//read stack and get the opcode of next
//bytecode
mov    -0x1c4(%ebp),%eax
shl      $8, %eax  //left shift
//BASE is an immediate value
addl    $BASE, %eax
//jump to deal with next bytecode
jmp    *%eax
```

Fig. 6 The native instructions associated with new bytecode dispatching.

## V. Experimental Evaluation

This new bytecode dispatching mechanism has been implemented under Ubuntu-12.04 on Intel(R) Core(TM) 2 CPU E6550 2.33GHz with 2 processors, 2G memory, 32k L1 cache and 4M L2 cache. When assess the optimized CPython, about 35 benchmarks are gathered from CPython-3.3.0 and Pyston-0.2. Two criterions are considered here: (1) correctness (optimized CPython can provide the same functionality as original one.), and (2) efficiency (optimized CPython can execute benchmarks faster than original one).

### A. Correctness

Taking benchmark *diff.py* as an example, file sysmodule.c and file _testcapimodule.c are chosen from CPython's source files randomly and are used as two parameters of *diff.py*. Then, *diff.py* is executed by optimized CPython and original CPython separately, and two result files are produced. Later, Linux command *diff* are used to compare these two result files, concluding that there is no difference between them.

In addition, the benchmarks, which are listed in Table I and cannot be executed successfully by Jython, IronPython, Pyston or PyPy, can be executed successfully by optimized CPython. Actually, optimized CPython can execute all these 35 benchmarks and the time spent on these benchmarks is shown in Fig. 7.

### B. Efficiency

Optimized CPython and original CPython are compiled with the same parameters, and both of them are used to execute these benchmarks separately. Each benchmark is executed one thousand times to reduce volatility, and the final results are shown in Fig. 7. As to some benchmarks taking too little time to run, we recode the time spent on their several times running. Taking *image.py* for example, it just takes original CPython 0.046 seconds to run this benchmark. So, running *image.py*100* takes 4.6 seconds. Similarly, running *queen.py/10* takes 10.23 seconds, for the reason that it takes CPython 102.3 seconds to finish the queen.py. It can be seen from Fig. 7 that all of these benchmarks achieve performance improvement with our new bytecode dispatching mechanism. The average performance improvement is about 8.5%. In particularly, benchmark *image.py* achieves up to 18% performance improvement.

The performance improvement happens for two main reasons. Firstly, this new interpreter replaces slow memory reading operations with quick operations on registers, and reduces the time spent on bytecode dispatching. Secondly, less memory reading operations cut down on the cache misses, especially for low specification machines. Perf [15] is used to measure L1 data cache misses for part of these benchmarks and the results are shown at Table V. Table V column 2 lists the cache misses reported by original CPython when it is used to execute these benchmarks, while Table V column 3 lists the cache misses reported by optimized CPython when it is used to do the same jobs. The last column in Table V shows that about 14.64 percent of cache misses are left out on average. The more memory reading operations it reduces, the greater chance that

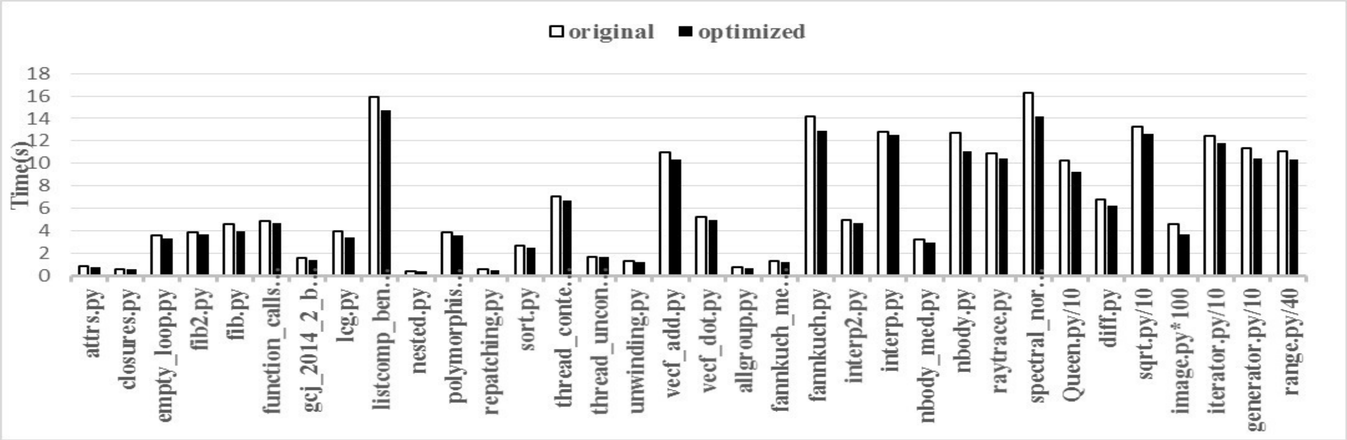


Fig. 7 Ratio of  benchmarks before and after optimization.

TABLE V
STATISTICS ON L1 DATA CACHE MISSES

| Benchmarks | ori-CPython | opti-CPython | 1-(opti/ori) |
|---|---|---|---|
| queen.py | 64M | 43M | 32.8% |
| pow.py | 3.93M | 3.54M | 9.9% |
| diff.py | 36.9M | 33.1M | 10.3% |
| sqrt.py | 3.46M | 2.56M | 26.0% |
| image.py | 1.37M | 1.14M | 16.8% |
| iterator.py | 83.5M | 80.7M | 3.35% |
| generator.py | 170M | 148M | 12.9% |
| range.py | 292M | 270M | 7.53% |
| while.py | 295M | 259M | 12.2% |
| average | - | - | 14.64% |

Note: “ori-CPython” stands for “original CPython”, “opti-CPython” stands for “optimized CPython”, and “1-(opti/ori)” stands for “1 - (optimized CPython / original CPython)”.

remaining memory reading operations hit the cache. In another word, the remaining memory reading operations can be finished in less time.

## VI. RELATED WORK

There are a large quantity of recent papers researching interpreter performance. Romer et al. [16] have reported the performance characteristics of some interpreters. Later, Ertl and Gregg [3] investigated the performance of recent efficient interpreters. Both of these two studies have found that almost every interpreters perform an exceptionally high number of indirect branches. Since most of indirect branches are caused by bytecode dispatching, their conclusion is consistent with this performance research reporting in section 3. Our research aims at reducing the time spent on the second instruction in Fig. 3 to a minimum, while their work aims at optimizing the third instruction in Fig. 3 and reducing indirect branches mispredictions. This is the main difference between us.

Several techniques are used to reduce indirect branches mispredictions. J Hoogerbrugge and L Augusteijn [17], [18] have proposed that software pipelining interpreters is a way to reduce dispatch branch cost on architectures with split indirect branches. In addition, Subroutine threading [19] has also been proposed to avoid the overheads of indirect branches in intrepreter implementations. Each bytecode is implemented with a particular C function. Instead of dispatching or interpreting bytecode, a simple JIT compiler generates executable code for a sequence of calls to these functions. This method can eliminate indirect branches at the cost of sacrificing both simplicity and portability.

Cache misses have a significant impact on the program performance [20]. Brunthaler [21] have proposed a formalization of interpreter opcode ordering (bytecode scheduling) for an interpreter with an extended opcode set, and concluded that high cache miss ratio is another bottleneck of interpreters. A lot of techniques, like feedback-guided technique [22], profile-guided technique [23], etc, are conducted to achieve better orderings, improving code locality and reducing cache misses. Jason McCandless and his co-worker [24] implement a metaheuristic (Monte Carlo) to generate better orderings, achieving considerable performance improvement. As shown in section 5, cache misses can also be reduced by our new mechanism, making the new interpreter perform better.

Another method which is widely used is combining sequences of VM instructions into super-instructions [25]. This technique focuses on reducing the number of bytecode dispatches, and has two variants: static super-instructions and dynamic super-instructions. The comparison between these two variants are shown in [5].

As far as we know, the nearest research to our work is [14]. The main difference is that their research plans for each bytecode a fixed-size block of instructions. In such a case, bytecodes with short instruction implementation will incur a lot of NOP instructions, which increases the code size of the interpreter dispatch loop and reduce cache hit ratio. In addition, bytecodes with long instruction implementation will lead to lots of extra jump instructions. Relatively speaking, our mechanism is more flexible and efficient than that, with much less NOP instructions and no extra jump.

## VII. CONCLUSIONS

In this paper, the performance research on CPython’s interpreter is carried out, figuring out that bytecode dispatching has a big influence on interpreters. Then, a novel bytecode dispatching mechanism is designed, aiming at removing memory reading operations during bytecode dispatching and reducing the time spent on this phase to a minimum. The final binary code of each kind of bytecodes is arranged in numerical order and each of them occupies several BSIZE memory spaces.

This novel mechanism is implemented, and its correctness and efficiency are demonstrated by a large number of benchmarks. Comparisons are made between original CPython and optimized CPython to show that the new mechanism achieves about 8.5 percent performance improvement on average. For some particular benchmarks, the maximum improvement is up to 18 percentages. This performance improvement happens for two main reasons: lesser memory reading operations and lesser cache misses. The novel mechanism proposed here can also be adopted by other interpreters, and will contribute to their performance improvement.

## REFERENCES

[1] M. Arnold, S. J. Fink, D. Grove, M. Hind, and P. F. Sweeney, "A survey of adaptive optimization in virtual machines," *Proc. of IEEE,* vol. *93*, no. *2*, pp. *449-466*, Feb., 2005. http://dx.doi.org/10.1109/jproc.2004.840305

[2] D. E. Knuth, "An empirical study of FORTRAN programs," *Softw.: Practice and experience,* vol. *1*, no. *2*, pp. *105-133*, Jun., 1971. http://dx.doi.org/10.1002/spe.4380010203

[3] M. A. Ertl, and D. Gregg, "The structure and performance of efficient interpreters," *JILP,* vol. *5*, pp. *1-25*, Mar., 2003.

[4] I. Piumarta, and F. Riccardi, "Optimizing direct threaded code by selective inlining," *ACM Sigplan Not.,* vol. *33*, no. *5*, pp. *291-300*, May, 1998. http://dx.doi.org/10.1145/277652.277743

[5] M. A. Ertl, and D. Gregg, "Optimizing indirect branch prediction accuracy in virtual machine interpreters," *ACM Sigplan Not.,* vol. *38*, no. *5*, pp. *278-288*, May, 2003. http://dx.doi.org/10.1145/780822.781162

[6] E. Gagnon, and L. Hendren, "Effective inline-threaded interpretation of Java bytecode using preparation sequences," in *Compiler Construction*, G. Goos, J. Hartmanis and J. v. Leeuwen, Ed., Heidelberg, DE: Springer, 2003, pp.170-184. http://dx.doi.org/10.1007/3-540-36579-6_13

[7] G. T. Sullivan, D. L. Bruening, I. Baron, T. Garnett, and S. Amarasinghe, "Dynamic native optimization of interpreters," in *Proc. 2003 workshop on Interpreters, virtual machines and emulators*, New York, 2003, pp. *50-57*. http://dx.doi.org/10.1145/858570.858576

[8] R. W. Bill, *Jython for Java programmers*, 1st. ed., USA: SAMS, 2001. ISBN 978-0735711112. http://file182.cordpdf.org/1juv4l_jython-for-java-programmers.pdf

[9] J. Hugunin, "IronPython: A fast Python implementation for .NET and Mono," in *PyCon.*, Washington, USA, 2004.

[10] M. L. Hetland, "Pedal to the Metal: Accelerating Python," in *Python Algorithms*, Trondheim, NO: Springer, 2014, pp.255-258. http://dx.doi.org/10.1007/978-1-4842-0055-1_12

[11] Pypy. (2006). PyPy is a fast, compliant alternative implementation of the Python language. [Online]. Available: http://pypy.org/.

[12] Forth-language for interactive computing. (1970). an technical report on Forth. [Online]. Available: http://mx1.1strecon.org/downloads/Forth_Resources/CM_ForthLanguageInteractiveComputing_1970.pdf.

[13] J. L. Hennessy, and D. A. Patterson, *Computer architecture: a quantitative approach*, Waltham, GB: Elsevier, 2012. ISBN 978-0-12-383872-8. http://www.cpp.edu/~kding/materials/Computer%20Architecture%20A%20Quantitative%20Approach%20(5th%20edition).pdf

[14] Y. Ye, C.-Q. Li, and J.-S. Hu, "Transplantation and Optimization of Dalvik Virtual Machine Based on CK610," *Comput. Eng.*, vol. *16*, pp. *100*, 2011. doi:10.3969/j.issn.1000-3428.2011.16.098

[15] A. Melo, "The new linux' perf' tools," in *17th Int. Linux Sys. Tech. Conf.*, Nuremberg, GE, 2010.

[16] T. H. Romer, D. Lee, G. M. Voelker, A. Wolman, W. A. Wong, J.-L. Baer, B. N. Bershad, and H. M. Levy, "The structure and performance of interpreters," *ACM Sigplan Not.,* vol. *31*, no. *9*, pp. *150-159*, 1996. http://dx.doi.org/10.1145/248209.237175

[17] J. Hoogerbrugge, and L. Augusteijn, "Pipelined Java Virtual Machine Interpreters," in *Compiler Construction*, Vol. 1781, D. A. Watt, Ed., Heidelberg, GE: Springer, 2000, pp.35-49. http://dx.doi.org/10.1007/3-540-46423-9_3

[18] J. Hoogerbrugge, L. Augusteijn, J. Trum, and R. v. d. Wiel, "A code compression system based on pipelined interpreters," *Softw.: Practice and Experience,* vol. *29*, no. *11*, pp. *1005-23*, 1999. http://dx.doi.org/10.1002/(sici)1097-024x(199909)29:11<1005::aid-spe270>3.0.co;2-f

[19] M. Berndl, B. Vitale, M. Zaleski, and A. D. Brown, "Context threading: A flexible and efficient dispatch technique for virtual machine interpreters," in *Proc. Int. Symp. Code Gen. Optim.*, Washington, USA, 2005, pp. *15-26*. http://dx.doi.org/10.1109/cgo.2005.14

[20] Ristov, and Sasko, "Performance impact of reconfigurable L1 cache on GPU devices," *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on,* Kraków, Poland, IEEE, 2013, pp. 507-510.

[21] S. Brunthaler, "Interpreter instruction scheduling," in *Compiler Construction*, J. Knoop, Ed., Heidelberg, GE: Springer, 2011, pp.164-178. http://dx.doi.org/10.1007/978-3-642-19861-8_10

[22] P. Zhao, and J. e. N. Amaral, "Feedback-directed switch-case statement optimization," in *Proc. 2005 Int. Conf. Parallel Process. Workshops*, Oslo, NO, 2005, pp. *295-302*. http://dx.doi.org/10.1109/icppw.2005.32

[23] K. Pettis, and R. C. Hansen, "Profile guided code positioning," *ACM Sigplan Not.,* vol. *25*, no. *6*, pp. *16-27*, 1990. http://dx.doi.org/10.1145/93548.93550

[24] D. Gregg, and J. Mccandless, "Optimizing interpreters by tuning opcode orderings on virtual machines for modern architectures," in *Conf. Princip. Prac. Program. Java*, Kongens Lyngby, DK, 2011, pp. *161-170*. http://dx.doi.org/10.1145/2093157.2093183

[25] Optimizations for a java interpreter using instruction set enhancement. (2005). Optimizations for a java interpreter using instruction set enhancement. [Online]. Available: https://www.scss.tcd.ie/publications/tech-reports/reports.05/TCD-CS-2005-61.pdf.

DOI: 10.15439/2015F217
ACSIS, Vol. 5

# Representation of a trend in OFN during fuzzy observance of the water level from the crisis control center

Jacek M. Czerniak
Casimir the Great University in Bydgoszcz
Institute of Technology
ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland
Email: jczerniak@ukw.edu.pl

Wojciech T. Dobrosielski, Łukasz Apiecionek, Dawid Ewald
Casimir the Great University in Bydgoszcz,
Institute of Technology,
ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland
Email: {wdobrosielski, lapiecionek, dawidewald }@ukw.edu.pl

***Abstract*—This paper presents the issue of fuzzy arithmetic calculations in two different notations. The well-known L-R notation proposed by Dubois-Prade, which enjoys a well-earned recognition of the researchers dealing with fuzzy logic was presented on one hand. On the other hand, a OFN notation introduced by Kosiński was discussed. Comparative calculations were performed using the data of the benchmark "Dam and Crisis control center paradox". That benchmark is available in two versions, where the opposite trend is visible at the dam and at the CCC (Crisis control center). In one of the versions, the water level at the observed area decreases and increases at the dam, while in the other version the situation is opposite. The trend difference detection can aid the short-term forecast of the situation change at the monitored area. Results of the applied calculations in OFN notation show that this arithmetic is sensitive to trend differences related to the order characteristic for those numbers. Relationship between the fuzzy logic and the trend of the observed phenomena is an added value to the generalization of OFN and it is also a good signal for the future development of applications of such fuzzy calculations, being their unique feature at the same time.**

## I. Introduction

THE HISTORY of artificial intelligence shows that new ideas were often inspired by natural phenomena. Many tourists come back with passion to beaches at the Oceanside and many sailors sail on tide water. The phenomenon of high and low tides, although well known, has been stimulating imagination and provoking reflexion on the perfection of the Creation for many ages. The casual observer in unable to precisely specify water level decline, but he or she can easily describe it using fuzzy concepts such as "less and less", "little" and "a bit". The same applies to increase of water in the observed basin. The observer can describe it such linguistic terms like "more", "lots of" or "very much". Such linguistic description of reality is characteristic to powerful and dynamically developing discipline of artificial intelligence like fuzzy logic. The author of Fuzzy logic is an American professor of the Columbia University in New York City and of Berkeley University in California - Lotfi A. Zadeh, who published the paper entitled "Fuzzy sets" in the journal "Information and Control" in 1965 [1]. He defined the term of a fuzzy set there, thanks to which imprecise data could be described using values from the interval (0,1). The number assigned to them represents their degree of membership in this set. It is worth mentioning that in his theory L.Zadeh used the article on 3-valued logic published 45 years before by a Pole - Jan Šukasiewicz [2]. That is why many scientists in the world regard this Pole as the "father" of fuzzy logic. Next decades saw rapid development of fuzzy logic. Another milestones of the history of that discipline should necessarily mention L-R representation of fuzzy numbers proposed by D.Dubois and H.Prade [3], [4], [5], which enjoys great successes today. Coming back to the original analogy, one can see some trend, i.e. general increase during rising tide or decrease during low tide, regardless of momentary fluctuations of the water surface level. This resembles a number of macro and micro-economic mechanisms where trends and time series can be observed. The most obvious example of that seem to be bull and bear market on stock exchanges, which indicate to the general trend, while shares of individual companies may temporarily fall or rise. The aim is to capture the environmental context of changes in economy or another limited part of reality. Changes in an object described using fuzzy logic seem to be thoroughly studied in many papers. But it is not necessarily the case as regards linking those changes with trend. Perhaps this might be the opportunity to apply generalization of fuzzy logic which are, in the opinion of authors of that concept, W.Kosiński [6] and his team, Ordered Fuzzy Numbers.

As the basis for experiments, let us assume the example of the dam and the impounding basin presented in the figure below (Fig. 1). Letter A indicates the water level measured during last evening measurement. Then there was a rapid surge of water in the night. Whereas morning measurement was marked using letter C. Measurements were imprecise to some extent due to rapid changes of weather conditions. It is also known that during the last measurement the safety valve $z2$, was open and then the valve $z1$ activated. The management of the dam faces the problem of reporting rapid surges of water to the disaster recovery centre.

978-83-60810-66-8/$25.00©2015, IEEE

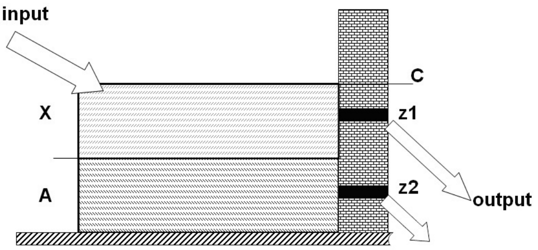


Fig. 1. The diagram of water flow in the impounding basin

## II. Theoretical background description of OFN

### A. Some definitions of OFN

Each operation on fuzzy numbers, regardless if it is addition, subtraction, division or multiplication, can increases the carrier value. Several operations performed on given L-R numbers can result in numbers that are too broad and as a result they can become less useful. Solving equations using conventional operations on fuzzy numbers [7] is usually impossible either. An $A+X=C$ equation can always be solved using conventional operations on fuzzy numbers only when A is a real number. First attempts to redefine new operations on fuzzy numbers were undertaken at the beginning of the 1990-ties by Witold Kosiński and his PhD student - P. SŞysz [8]. Further studies of W. Kosiński published in cooperation with P. Prokopowicz and D. Ślęzak [9], [7], [10] led to introduction of the **ordered fuzzy numbers model - OFN**.

*Definition 1: An* ***ordered fuzzy number*** *A was identified with an ordered pair of continuous real functions defined on the interval* $[0,1]$*, i.e.,* $A=(f,g)$ *with* $f,g:[0,1]\longrightarrow R$ *as continuous functions.*

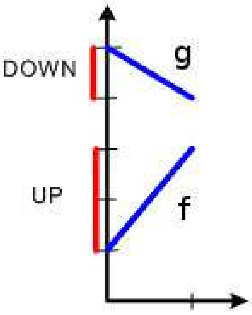


Fig. 2. Ordered fuzzy number

We call f and g the up and down-parts of the fuzzy number $A$, respectively. To be in agreement with the classical denotation of fuzzy sets (numbers), the independent variable of both functions $f$ and $g$ is denoted by $y$, and their values by $x$. [6]

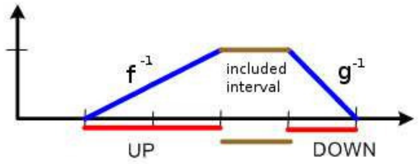


Fig. 3. OFN presented in a way referring to fuzzy numbers

Continuity of those two parts shows that their images are limited by specific intervals. They are named respectively: $UP$ and $DOWN$. The limits (real numbers) of those intervals were marked using the following symbols: $UP=(l_A,l_A^-)$ and $DOWN=(l_A^+,p_A)$. If both functions that are parts of the fuzzy number are strictly monotonic, then there are their inverse functions $x_{up}^{-1}$ and $x_{down}^{-1}$ defined in respective intervals $UP$ and $DOWN$. Then the following assignment is valid:

$$\begin{array}{c} l_A := x_{up}(0),\;\; l_A^- := x_{up}(1), \\ l_A^+ := x_{down}(1),\;\; p_A := x_{down}(0) \end{array} \tag{1}$$

If a constant function equal to 1 is added within the interval $[1_A^-,1_A^+]$ we get UP and DOWN with one interval (Fig. 2), which can be treated as a carrier. Then the membership function $\mu_A(x)$ of the fuzzy set defined on the R set is defined by the following formulas:

$$\begin{array}{lll} \mu_A(x)=0 & for & x\notin[l_A,p_A] \\ \mu_A(x)=x_{up}^{-1}(x) & for & x\in UP \\ \mu_A(x)=x_{down}^{-1}(x) & for & x\in DOWN. \end{array} \tag{2}$$

The fuzzy set defined in that way gets an additional property which is called order. Whereas the following interval is the carrier:

$$UP\,\cup\,[1_A^+,1_A^-]\,\cup\,DOWN \tag{3}$$

The limit values for up and down parts are:

$$\begin{array}{l} \mu_A(l_A)=0 \\ \mu_A(1_A^-)=1 \\ \mu_A(1_A^+)=1 \\ \mu_A(p_A)=0 \end{array} \tag{4}$$

Generally, it can be assumed that ordered fuzzy numbers are of trapezoid form. Each of them can be defined using four real numbers:

$$A\;=\;(l_A,1_A^-,1_A^+,p_A). \tag{5}$$

The figures below (Fig. 4) show sample ordered fuzzy numbers including their characteristic points.

Functions $f_A$, $g_A$ correspond to parts $up_A$ , $down_A\subseteq R^2$ respectively, so that:

$$up_A=(f_A(y),y):y\in[0,1] \tag{6}$$

$$down_A=(g_A(y),y):y\in[0,1] \tag{7}$$

The orientation corresponds to the order of graphs $f_A$ and $g_A$. The figure below (Fig. 5) shows the graphic interpretation

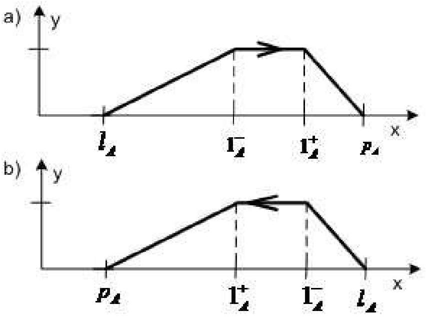


Fig. 4. Fuzzy number that is ordered a) positively b) negatively

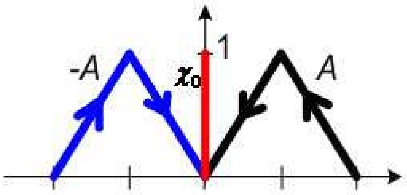


Fig. 5. Opposite numbers and the real number

of two opposite fuzzy numbers and the real number $\chi_0$. Opposite numbers are reversely ordered [11].

*Definition 2: A **membership function of an ordered fuzzy number** $A$ is the function $\mu_A : R \to [0,1]$ defined for $x \in R$ as follows [9], [6]:*

$$\mu(x) = \begin{cases} f^{-1}(x) & \text{if} \quad x \in [f(0), f(1)] = [l_A, 1_A^-] \\ g^{-1}(x) & \text{if} \quad x \in [g(1), g(0)] = [l_A^+, p1_A] \\ 1 & \text{if} \quad x \in [l_A^-, 1_A^+] \end{cases} \tag{8}$$

The above membership function can be used in the control rules similarly to the way membership of classic fuzzy numbers is used. All quantities that can be found in the fuzzy control describe selected part of the reality. Process of determining this value is called **fuzzy observation**.

### B. Arithmetic operations in OFN

The operation of adding two pairs of such functions is defined as the pair-wise addition of their elements, i.e., if $(f1, g1)$ and $(f2, g2)$ are two ordered fuzzy numbers, then $(f1 + f2, g1 + g2)$ will be just their sum. It is interesting to notice that as long as we are dealing with an ordered fuzzy number represented by pairs of affine functions of the variable $y \in [0,1]$, its so-called classical counterpart, i.e., a membership function of the variable $x$ is just of trapezoidal type. For any pair of affine functions $(f, g)$ of $y \in [0,1]$ we form a quaternion of real numbers according to the rule $[f(0), f(1), g(1), g(0)]$ which correspond to the four numbers $[l_A, 1_A^-, 1_A^+, p_A]$ as was mentioned in previous paragraph. If $(f, g) = A$ is a base pair of affine functions and $(e, h) = B$ is another pair of affine functions, then the set of typical operation will be uniquely represented by the following formulas respectively:

- addition $A + B = (f + e, g + h) = C$,
$$C \to [f(0)+e(0), f(1)+e(1), g(1)+h(1), g(0)+h(0)] \tag{9}$$
- scalar multiplication $C = \lambda A = (\lambda f, \lambda g)$,
$$C \to [\lambda f(0), \lambda f(1), \lambda g(1), \lambda g(0)] \tag{10}$$
- subtraction $A - B = (f - e, g - h) = C$
$$C \to [f(0)-e(0), f(1)-e(1), g(1)-h(1), g(0)-h(0)] \tag{11}$$
- multiplication $A * B = (f * e, g * h) = C$
$$C \to [f(0)*e(0), f(1)*e(1), g(1)*h(1), g(0)*h(0)] \tag{12}$$

### C. Association of OFN order with the environmental trend

In order to explain calculations presented in this sections, authors made the following assumptions concerning the context of changes taking place in the studied object (the impounding basin).

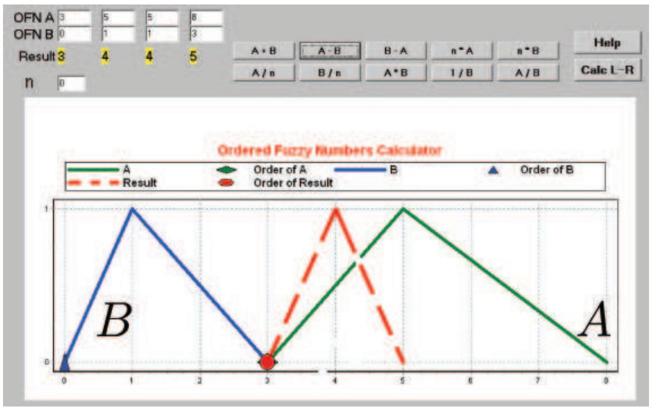
- **close context** - understood us the trend visible locally in the basin. It defines the trend of the object, i.e. if it is gradually filled or if the water level gradually falls. It is defined locally by the management of the dam,
- **further (environmental) context** - understood as the global trend for the observed area. It specifies trend of specific section of the river, intensity of precipitation as well as the set of other regional data which give better image of the environment. It is defined in the disaster recovery centre.

Two context types described above will be associated with the order of OFN numbers as follows. As the current water level state is reported to the disaster recovery centre, numbers that represent that state will always be oriented according to the environmental trend (reported to the centre). Whereas order of fuzzy numbers that represent changes of the water level in the impounding basin will be consistent with the local trend defined by the management of the dam. As a result, the order will be positive when the basin will be gradually filled, regardless the rate of that process. Whereas negative order can be observed when the outflow of water from the basin starts. Trend order changes themselves, as well as boundary conditions of the moment when they should occur, will make a separate process defined be the disaster recovery centre and the dam management respectively. Nothing stands in the way to ultimately use known segmentation methods for those processes and to determine the trend. Such issues are currently present in numerous publications and their detailed description is beyond the scope of this paper. Authors of that study assumed that the order equivalent to the trend changes is provided by a trusted third party and represents the expert's opinion concerning short-term weather forecast in the region important for the water level in the impounding basin as well as for the basin itself.

## III. An experimental comparison of fuzzy numbers arithmetic

### A. Elementary arithmetic operations

To automate computational experiments, authors of this study developed dedicated programme Ordered FN, which also efficiently supports graphic interpretation of operations being performed on ordered fuzzy numbers. Additionally, it is equipped with the Calc L-R module.



OFN A 3 5 5 8
OFN B 0 1 1 3
Result 3 4 4 5
n 0

Fig. 6. Main screen of the programme

Ordered FN is equipped with an additional module which is started using the Calc L-R button. It includes procedures to calculate the sum, the difference and the product of L-R numbers. This allows to compare some results obtained from operation on L-R and OFN numbers. The introduced fuzzy number has a form of $(m, \alpha, \beta)$ where $\alpha$ and $\beta$ are left and right-side dispersions. Shown here is an example of the software application. It includes the summary of simple arithmetic operation results in OFN and L-R notation. As shown in the attached figures (Fig. 6), subtraction of 3-3 is different for L-R than for OFN numbers.

### B. Comparison of calculations on ***L-R*** and ***OFN*** numbers

Majority of operation performed on L-R numbers, regardless if it is addition or subtraction, increases the carrier value, i.e. the areas of non-accuracy. Hence performance of several operations on those numbers can cause such big fuzziness that the resulting quantity will be useless. It is impossible to solve an $A + X = C$ equation using L-R representation through operations because for that interpretation $X + A + (-A) \neq X$ and $X*A*A^{-1} \neq X$. Whereas every inverse operation on L-R numbers will increase the carrier. It is often impossible to solve the equations using analytical (computations) method either. However, it is possible to break the stalemate using certain empirical methods. The situation is totally different in case of operation on ordered fuzzy numbers. It is possible to solve the above mentioned equation using an analytical method.

*Example 3.1:* Problem: There was a rapid surge of water in the impounding basin at night. The management of the dam has to send reports to the disaster recovery centre including the value of the water level change comparing to the previous state. Unfavourable weather conditions do not allow for precise measurement.

Data:

| | |
|---|---|
| $A(1, 1, 2)$ | - previous measurement [mln $m^3$] |
| $C(5, 2, 3)$ | - current measurement in [mln $m^3$] |

Mathematic interpretation:Hence the problem comes down to determining $X$ number that satisfies the equation $A+X = C$ as $A(1, 1, 2) + X = C(5, 2, 3)$

| **Solution versions:** |
|---|

| | |
|---|---|
| **Version I:** | *Solution using computational*<br>*method of OFN arithmetic.* |

$$A(0, 1, 1, 3) + X = C(3, 5, 5, 8)$$
$$X = C - A$$
$$X = (3, 4, 4, 5)$$

***Verification:*** $A + X = C$

$$A(0, 1, 1, 3) + X(3, 4, 4, 5) = C(3, 5, 5, 8)$$

| | |
|---|---|
| **Version II:** | *Solution using computational*<br>*method of L-R arithmetic.* |

$$A(1, 1, 2) + X = C(5, 2, 3)$$
$$X = C - A$$
$$X = (4, 4, 4)$$

***Verification:*** $A + X = C$

$$A(1, 1, 2) + X(4, 4, 4) \neq C(5, 2, 3)$$

As a result of operations on L-R numbers we obtain the outcome $(4, 4, 4)$. However, the verification through addition $(A + X)$ gives the result different from $C$. Correct result $(4, 1, 1)$ can only be achieved using empirical method. It is problematic and not always feasible.

## IV. Conclusion

Operations performed on ordered fuzzy numbers are often more accurate than operations performed on classic fuzzy numbers. Results of operations performed on them are the same as those obtained from operations on real numbers. Performing multiple not necessarily causes large increase of the carrier. The situation is different for L-R fuzzy numbers, where several operations often lead to numbers of high fuzziness. An infinitesimal carrier is interpreted as a real number and thus for OFN numbers one can apply commutative and associative property of multiplication over addition. The possibility to perform back inference on them allows to reproduce input data by solving an appropriate equation. This very property is an added value that makes this fuzzy logic extension worth to promulgate. Calculations performed on ordered fuzzy numbers are easy and accurate. It is worth mention the multiplication here, where the same procedure is used for all ordered fuzzy numbers regardless their sign. Whereas multiplication of L-R numbers is different for two positive numbers than for two negative ones. Another completely different procedure is used for multiplication of numbers of indefinite signs and for fuzzy zeros. It also seems to be very interesting to associate OFN numbers with trend of changes taking place for studied part of the reality. We are convinced that new applications of this property of OFN, shown here in the example of the fuzzy observation of the impounding basin in unfavourable weather conditions, will be introduced with the passing of time. Hence it seems that introduction of OFN gives new possibilities for designers of highly dynamic systems. With this approach it is possible to define trend of changes, which gives new possibilities for the development of fuzzy control and it charts new ways of research in the fuzzy logic discipline. Broadening it by the theory of ordered fuzzy numbers seems to allow for more efficient use of imprecise operations. Simple algorithmization of ordered fuzzy numbers allows to use them in a new control model. It also inspires researchers to search for new solutions. Authors did not use defuzzyfication operators [12] in this paper, which in themselves are interesting subject of many researches. They will also contribute to development of the comparative calculator created here. Although authors of this study are not so enthusiastic like the creators of OFN as regards excellent prospects of this new fuzzy logic idea, but they are impressed by possibilities for arithmetic operations performed using this notation. Even sceptics who treat OFN with reserve as it is generalization of fuzzy logic, can benefit from this arithmetic. After all OFN can be treated as internal representation of fuzzy numbers (heedless of it's authors' intention). With this new kind of notation for fuzzy numbers and fuzzy control, it is possible to achieve clear and easily interpreted calculation, which can be arithmetically verified regardless of the input data type. Perhaps the OFN idea will become another paradigm of fuzzy logic, just like the object oriented programming paradigm has become dominant in software engineering after the structured programming paradigm. Whichever scenario wins, at least some aspects of OFN arithmetic seem to be hard to ignore a priori.

## REFERENCES

[1] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338 – 353, 1965.

[2] J. Łukasiewicz, *O logice trójwartościowej*, 1988.

[3] D. Dubois and H. Prade, "Operations on fuzzy numbers," *International Journal of systems science*, vol. 9, no. 6, pp. 613–626, 1978.

[4] ——, "Fuzzy elements in a fuzzy set," in *Proc. IFSA*, vol. 5, 2005, pp. 55–60.

[5] ——, "Gradual elements in a fuzzy set," *Soft Computing*, vol. 12, no. 2, pp. 165–175, 2008.

[6] W. Kosiński, "On fuzzy number calculus," *Int. J. Appl. Math. Comput. Sci*, vol. 16, no. 1, pp. 51–57, 2006.

[7] W. Kosiński, P. Prokopowicz, and D. Ślęzak, "On algebraic operations on fuzzy numbers," in *Intelligent Information Processing and Web Mining*. Springer, 2003, pp. 353–362.

[8] W. Kosiński and D. SŞysz, "Fuzzy numbers and their quotient space with algebraic operations," *Bull. Polish Acad. Sci.Ser. Tech. Sci.*, vol. 41, pp. 285–295, 1993.

[9] W. Kosiński, P. Prokopowicz, and D. Ślęzak, "Ordered fuzzy numbers," *Bulletin of the Polish Academy of Sciences, Ser. Sci. Math*, vol. 51, no. 3, pp. 327–338, 2003.

[10] W. Kosiński, P. Prokopowicz, and K. Frischmuth, *On Algebra of Ordered Fuzzy Numbers, Soft Computing Foundations and Theoretical Aspects*, J. K. Krassimir T. Atanassow, Olgierd Hryniewicz, Ed. EXIT, 2004.

[11] P. Piotr, "Algorytmization of operations on fuzzy numbers and its applications (in polish)," Ph.D. dissertation, 2005.

[12] T. Bednarek, W. Kosiński, and K. Węgrzyn-Wolska, "On orientation sensitive defuzzification functionals," in *Artificial Intelligence and Soft Computing*. Springer, 2014, pp. 653–664.

[13] G. Gerla, "Fuzzy logic programming and fuzzy control," *Studia Logica*, vol. 79, no. 2, pp. 231–254, 2005.

[14] S. Gottwald, "Mathematical aspects of fuzzy sets and fuzzy logic: Some reflections after 40 years," *Fuzzy sets and systems*, vol. 156, no. 3, pp. 357–364, 2005.

[15] C. L. Walker and E. A. Walker, "The algebra of fuzzy truth values," *Fuzzy Sets and Systems*, vol. 149, no. 2, pp. 309–347, 2005.

[16] I. Couso and S. Montes, "An axiomatic definition of fuzzy divergence measures," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 16, no. 01, pp. 1–17, 2008.

[17] J. Dombi, "Towards a general class of operators for fuzzy systems," *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 2, pp. 477–484, 2008.

[18] L. A. Zadeh, "Is there a need for fuzzy logic?" *Information sciences*, vol. 178, no. 13, pp. 2751–2779, 2008.

[19] A. Grabowski, "On the computer certification of fuzzy numbers," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha, L. Maciaszek, Ed. IEEE, 2013, pp. 51–54.

[20] I. Bošnjak, R. Madarász, and G. Vojvodić, "Algebras of fuzzy sets," *Fuzzy Sets and Systems*, vol. 160, no. 20, pp. 2979–2988, 2009.

[21] Z. Xu, S. Shang, W. Qian, and W. Shu, "A method for fuzzy risk analysis based on the new similarity of trapezoidal fuzzy numbers," *Expert Systems with Applications*, vol. 37, no. 3, pp. 1920–1927, 2010.

[22] T. Nawarycz, K. Pytel, M. Gazicki-Lipman, W. Drygas, and L. Ostrowska-Nawarycz, "A fuzzy logic approach to the evaluation of health risks associated with obesity," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha, L. Maciaszek, Ed. IEEE, 2013, pp. 231–234.

[23] B. Rębiasz, B. Gaweł, and I. Skalna, "Fuzzy multi-attribute evaluation of investments," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha, L. Maciaszek, Ed. IEEE, 2013, pp. 977–980.

DOI: 10.15439/2015F313
ACSIS, Vol. 5

# Synthesis of Power Aware Adaptive Schedulers for Embedded Systems using Developmental Genetic Programming

Stanisław Deniziak
Department of Computer Science
Kielce University of Technology
Kielce, Poland
Email: s.deniziak@tu.kielce.pl

Leszek Ciopiński
Department of Computer Science
Kielce University of Technology
Kielce, Poland
Email: l.ciopinski@tu.kielce.pl

***Abstract*—In this paper we present a method of synthesis of adaptive schedulers for real-time embedded systems. We assume that the system is implemented using multi-core embedded processor with low-power processing capabilities. First, the developmental genetic programming is used to generate the scheduler and the initial schedule. Then, during the system execution the scheduler modifies the schedule whenever execution time of the recently finished task occurred shorter or longer than expected. The goal of rescheduling is to minimize the power consumption while all time constraints will be satisfied. We present real-life example as well as some experimental results showing advantages of our method.**

## I. Introduction

BESIDES the cost and performance, power consumption is one of the most important issue considered in the optimization of embedded systems. Design of energy-efficient embedded systems is important especially for battery-operated devices. Although the minimization of power consumption is always important, since it reduces the cost of running and cooling the system.

Embedded systems are usually real-time systems, i.e. for some tasks time constraints are defined. Therefore, power optimization should take into consideration that all time requirements should be met. Performance and power consumption are orthogonal features, i.e. in general, higher performance requires more power. Hence, the optimization of embedded system should consider the trade-off between power, performance, cost and perhaps other attributes.

Performance of the system may be increased by applying a distributed architecture. The function of the system is specified as a set of tasks, then during the co-design process, the optimal architecture is searched. Distributed architecture may consist of different processors, dedicated hardware modules, memories, buses and other components. Recently, the advent of embedded multicore processors has created an interesting alternative to dedicated architectures. First, the co-design process may be reduced to task scheduling. Second, advanced technologies for power management, like DVFS (Digital Voltage and Frequency Scaling) or big.LITTLE [1], create new possibilities for designing low-power embedded systems.

Optimization of embedded systems is based on assumptions that certain system properties are known. For example, to estimate the performance of the system, execution times for all tasks should be known. Sometimes it is difficult to precisely predict all required information. Therefore, to guarantee the proper design, the worst case estimation is used. During the operation of the system it may occur that certain system properties may significantly differ from estimations or may dynamically change. It may be caused by too pessimistic estimation, by data-dependence or by some unpredictable events. In such cases the idea of self-adaptivity may be used to optimize some system properties.

In this paper we present the novel method for synthesis of the power-aware scheduler for real-time embedded systems. We assume that the function of the system is specified using the task graph that should be executed by the multicore processor supporting the big.LITTLE technology. The scheduler is generated automatically using the developmental genetic programming (DGP). The scheduler is self-adaptive, i.e. it dynamically reschedules tasks whenever any task finished its execution earlier or later than expected. In the first case the goal of the rescheduling is the reduction of power consumption by moving some tasks to low-power cores. In the second case, the system is rescheduled to satisfy all time constraints by moving some tasks to high-performance cores. Example shows the benefits of using our methodology.

The rest of the paper is organized as follows. Next section presents the related work. Section III presents the concept of the developmental genetic programming with respect to other genetic approaches. In Section IV we present our method. Section V describes an example and experimental results. The paper ends with conclusions.

## II. Related Work

Although there are a lot of synthesis methods for low-power embedded systems [2], the problem of optimal mapping of a task graph onto the multicore processor is rather a variant of the resource constrained project scheduling (RCPSP)[3] one, than the co-synthesis. Since the RCPSP is NP-complete, only

978-83-60810-66-8/$25.00©2015, IEEE

heuristic approach may be applied to real-life systems. Among the proposed heuristics for solving RCPSP, ones of the most efficient are methods based on genetic algorithms [4][5][6].

For systems that may dynamically change during operation, some methods of rescheduling was proposed. In [7] the number of tasks that receives a new start time, after rescheduling, is minimized. Another approach [8] proposes to reschedule the remaining tasks, such that the sum of deviations of the new finishing times from the original ones is minimized. In [9] the sum of deviations of starting and finishing times of all tasks is minimized. Proactive scheduling [10] does not perform rescheduling, but it minimizes the perturbations, caused by delays, by maximization of the minimum or total free slacks of task executions.

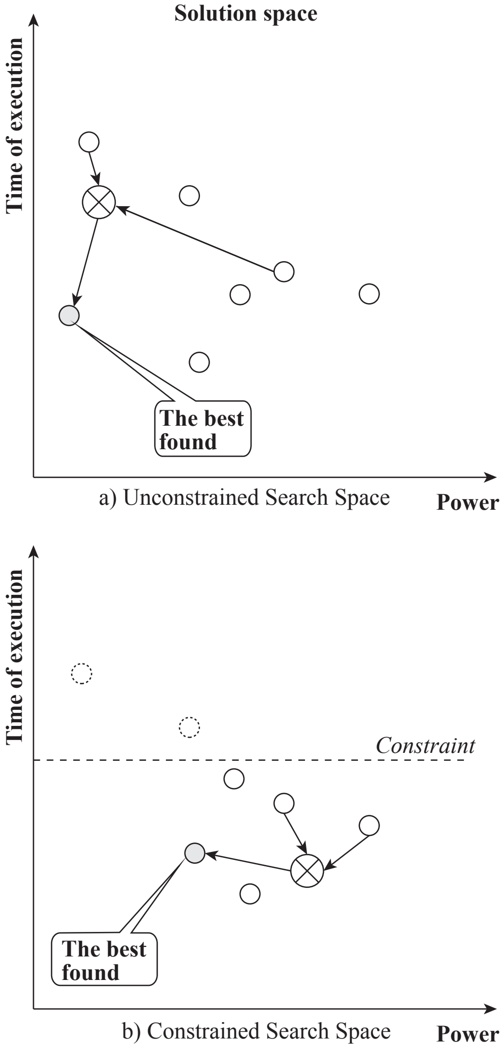
Three different methods of applying big.LITTLE technology for minimizing the power consumption were proposed[11]. In the cluster switching, low-power cores are grouped into “little cluster", while high performance cores are arranged into “big cluster". The system uses only one cluster at a time. If at least one high performance core is required then the system switches to the “big cluster", otherwise the “little cluster" is used. Unused cluster is powered off. In CPU migration approach, low-power and high-performance cores are paired. At a time only one core is used while the other is switched off. At any time it is possible to switch paired cores. The most powerful model is a global task scheduling. In this model all cores are available at the same time.

The big.LITTLE technology is quite new and is mainly used in mobile devices. According to our best knowledge there are no applications of this technology to design low power real-time embedded systems, as well as an adaptive scheduling method for such systems.

## III. Developmental Genetic Programming

Genetic algorithms (GA) [12] are very commonly used in wide spectrum of optimisation problems. Main advantage of GA approach is the possibility of getting out from the local minima of optimization criterion. Thus, GA is efficient for global optimization of complex problems, like RCPSP or multi-objective optimization of distributed real-time systems [13].

Although GA approach usually give satisfactory results it may be inefficient for hard constrained problems. In these cases a lot of individuals obtained using genetic operators correspond to not feasible solutions (e.g. schedule that exceeds required deadline or incorrect schedule, in the RCPSP). Such individuals should not be considered during the evolution. It is provided by defining the constrained genetic operators, that produce only correct solutions. But such constrained operators may create infeasible regions in the search space. Such regions may contain optimal or close to optimal solutions. This problem is illustrated on Fig. 1. Assume that we optimize the power consumption of the real-time system. Due to constraint violation, the solutions above the dotted line are not valid, hence they are never produced during the evolution. But it is possible that such “forbidden" individuals may be used during the crossover or mutation to produce highly optimized solutions (Fig. 1a). Unfortunately, the search space should be limited to consider only valid solutions (Fig.  1b) and the optimal solution may never be obtained.



a) Unconstrained Search Space

b) Constrained Search Space

Fig. 1. Search space in GA

Above problem may be eliminated by using the Developmental Genetic Programming (DGP). DGP is an extension of the GA by adding the developmental stage. This method first time was applied to optimize analog circuits [14]. The main difference between DGP and GA is that in the DGP genotypes represent the method building the solution, while in the GA genotypes describe the solution. Thus, during the evolution, a method of building a target solution is optimized, instead of a solution itself.

In the DGP the search space (genotypes) is separated from the solution space (phenotypes). The search space is not constrained, all individuals are evolved. Thus, all of them may take part in the reproduction, crossover or mutation. There is no “forbidden" genotypes.

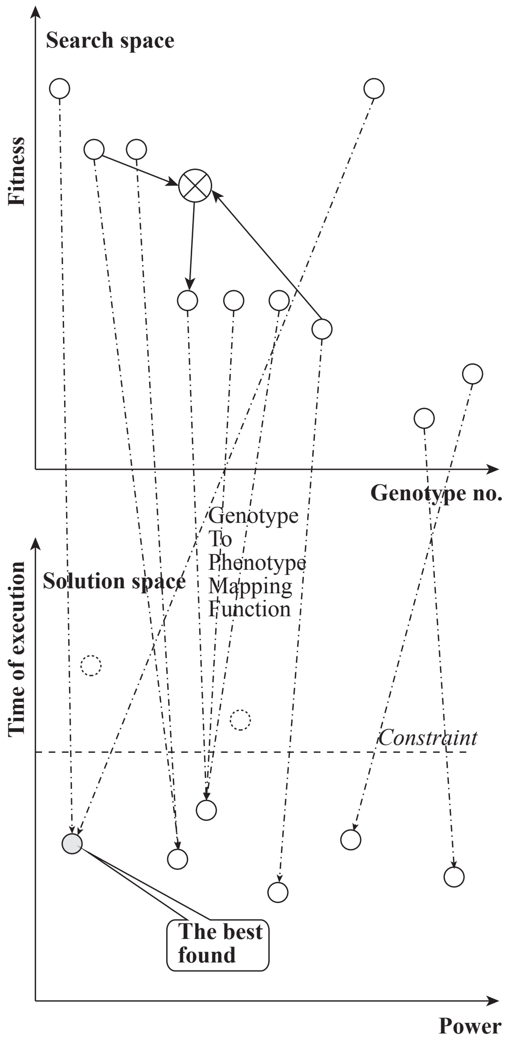


Fig. 2. Genotype to Phenotype mapping

Phenotypes are created by using genotype-to-phenotype mapping function, which always produces a valid solution. During the evolution, the fitness of the genotype is evaluated according to the quality of the corresponding phenotype. Fig. 2 presents the idea of the DGP. The search space consists of the genotypes that evolve without any restrictions. Any genotype may be mapped onto phenotype representing valid solution. This can be assured by using constrained mapping function. This function never produces solutions above the constraint line.

This idea of DGP is taken from biology, where the genotype corresponds to the chromosome containing information used for synthesis of proteins. The application of DGP occurred successful in many domains [15], where human-competitive results were obtained. High efficiency of the DGP-based optimization was also proved for hardware-software codesign[16] and cost minimization in real-time cloud computing[17].

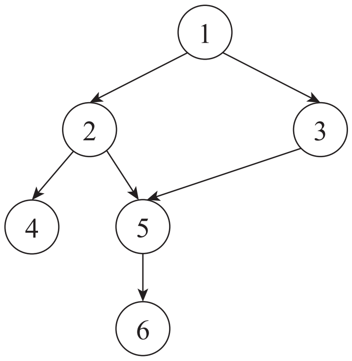


Fig. 3. A sample task graph

## IV. Synthesis of Adaptive Scheduler

Idea of our approach is based on the observation that when the DGP will be applied for the RCPSP problem, then except the final schedule we also obtain the scheduler dedicated to the optimized system. Thus, instead of the implementation of static schedule we may implement this scheduler, which may adapt to any perturbation during the system operation. We assume that the system is specified as a task graph. This is very widely used method of specification of real-time embedded systems. We also assume that for each task, the time of execution and the average power consumption is known for each available processor core. Usually these parameters are estimated using the worst case estimation methods. During the system operation, the scheduler will dynamically modify the schedule to minimize the power consumption whenever it will be possible to move some tasks to low-power cores, i.e. when execution time of finished tasks will occur shorter than estimated. In our method it is also possible to use average execution time, instead of the worst case estimation. When same task will be delayed then the scheduler will try to find the new schedule that satisfies the time requirements. We use ARM multicore processors with big.LITTLE technology for implementation of the target systems. Such system consists of two processors, usually quad-core. The first of them has higher performance (about 40%), but consumes more power. The second one is slower, but it is optimized to use much less energy (about 75%), to execute the same task. The goal of optimization is to find the makespan for which the power consumption is as small as possible, while all time constraints are met.

### *A. Task Graph*

The function of an embedded system is specified as a set of tasks. Between certain tasks may be the relationship that specifies the order of their execution. This may be specified as a task graph, which is the acyclic directed graph where nodes correspond to tasks and edges describe required order of execution. A sample task graph is given on Fig.3.

TABLE I
A SAMPLE LIBRARY OF RESOURCES

| Task # | Core # | Execution time [ns] | Power Consumption [mJ] |
|---|---|---|---|
| 1 | 0 | 537 | 5 |
| 1 | 1 | 537 | 5 |
| 1 | 2 | 537 | 5 |
| 1 | 3 | 537 | 5 |
| 1 | 4 | 671 | 3 |
| 1 | 5 | 671 | 3 |
| 1 | 6 | 671 | 3 |
| 1 | 7 | 671 | 3 |
| 2 | 0 | 1072 | 11 |
| ... | ... | ... | ... |
| 6 | 7 | 176 | 1 |

TABLE II
SCHEDULER'S PREFERENCES

| Step | Option | P |
|---|---|---|
| 1 | a. The highest performance<br>b. The lowest power<br>c. The lowest time * power<br>d. Determination by second gene<br>e. The fastest starting core<br>f. The fastest finishing core | 0.16(6)<br>0.16(6)<br>0.16(6)<br>0.16(6)<br>0.16(6)<br>0.16(6) |
| 2 | List scheduling | 1 |

### B. Resources

Estimated execution parameters are given in a library of available resources. A resource is a core of a processor, which is able to execute a task. For each core the execution time and the power consumption are given. Part of a sample ARM Cortex-A15/Cortex-A7 database, for task graph from Fig.3, is presented in Table I.

### C. Strategies of scheduling

The scheduler creates a makespan in two steps:

1) task assignment: tasks are assigned to cores according to preferences specified for each group of tasks (Table II),
2) task scheduling: this step is executed only when more than one task is assigned to the same core. During this step a selected group of tasks is scheduled using the scheduling strategy specified for this group (Table II).

Initial population consists of randomly generated genotypes. During initialization, preferences defining the decision table for the scheduler are assigned to each gene. Table II contains the set of possible preferences that the scheduler may choose. The last column in Table II shows a probability of the selection.

The first option prefers the core with the highest performance. Second one prefers a core with the lowest power consumption. Third option prefers a core with the best ratio of the power consumption to the time of execution. Fourth option allows using a core, that cannot be obtained as a result of the remaining options. The next option prefers a core, which could start an execution of the task as soon as possible (other cores might be busy). The last option prefers a core which could be the first to finish a task (be freed). For the second step only one option is available, the list scheduling method.

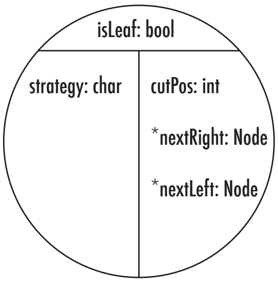
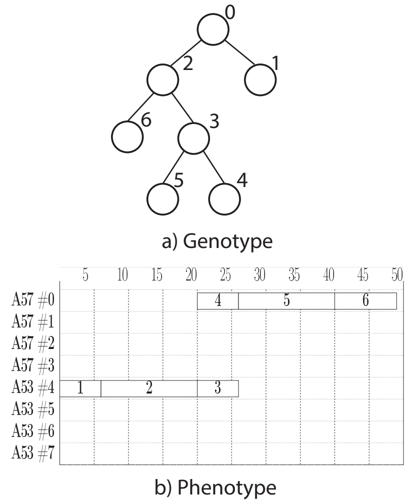


Fig. 4. A node of the genotype



a) Genotype

b) Phenotype

Fig. 5. A sample genotype (a) and the corresponding phenotype (b)

### D. Genotype

The genotype has a form of binary tree corresponding to the certain procedure of task scheduling[18]. Every node in the genotype has the structure presented on Fig. 4.

The first field *isLeaf* determines a type of the node in a tree. When the node is a leaf this field equals true. Then, the field named *"strategy"* defines the strategy of scheduling for group of tasks assigned to this node. All possible strategies are given in Table II. In this case, information from the other fields is omitted. When the node is not a leaf, a content of the field *"strategy"* is neglected. In this case, *cutPos* contains a number describing which group of tasks should be assigned and scheduled by the left node and which one by the right one. Thus, *nextLeft* and *nextRight* must not be null pointers.

The simplest genotype consists of only one node, which is also a leaf and a root. A sample genotype and the corresponding phenotype are presented on Fig. 5.

During the evolution a genotype may grow up but the size of the tree is limited. If a tree will be too large, the performance of genotype to phenotype mapping, required for the fitness evaluation, would be slightly decreased. Size limit of the genotype also avoids constructing too many unused branches.

An initial genotype tree may grow up as an effect of genetic operators: mutation and crossover. An action associated with

TABLE III
THE RULES OF MUTATIONS

| Is a leaf? | | | |
|---|---|---|---|
| Yes | | No | |
| Draw: switch leaf/node or not? | | | |
| Yes | No | Yes | No |
| Set *isLeaf* as FALSE. If *nextLeft* or *nextRight* is NULL - create a new leaf for it. | Draw new strategy | Set isLeaf as TRUE | Change value for a randomly chosen field: *cutPos*, *nextLeft* or *nextRight* |

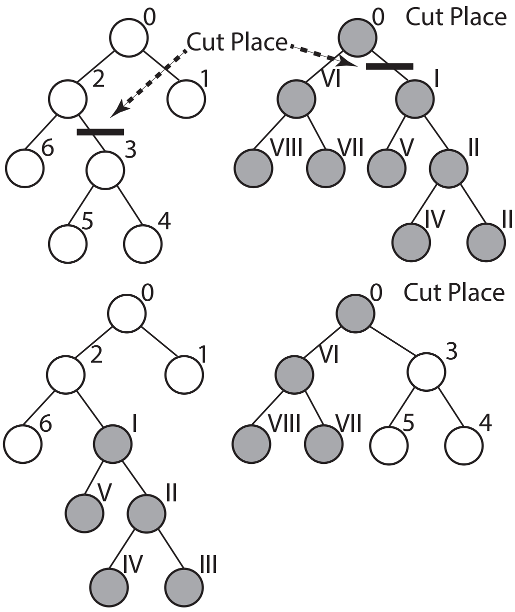


Fig. 6. An example of the crossover

the mutation depends on the type of the node and is presented in the Table III [18].

The crossover is used to create new individuals that are a combination of genes of parent genotypes. First, points of cut for both trees are drawn, then the cut branches are exchanged. In this way two new genotypes are created. A sample crossover is presented on Fig. 6.

With every genotype an array is associated. Its size is equal to the number of tasks and contains indexes of cores. If for given task, strategy 'd' is chosen, the core with an index taken from the array is used. During the mutation, a position in the array is randomly chosen. Then, a new index is randomly generated. During the crossover, parts of the arrays from both genotypes are swapped. The array defines the alternative scheduling strategy that is not driven by the performance or power consumption.

### E. Genotype to phenotype mapping

The first step, during the genotype-to-phenotype mapping, is to assign strategies to tasks (i.e. preferences for assigning tasks to resources). For the example from Fig. 11, this step is illustrated on Fig. 7. Node 0 groups tasks into two sets: {1,2,3} and {4,5,6}. The first set is partitioned by node 2 into next two groups. In the first one, there is only task 1. Tasks 2 and 3 belong to the second one, which is partitioned again by node 3. The *cutPos* parameter of node 3 equals 4, this means that tasks should be partitioned into group from 2 to 5 and the rest. But tasks 5 and 6 are outside of the set assigned to node 3, these tasks are assigned to node 1. Thus, there is only one group of tasks {2, 3} assigned to node 5.

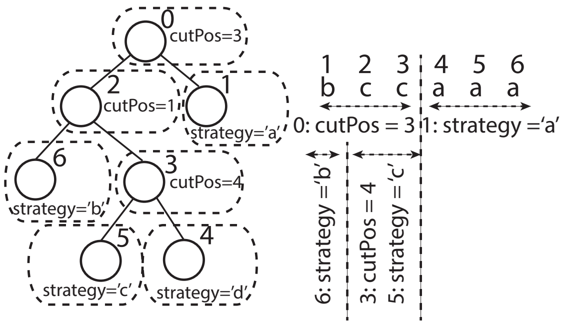


Fig. 7. The first step in genotype-to-phenotype mapping

In the second step, all tasks without any predecessor in the task graph, or with predecessors having already assigned core, are being searched for. These tasks are assigned to cores according to preferences determined in the first step. This step is repeated as long as there are tasks without assigned cores. In the third step, the total power consumption of the solution is calculated. For this purpose, the resource library (Table I) is used.

### F. Parameters of DGP

During the evolution, new populations of schedulers are created using genetic operations: reproduction, crossover (recombination) and mutation. After the genetic operations are performed on the current population, a new population replaces the current one. The evolution is controlled by the following parameters:

- *population size*: the number of individuals in each population is always the same. The value of this parameter is determined according to the value of "number of tasks" * "number of cores",
- *reproduction size*: number of individuals created using the reproduction,
- *crossover size*: the number of individuals created using the crossover,
- *mutation size*: the number of individuals created using the mutation.

Finally, the selection of the best individuals by a tournament is chosen [12]. In this method, chromosomes (genotypes) are drawn with the same probability in quantity defined as a size of the tournament. The best one is taken to the next generation. Hence, the tournament is repeated as many times as the number of chromosomes for a reproduction, crossover and mutation is required. A size of the tournament should be defined carefully. It should not be too high, because the selection pressure is too strong and the evolution will be too greedy. It also should not be too low, because the time of finding any better result would be too long.

### G. Self-adaptability of the scheduler

Finding the best makespan for low-power real-time embedded system is not the only goal of our approach. DGP methods are very effective in solving optimization problems and very often give the optimal solution. From the other side, they give results in relatively long time, thus genetic approach can not be used for rescheduling in real-time systems.

In the DGP the scheduling is performed during the genotype to phenotype mapping. This process is very fast, therefore it can be executed during the system operation. So, instead of implementing the final schedule we implement the method which creates this schedule. We observed that such approach has great self-adaptability capabilities.

Since the DGP has to consider only valid makespans. The genotype to phenotype mapping is a constrained process. If for a set of preferences defined by the genotype, it is not possible to obtain the valid phenotype then the mapping selects the next matching resource. Thus, the preferences specified by the genotype need not be strictly adhered. For example, if for given task preference suggest assigning this task to the low-power core, then if this decision will lead to an infeasible makespan, the scheduler will choose another, faster core that best matches to this preference.

Therefore, scheduler is able not only to build a correct solution, but also modify it if any unpredictable events will occur. E.g. if a task execution will be longer than expected, then the scheduler could move some tasks from a slower to a faster core, to fulfill time requirements. Similarly, if a task will be finished before its predicted end time, scheduler can move other tasks from a faster core to slower one, to save some energy.

### H. Fitness function

A fitness function determines the optimization goal of the DGP. In the presented approach, two options are possible. In the first one, the cheapest solution which has to be finished before a deadline is searched for. Such fitness function is applied when hard real time constraints have to be satisfied. In the second one, the DGP should find the fastest solution, which does not exceed a given power consumption. This case concerns systems with soft real-time requirements.

## V. Example and Experimental Results

We have verified advantages of the presented method using example of the complex multimedia system, which was described in [19]. The result has been compared with the method based on Least-Laxity-First Scheduling Algorithm [20].

### A. Task Graph and Run-time Parameters

The sample system is a multimedia player implemented as a real-time embedded system. The specification of the system consists of 40 tasks. Fig. 8 presents the task graph describing the system, details are given in [19]. We consider shared memory architecture, thus the communication between tasks may be neglected. The execution time is critical - it is typical soft real-time system. If the deadline is only slightly exceeded, the quality of the system is decreased, but the solution may be accepted. If the system exceeds hard deadline, then result is unacceptable and system should be redesigned.

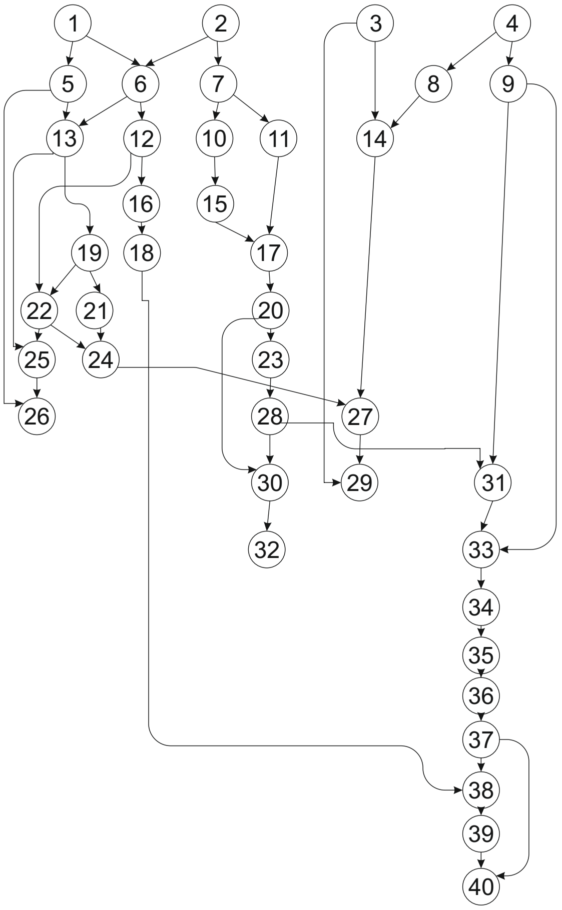


Fig. 8. Task graph of the multimedia system

We assumed, that the application will be implemented in software running on system consisting of two 4-core processors. One of them is ARM Cortex A57 and the second is ARM Cortex A53. The first processor is faster for about 25%, but consumes about 50% more power. Both processors support ARM big.LITTLE technology. Run-time parameters for both processors are given in Table IV.

The goal of our methodology is to create a self-adapting scheduler, which run the program tasks, balancing them between the cores, to minimize the power usage. The scheduler should be able to reschedule remaining tasks, whenever any task will finish its execution before or after expected time frame.

TABLE IV
EXECUTION TIME AND POWER CONSUMPTION

| **Task** | **Processor cores** | | | |
|---|---|---|---|---|
| | ***A57 (high performance)*** | | ***A53 (energy efficient)*** | |
| | ***energy*** | ***time*** | ***energy*** | ***time*** |
| 1 | 5 | 537 | 3 | 671 |
| 2 | 11 | 1072 | 6 | 1340 |
| 3 | 5 | 537 | 3 | 671 |
| 4 | 4 | 376 | 2 | 470 |
| 5 | 73 | 7337 | 37 | 9171 |
| 6 | 11 | 1072 | 6 | 1340 |
| 7 | 110 | 10958 | 55 | 13698 |
| 8 | 74 | 7358 | 37 | 9198 |
| 9 | 11 | 1051 | 6 | 1314 |
| 10 | 6 | 559 | 3 | 699 |
| 11 | 5 | 486 | 3 | 608 |
| 12 | 3 | 286 | 2 | 358 |
| 13 | 13 | 1298 | 7 | 1623 |
| 14 | 37 | 3679 | 19 | 4599 |
| 15 | 21 | 2065 | 11 | 2581 |
| 16 | 53 | 5253 | 27 | 6566 |
| 17 | 75 | 7523 | 38 | 9404 |
| 18 | 11 | 1076 | 6 | 1345 |
| 19 | 4 | 409 | 2 | 511 |
| 20 | 4 | 409 | 2 | 511 |
| 21 | 11 | 1076 | 6 | 1345 |
| 22 | 2 | 157 | 1 | 196 |
| 23 | 260 | 26018 | 130 | 32523 |
| 24 | 2 | 176 | 1 | 220 |
| 25 | 2 | 197 | 1 | 246 |
| 26 | 260 | 26018 | 130 | 32523 |
| 27 | 236 | 23607 | 118 | 29509 |
| 28 | 6 | 559 | 3 | 699 |
| 29 | 11 | 1072 | 6 | 1340 |
| 30 | 110 | 10958 | 55 | 13698 |
| 31 | 5 | 486 | 3 | 608 |
| 32 | 3 | 286 | 2 | 358 |
| 33 | 11 | 1072 | 6 | 1340 |
| 34 | 4 | 409 | 2 | 511 |
| 35 | 4 | 409 | 2 | 511 |
| 36 | 236 | 23607 | 118 | 29509 |
| 37 | 74 | 7414 | 37 | 9268 |
| 38 | 3 | 253 | 2 | 316 |
| 39 | 2 | 179 | 1 | 224 |
| 40 | 2 | 176 | 1 | 220 |

### B. Genetic parameters

We assumed that the deadline for the system from Fig.4 is equal to 100000 ns. The Power Aware Scheduler and the optimized makespan were generated using DGP. During the experiments, the following values of genetic parameters were used:

- the evolution was stopped after 100 generations,
- each experiment was repeated 7 times,
- the population size was equal to 128,
- tournament size was equal to 10,
- the number of mutants in each generation, was equal to 20%,
- the crossover was applied for creation of 40% genotypes,
- 20% of individuals were created using reproduction.

The values of parameters described above were tuned according to method described in our previous work [18], thus we will describe it here very shortly. In the first step, we estimated an influence of the tournament size. When this parameter was too small, the evolution got stuck. When the tournament size was too big, the DGP found semi optimal solution very fast, but a further optimization was not possible. Next, the influence of crossover and mutation for obtaining the best solution has been tested. It has been done by searching for the best solution using different combination of these parameters. Thus the best values of these parameters have been chosen. Finally, the best combination of other evolution parameters has been evaluated.

### C. Least-Laxity-First Algorithm

One of the most known algorithms for scheduling tasks in real-time embedded systems is the Least-Laxity-First Algorithm (LLF). Basic LLF method schedules task according to the least laxity (slack time). The laxity is defined as a difference between an execution time and a task deadline. The goal of LLF is to find the schedule that satisfies all deadlines. It does not take into account power or cost optimization. Therefore, we modify this method by favouring energy-efficient cores. In other words, during scheduling, the method first tries assign a task to low-power core, only when it will violate the time constraint, the task will be assigned to more efficient core. Our Low Power LLF (LPLLF) method is used only for reference, to verify that the DGP is efficient also for power optimization in real-time embedded systems.

### D. Power-aware Scheduling

The makespans obtained using DGP and LPLLF methods are presented on Fig. 9. On Y-axis different cores are represented, while the time of execution is represented by the X-axis. Numbers correspond to the following tasks. The experimental results proved that the presented method is more efficient than LPLLF. Energy consumption for the system scheduled using DGP equals 990mJ, while the same example scheduled using LPLLF requires 1018mJ. To meet the deadline, the LPLLF method assigned the long task 36 to the most efficient core. But in the DGP, more energy-efficient solution was found by assigning some shorter tasks, that in total consume less power than task 36, to the faster core.

Above experiment showed that the scheduler constructed using DGP is able to find highly optimized solutions. More experiments proving this remark are given in our previous work [16][17][18].

### E. Self-adaptivity capabilities

Static scheduling is based on estimation of execution times for all tasks. During the system operation, time of execution may significantly be shorter (e.g. if the worst case estimation was applied) or longer (e.g. in case of the most likely estimation). Therefore, the system may be additionally optimized during run-time, by using self-adaptive scheduling. In this way certain system parameters (power consumption, performance) may be improved. Scheduler generated using our method consists of series of system construction functions, corresponding to each gene. These functions are flexible, i.e. design decisions

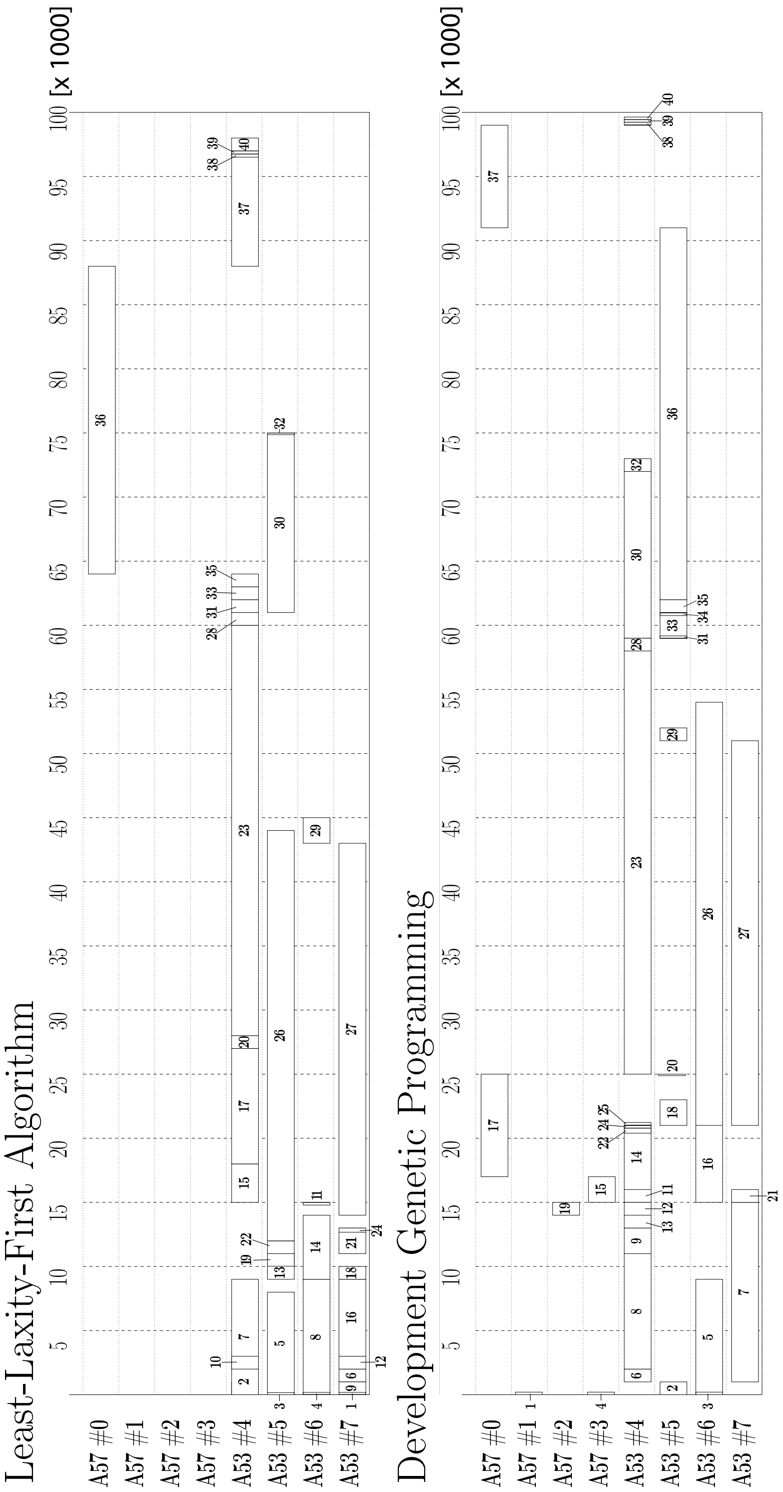


Fig. 9. Makespans obtained using LLF and DGP

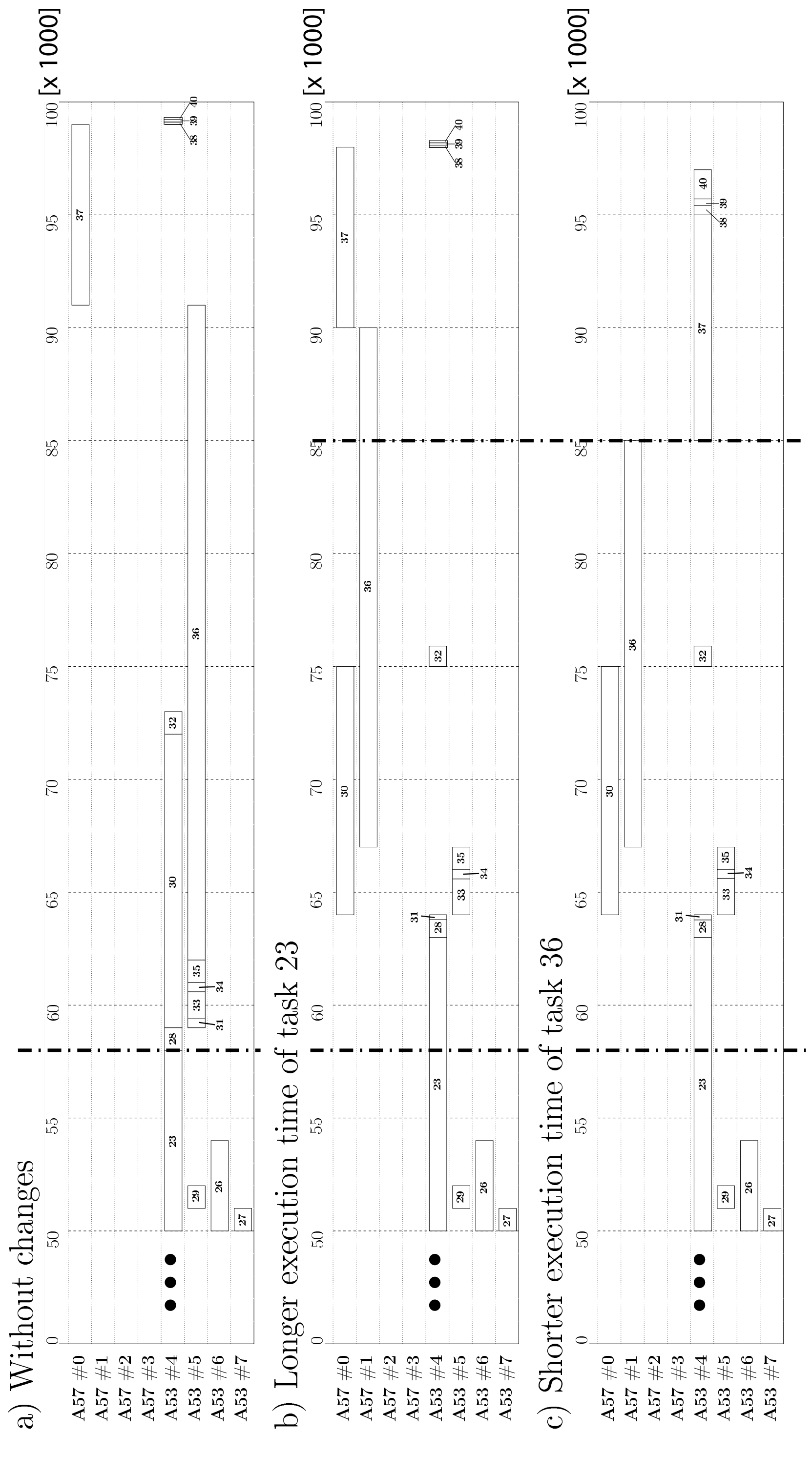


Fig. 10. Self-adaptation capabilities of power-aware scheduler

are driven by preferences, they are not strictly defined. This is necessary to assure that only feasible makespan will be created. Flexibility of the system construction functions provides to self-adaptivity capabilities of the scheduler.

An example of the self-adaptivity is presented on Fig. 10. If the execution time of task 23 will be too long, then all succeeding tasks should be postponed and the system would exceed the deadline. But our scheduler adapts to the delayed end time of task 23, and despite the fact that it uses the same construction functions, some tasks (tasks 30 and 36 in our example) will be assigned to more efficient cores (Fig. 10b). In other case, if an execution time of a task 36 will be shorter than it was expected, the scheduler will assign some tasks (task 37 in our example) to low power core (Fig. 10c). In this way power consumption will be reduced.

To verify the capabilities of self-adaptivity of the scheduler we performed some simulations of different changes in execution times for some task. Table V presents results obtained for cases when execution times occurred longer than estimated. In all cases our scheduler was able to adapt to this situation and new makespans that satisfied the deadline were created.

Table VI presents results obtained for another cases, where execution times for some tasks occurred shorter than expected, i.e. estimation was too pessimistic. In such case the scheduler has an opportunity to reduce the power consumption. It should be noticed that in some cases the scheduler found more energy-efficient makespans, but a lot of makespans were not changed. The main reason is that in the most cases there was not possible to improve the power consumption because all remaining tasks were already assigned to low-power cores. It is visible when we compare results obtained for different deadlines. When the deadline is shorter than 95 000 there is still a possibility for improvement. For longer deadlines, even when all tasks were faster it was not possible to decrease the power consumption.

## VI. Conclusions

In this paper the method of automatic synthesis of power-aware schedulers for real-time distributed embedded systems was presented. Starting from the system specification in the form of the task graph, we use developmental genetic programming to optimize the scheduling strategy that minimizes the power consumption. Finally, the best makespan as well as the optimized scheduler are generated. The scheduler has powerful capabilities of self-adaptation. This feature may be used to dynamically minimize the power consumption as well as to increase the system performance.

The presented method is dedicated to ARM big.LITTLE technology, developed for a low power systems. But, since we use general optimization method, it would be easily adapted to other energy-efficient architectures.

The computational experiments confirmed, that the schedulers, generated using DGP, are efficient and flexible. For the sample system our method gave better results than results obtained using LLF-based method. Moreover, simulations showed that the scheduler is able to quickly and effectively react to any changes of task execution times, by rescheduling remaining tasks.

Despite the above advantages of our method, there is still possible to improve the methodology. In the future work, we will consider special types of adaptive genes, that could support more possibilities for self-adaptation, we will also consider using other scheduling methods, alternative to list scheduling, e.g. based on mathematical/constrained programming [21].

## References

[1] big.LITTLE Processing with $ARMCortex^{TM}$ - A15 & Cortex-A7, ARM Holdings, September 2013, http://www.arm.com/files/downloads/big.LITTLE_Final.pdf.

[2] J.Luo, N.K. Jha, Low Power Distributed Embedded Systems: Dynamic Voltage Scaling and Synthesis, Proc. 9th Int. Conference High Performance Computing - HiPC 2002, Lecture Notes in Computer Science, vol. 2552, 2002, pp. 679-693. http://dx.doi.org/10.1007/3-540-36265-7_63

[3] Hartmann S., Briskorn D., A survey of variants and extensions of the resource-constrained project scheduling problem, European journal of operational research : EJOR. - Amsterdam : Elsevier, Vol. 207., 1 (16.11.), pp. 1-15 (2010). http://dx.doi.org/10.1016/j.ejor.2009.11.005

[4] Hartmann, S. (1998). An competitive genetic algorithm for resource-constrained project scheduling. Naval Research Logistics, 45(7), 733-750. http://dx.doi.org/10.1002/(SICI)1520-6750(199810)45:7%3C733::AID-NAV5%3E3.3.CO;2-7

[5] Xiang Li, Lishan Kang, Wei Tan, "Optimized Research of Resource Constrained Project Scheduling Problem Based on Genetic Algorithms", Lecture Notes in Computer Science, Vol. 4683, 2007, pp 177-186. http://dx.doi.org/10.1007/978-3-540-74581-5_19

[6] Hossein Zoulfaghari, Javad Nematian, Nader Mahmoudi, and Mehdi Khodabandeh. 2013. A New Genetic Algorithm for the RCPSP in Large Scale. Int. J. Appl. Evol. Comput. 4, 2 (April 2013), 29-40. http://dx.doi.org/10.4018/jaec.2013040103

[7] K.M. Calhoun, R.F. Deckro, J.T. Moore, J.W. Chrissis, J.C.V. Hove, Planning and re-planning in project and production scheduling, Omega The international Journal of Management Science 30 (3) (2002) 155-170. http://dx.doi.org/10.1016/S0305-0483(02)00024-5

[8] S. Van de Vonder, E.L. Demeulemeester, W.S. Herroelen, A classification of predictive-reactive project scheduling procedures, Journal of Scheduling 10 (3) (2007) 195-207. http://dx.doi.org/10.1007/s10951-007-0011-2

[9] H. Sakkout, M. Wallace, Probe backtrack search for minimal perturbation in dynamic scheduling, Constraints 5 (4) (2000) 359-388. http://dx.doi.org/10.1023/A:1009856210543

[10] M. Al-Fawzan, M. Haouari, A bi-objective model for robust resource-constrained project scheduling, International Journal of Production Economics 96 (2005) pp.175-187. http://dx.doi.org/10.1016/j.ijpe.2004.04.002

[11] Brian Jeff, "Ten Things to Know About big.LITTLE". ARM Holdings, 2013,http://community.arm.com/groups/processors/blog/2013/06/18/ten-things-to-know-about-biglittle

[12] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag Berlin Heidelberg, 1996. http://dx.doi.org/10.1007/978-3-662-03315-9

[13] Dick, R.P., Jha, N.K.: MOGAC: A Multiobjective Genetic Algorithm for the CoSynthesis of Hardware-Software Embedded Systems. IEEE Trans. on ComputerAided Design of Integrated Circuits and Systems 17(10), 920-935 (1998). http://dx.doi.org/10.1109/43.728914

[14] Koza, J., Bennett III , F. H., Andre, D., Keane, M. A., 1998. Evolutionary Design of Analog Electrical Circuits Using Genetic Programming. In: I. C. Parmee (ed.), Adaptive Computing in Design and Manufacture. http://dx.doi.org/10.1007/978-1-4471-1589-2_14

[15] J.R.Koza, R.Poli, "Genetic Programming", In Edmund Burke and Graham Kendal, editors. "Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques", Chapter 5. Springer, 2005. http://dx.doi.org/10.1007/0-387-28356-0_5

[16] S.Deniziak, A.Górski, "Hardware/Software Co-Synthesis of Distributed Embedded Systems Using Genetic Programming", Lecture Notes in Computer Science, Springer-Verlag, 2008, pp.83-93. http://dx.doi.org/10.1007/978-3-540-85857-7_8

TABLE V
SELF-ADAPTATION FOR DIFFERENT DELAYS

| Case | Delay | Deadline [ns] | Time without resche-duling | Time after resche-duling | Time-out [%] | Energy [mJ] |
|---|---|---|---|---|---|---|
| 0 | (none) | 90000 | 89756 | - | 0 | 1494 |
| 1 | T2 + 53% | 90000 | 90466 | 88612 | 0,52 | 1586 |
| 2 | T8+20% | 90000 | 91595 | 89741 | 1,8 | 1586 |
| 3 | T17+3% | 90000 | 90038 | 89994 | 0,04 | 1550 |
| 4 | T28+36% | 90000 | 90007 | 89964 | 0,01 | 1550 |
| 5 | T36+1,3% | 90000 | 90056 | 89967 | 0,06 | 1496 |
| 0 | (none) | 95000 | 94463 | - | 0 | 1275 |
| 1 | T8 + 32% | 95000 | 96817 | 93404 | 1,9 | 1312 |
| 2 | T15 + 20% | 95000 | 94903 | 92766 | 0 | 1265 |
| 3 | T29 + 51% | 95000 | 94903 | 94903 | 0 | 1275 |
| 4 | T33 + 43% | 95000 | 94995 | 95039 | 0,04 | 1276 |
| 0 | (none) | 100000 | 99846 | - | 0 | 990 |
| 1 | T7: +30% | 100000 | 103955 | 99772 | 3,95 | 1319 |
| 2 | T14: +50% | 100000 | 100765 | 99846 | 0,77 | 1238 |
| 3 | T15: +20% | 100000 | 100259 | 96211 | 0,26 | 1071 |
| 4 | T17: +35% | 100000 | 102479 | 98431 | 2,48 | 1071 |
| 5 | T23: +15% | 100000 | 104724 | 98822 | 4,72 | 1163 |

TABLE VI
SELF-ADAPTATION FOR DIFFERENT EXTRA TIMES

| Case | Time decrease | Deadline [ns] | time | energy [mJ] | Energy without rescheduling [mJ] |
|---|---|---|---|---|---|
| 0 | (none) | 90000 | 89756 | 1494 | 1494 |
| 1 | T17 - 43% | 90000 | 89987 | 1464 | 1494 |
| 2 | T23 - 34% | 90000 | 89649 | 1376 | 1494 |
| 3 | T27 - 15% | 90000 | 89756 | 1494 | 1494 |
| 4 | T30 - 25% | 90000 | 89756 | 1494 | 1494 |
| 5 | T36 - 38% | 90000 | 80785 | 1494 | 1494 |
| 6 | All -30% | 90000 | 88916 | 1116 | 1494 |
| 7 | All -50% | 90000 | 83398 | 998 | 1494 |
| 0 | (none) | 95000 | 94463 | 1275 | 1275 |
| 1 | T7 - 21% | 95000 | 94463 | 1281 | 1275 |
| 2 | T16 - 53% | 95000 | 91810 | 1275 | 1275 |
| 3 | T30 - 48% | 95000 | 94463 | 1275 | 1275 |
| 4 | All -30% | 95000 | 94543 | 1275 | 1275 |
| 5 | All -50% | 95000 | 69690 | 1275 | 1275 |
| 0 | (none) | 100000 | 99846 | 990 | 990 |
| 1 | T7: -30% | 100000 | 99846 | 990 | 990 |
| 2 | T14: -50% | 100000 | 99846 | 990 | 990 |
| 3 | T15: -20% | 100000 | 99433 | 990 | 990 |
| 4 | T17: -35% | 100000 | 99324 | 953 | 990 |
| 5 | T36: -35% | 100000 | 91371 | 953 | 990 |
| 6 | All -30% | 100000 | 81970 | 953 | 990 |
| 7 | All -50% | 100000 | 89570 | 953 | 990 |

[17] S. Deniziak, L. Ciopiński, G. Pawiński, K.Wieczorek and S. Bąk "Cost Optimization of Real-Time Cloud Applications Using Developmental Genetic Programing", Proc. of the 7th IEEE/ACM International Conference on Utility and Cloud Computing, 2014, pp.774-779. http://dx.doi.org/10.1109/UCC.2014.126

[18] K.Sapiecha, L. Ciopiński, and S. Deniziak. "An application of developmental genetic programming for automatic creation of supervisors of multi-task real-time object-oriented systems." IEEE Federated Conference on Computer Science and Information Systems (FedCSIS), 2014. http://dx.doi.org/10.15439/2014F208

[19] Hu, Jingcao, and Radu Marculescu. "Energy-and performance-aware mapping for regular NoC architectures." Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on 24.4 (2005): 551-562. http://dx.doi.org/10.1109/TCAD.2005.844106

[20] Han, Sangchul and Park, Minkyu, Predictability of Least Laxity First Scheduling Algorithm on Multiprocessor Real-Time Systems, Proc. of EUC Workshops, Lecture Notes in Computer Science, vol.4097, 2006, pp.755-764 http://dx.doi.org/10.1007/11807964_76

[21] Sitek, P. "A hybrid CP/MP approach to supply chain modelling, optimization and analysis." Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on. IEEE, 2014. http://dx.doi.org/10.15439/2014F89

DOI: 10.15439/2015F255
ACSIS, Vol. 5

# Simulated annealing with constraints aggregation for control of the multistage processes

Paweł Drąg
Department of Control Systems and Mechatronics
Wrocław University of Technology
Janiszewskiego 11-17, 50-372 Wrocław, Poland
Email: pawel.drag@pwr.edu.pl

Krystyn Styczeń
Department of Control Systems and Mechatronics,
Wrocław University of Technology
Janiszewskiego 11-17, 50-372 Wrocław, Poland
Email: krystyn.styczen@pwr.edu.pl

***Abstract*—In the article the control and optimization of multistage technological processes were discussed. In the presented research, it was assumed, that in the complex technological processes, the multistage differential-algebraic constraints with unknown consistent initial conditions were considered. To rewrite the infinite-dimensional optimal control problem into the finite-dimensional optimization task, the direct shooting method was applied. Simulated annealing algorithm was proposed as the method for solving nonlinear optimization problem with constraints. Stretching function was used to allow us to locate the globally optimal solution. The complex process constraints were treated using constraints aggregation methods. The presented methodology was tested with optimal control problem of the two-reactors system. The numerical simulations were executed in MATLAB environment using Wroclaw Center for Networking and Supercomputing.**

***Index Terms*—optimal control, DAE systems, simulated annealing, constraints aggregation, stretching function.**

## I. Introduction

NOWADAYS, technological processes have been modeled using more general and complex systems of equations. To model technological processes in such branch of industry, like chemical engineering, biotechnology and aerospace engineering, both dynamics and physical conservation laws have been under consideration [2], [3], [8], [16]. Then, a validated model of the process can be applied to control and optimize the technological systems. In the other words, to ensure an efficient and trouble free technological processes, an optimization problem subject to nonlinear differential-algebraic constraints need to be solved [4], [13].

In the last years, deterministic, as well as stochastic optimization algorithms, are adjusted to solve new advanced technological problems. New methods in modeling, algorithmic procedures and globalization of the obtained solutions can be observed in all areas of optimization [9], [10], [26]. In the presented work we would like to indicate the simulated annealing algorithm, which has been seen as one of the main global optimization procedures . In this work the new aspects of the simulated annealing algorithm were given and discussed.

Direct shooting method enables us to transform the optimal control problem into medium- or large-scale nonlinear optimization problem. Then, using aggregation and disaggregation procedures, the considered model can be represented by other system, which reflects the most important features of the original process, but is better adjusted for computational optimization methods [14]. Then, the obtained results can be applied to control the original system.

The presentation of new methodology was performed in the following way. In the next section optimal control with differential-algebraic constraints was introduced. Then, in section III, optimization with aggregated constraints was discussed. ”Stretching” function technique in optimization procedures was presented in section IV. Then, a new simulated annealing algorithm with constraints was given and tested in sections IV and V. The considerations were concluded in section VI.

## II. Optimal control with differential-algebraic constraints

The optimal control algorithms are highly connected with the available modeling methods. The main objective of the control is always the same - we need to find the control function, which optimizes one of the known forms of the process performance index. But the assumptions, conditions and complexity of the considered processes are increasingly challenging and difficult to solve.

In the last years, new particular types of systems have become popular due to many practical applications. There are the multistage dynamical systems and processes with differential-algebraic equations, which are more general, than the pure dynamical processes.

The most important common features of the systems mentioned above, is the presence of the dynamics in the model. The difference, which has a far-reaching consequences in the calculation methods comes from the non-dynamical part of the model.

In the multistage processes, the state variables have to be continuous across the stages. It means, that in the known points, at the interface between the processes, algebraic continuity constraints are introduced. Therefore, the additional algebraic constraints have a pointwise character.

In processes with differential-algebraic constraints, the algebraic relations are justified by physical laws, which take place in the process and are reflected in the model. In this way, both the differential and algebraic equations have a continuous

978-83-60810-66-8/$25.00©2015, IEEE

character and can be treated as a one differential-algebraic system [6].

The presented work concerns the case, when the process is multistage and each stage was described using its own system of differential-algebraic equations. In this way, new theoretical assumptions have to be made and new computational concepts should be designed and implemented.

The features presented above lead us to the optimal control problems in the known form. At first, the process performance index, which can be treated as the measure of the control quality, has to be defined as follows

$$\min_{u(t)} Q = \int_0^{t_F} \mathcal{L}(y(t), z(t), u(t), p, t)dt + \mathcal{E}(y(t_F), z(t_F), t_F) \tag{1}$$

with single system of differential-algebraic constraints in semi-explicit form

$$\begin{array}{rcl} B(t)\dot{y}(t) & = & f(y(t), z(t), u(t), p, t) \\ 0 & = & g(y(t), z(t), u(t), p, t), \end{array} \tag{2}$$

where $y(t) \in \mathcal{R}^{n_y}$ is a differential state, $z(t) \in \mathcal{R}^{n_z}$ is an algebraic state and $u(t) \in \mathcal{R}^{n_u}$ denotes the unknown control function. The independent variable (e.g. time or length of the chemical reactor) is denoted as $t \in \mathcal{R}$. As $p \in \mathcal{R}^{n_p}$ was denoted the vector of global parameters constant in time. The variables in the process are defined by vector-valued functions:

$$f : \mathcal{R}^{n_y} \times \mathcal{R}^{n_z} \times \mathcal{R}^{n_u} \times \mathcal{R}^{n_p} \times \mathcal{R} \to \mathcal{R}^{n_y} \tag{3}$$

and

$$g : \mathcal{R}^{n_y} \times \mathcal{R}^{n_z} \times \mathcal{R}^{n_u} \times \mathcal{R}^{n_p} \times \mathcal{R} \to \mathcal{R}^{n_z}. \tag{4}$$

To ensure, that the system (2) consists of differential-algebraic equations in semi-explicit form, it is assumed, that the matrix $B(t)$ is invertible for all values of $t$.

As it was mentioned, the quality of the process is measured by the value of $Q \in \mathcal{R}$. It is an important assumption, that quality of the whole complex process can be specified by only one real number.

When the multistage processes are considered, then each stage can be described using its own system of differential-algebraic equations

$$\begin{array}{rcl} B^i(t)\dot{y}^i(t) & = & f^i(y^i(t), z^i(t), u^i(t), p, t) \\ 0 & = & g^i(y^i(t), z^i(t), u^i(t), p, t), \\ & & i = 1, \cdots, NS, \end{array} \tag{5}$$

where $NS$ denotes number of the stages in the considered process.

One of the most important progresses in the optimal control methods has been connected with parametrization of the control problem. Because in this way the optimal control problem can be treated as a nonlinear optimization problem, efficient numerical optimization algorithms can be applied.

The parametrization of the control function has been proposed in the article [24] in 1994. Today, the piecewise continuous control parametrization using constant, linear or quadratic functions has been a commonly used approach.

In this way, instead to search for optimal solution using advanced analytical methods, the efficient algorithm of numerical optimization can be applied to obtain the best solution in the assumed class of piecewise continuous functions [7].

The parametrization of the control function in optimal control problems has been often used together with the multiple shooting technique. The multiple shooting method is appropriate to decompose the complex technological processes as well as to stabilize the solutions of both the unstable and highly nonlinear systems.

The multiple shooting method together with the control function parametrization lead us to the direct shooting approach. In this methodology, at the first step, the independent variable domain is divided into the assumed number of intervals

$$t = \bigcup_{i=1}^{NS-1} [t_{i-1} \quad t_i) \cup [t_{NS-1} \quad t_{NS}], \tag{6}$$

where $t_0$ is the time instant, when the process starts and $t_{NS}$ is the final time.

Therefore, the control function as well as the differential-algebraic system can be parametrized in each interval. In practical applications, when the control function is parametrized as piecewise constant, then

$$u^i(t) = u_i, \qquad i = 1, \cdots, NS. \tag{7}$$

The differential-algebraic model of the technological process is parametrized in the sense of the unknown initial conditions. This approach enables us to solve the DAE system using the efficient numerical procedures [17]. The initial conditions of differential-algebraic system can be parametrized in the following way

$$\begin{array}{rcl} y^i(t_{i-1}) & = & s_y^i \\ z^i(t_{i-1}) & = & s_z^i \end{array} \tag{8}$$

for $i = 1, \cdots, NS$.

Therefore, the differential-algebraic constraints have been obtain the new form

$$\begin{array}{rcl} B^i(t)\dot{y}^i(t) & = & f^i(y^i(t), z^i(t), u^i, p, t^i) \\ 0 & = & g^i(y^i(t), z^i(t), u^i, p, t^i), \end{array} \tag{9}$$

with $t^i \in [t_{i-1} \quad t_i]$ and $i = 1, \cdots, NS$. Because the consistent initial conditions are needed to solve the DAE systems, the algebraic part can be extended by damping factor $\alpha$

$$0 = g^i(y^i(t), z^i(t), u^i, p, t^i) + \alpha^i(t) g^i(s_y^i, s_z^i, u^i, p, t^i) \tag{10}$$

for $t^i \in [t_{i-1} \quad t_i]$ and $i = 1, \cdots, NS$.

The process performance index has been rewritten as follows

$$\min_{s_y^i, s_z^i, u^i} \hat{Q} = \sum_{i=1}^{NS} \int_{t_{i-1}}^{t_i} \mathcal{L}(s_y^i, s_z^i, u^i, p, t)dt + \mathcal{E}(s_y^{NS}, s_z^{NS}, t_F). \tag{11}$$

It is worth to note, that the trajectories of the differential states are continuous. It means, that the additional relations have to be incorporated to the model

$$y^i(t_i) = y^{i+1}(t_i) = s_y^{i+1}, \qquad i = 1, \cdots, NS-1, \tag{12}$$

which results as additional pointwise algebraic constraints of the following form

$$y^i(t_i) - s_y^{i+1} = 0, \qquad i = 1, \cdots, NS-1. \tag{13}$$

In the most practical situations, the unknown parameters have their physical interpretation. Among them can be distinguished an unknown volume of the chemical reactor, concentration of the substrates or value of the temperature. The interpretation of the decision variables and technological restrictions defines the feasible region and can be rewritten as the inequality constraints in the form of lower and upper bounds.

Direct shooting approach enables us to transform the optimal control problem into the nonlinear programming problem with the vector of decision variables designed as

$$x = [s_y^i \quad s_z^i \quad u_i \quad p]^T, \tag{14}$$

the process performance index in the form of the objective function

$$\min_x \hat{Q}(x) \tag{15}$$

differential-algebraic constraints

$$\begin{array}{rcl} B^i(t)\dot{y}^i(t) & = & f^i(y^i(t), z^i(t), u^i, p, t^i) \\ 0 & = & g^i(y^i(t), z^i(t), u^i, p, t^i), \end{array} \tag{16}$$

consistency constraints

$$0 = g^i(y^i(t), z^i(t), u^i, p, t^i) + \alpha^i(t) g^i(s_y^i, s_z^i, u^i, p, t^i) \tag{17}$$

continuity constraints

$$y^i(t_i) - s_y^{i+1} = 0, \tag{18}$$

and the constraints of upper and lower bounds type

$$x_L \leq x \leq x_U \tag{19}$$

for $t^i \in [t_{i-1} \quad t_i]$ and $i = 1, \cdots, NS$.

This causes, that the various numerical optimization algorithms can be treated as an important part of the control algorithms [18]. This transformation is a main effect of the direct shooting method applied in the optimal control problems.

## III. Optimization with aggregated constraints

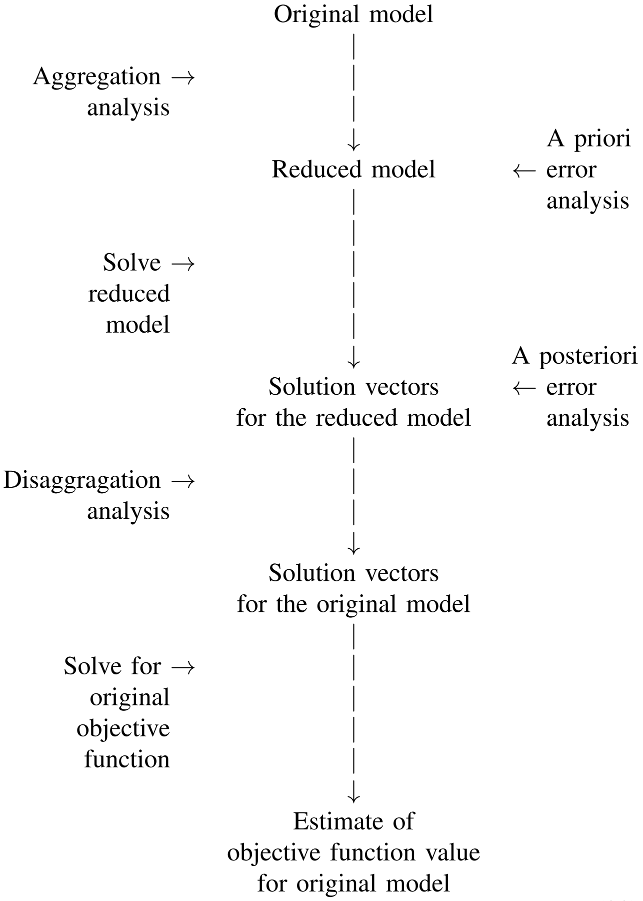
The complexity of the considered models and large number of decision variables makes, that the analytical solutions and some intuitive numerical approaches can be inappropriate in the control of the real-life technological processes.

One of the most important question in modern optimization and control theory is, how to solve the problems with thousands or millions variables and comparable number of equality and inequality constraints. The second important question is, how to compare two unfeasible solutions and to make a decision, which one proposed solution is better than the others.

To analyze the complex process, both the aggregation and disaggregation techniques were used. There is a general methodology, how to obtain useful information about the process by a possibly little model modifications and computations amount [5].

In the previous section, the optimal control problem was reformulated as a middle- or large-scale nonlinear optimization problem. To do this, the direct shooting method was applied. At this point, the consecutive question cannot more wait for the answer - how to treat optimization problems with so many variables and subject to huge number of constraints? The question, which has been just imposed, is considered in this section.

The general way to obtain the useful information about the complex process can be constructed as follows [23]


Original model
Aggregation → analysis
A priori ← error analysis
Reduced model
Solve → reduced model
A posteriori ← error analysis
Solution vectors for the reduced model
Disaggragation → analysis
Solution vectors for the original model
Solve for → original objective function
Estimate of objective function value for original model

(20)

In a general sense, in model aggregation, large-scale optimization models are reformulated as less complex systems. The obtained in this way models reflect the most important features of the original systems, but should be much more suitable to perform numerical simulations.

In the other words, the aggregation technique is a method to specified parts of the model, which can be described using only one single element. At this step, the new single element should be explicitly defined. Therefore, aggregation analysis consists of two levels:

- process of determining, which data are in some sense similar and can be considered as elements in the same group; this step is known as cluster analysis,
- the method of combining the clustered data to define the reduced model.

In contrast to the aggregation analysis, the disaggregation is a method, which can derive the information about the complex model using results obtained by reduced system analysis. The disaggregation analysis, often known as the reversal aggregation analysis, enables us to estimate the solution of the original system using only results obtained from solving the reduced model.

The last stage, which is connected with aggregation and disaggregation methodology, is an error analysis. The error analysis determines the resulting error, which can be introduced by algorithms with aggregate models and disaggregate solutions. From practical point of view, the error analysis can be useful for selection of appropriate aggregation and disaggregation procedures. In the literature two main types of error are defined

a) a priori error bounds, which are the bounds placed upon the optimal value of an original optimization model after aggregation stage, but before the aggregate model is solved,

b) a posteriori error bounds are the bounds placed upon the optimal value of an original optimization model after the aggregated model has been formed and solved.

In practical applications, the iterative optimization algorithms with successive aggregation-disaggregation schemes are used. Such procedures are known as an Iterative Aggregation-Disaggregation (IAD) technique.

As it was mentioned, the direct shooting method allows us to transform the optimal control problem into a nonlinear optimization problem. Therefore, solution of the original infinite-dimensionally task can be obtained by solving finite-dimensional optimization problem. But the question posed in this section does not refer to the optimal control problems. It concerns the much more fundamental task - how to solve medium- and large-scale optimization problems?

It is expected, that the efficient numerical optimization methods enable us to solve optimal control problems.

Therefore, we would like to propose aggregation methods, which can be useful to define new reduced process performance index, as well as reduced constraints.

At this place, we would like to focus on the constraints function aggregation. It was mentioned, that two main types of the constraints need to be considered in the process - there are continuous and pointwise constraints. In particular, it means, that some group of functions are connected with process description by continuous differential-algebraic constraints. There is the second group of constraints, which denotes the additional constraints at the prescribed time instances in the process.

Let us consider the system of the continuous differential-algebraic equations. This particular system can be solved by implementation of efficient numerical algorithms. After that, the process performance index, as well as fulfillment of the constraint can be determined. There is an important question, which refers to consistent initial conditions, which enables us to solve the differential-algebraic system. Although the initial conditions are defined at the specific point, they strictly refer to the continuous process constraints and should be perceived with continuous differential-algebraic constraints. In this work, we refer to this type of constraints as *the consistency constraints* and denote as

$$c_{cons}(\{s_y^i, s_z^i, u^i, p, t_{i-1}\}_{i=1}^{NS}) =$$

$$= \begin{bmatrix} g^1(s_y^1, s_z^1, u^1, p, t_0) \\ g^2(s_y^2, s_z^2, u^2, p, t_1) \\ \vdots \\ g^{NS}(s_y^{NS}, s_z^{NS}, u^{NS}, p, t_{NS-1}) \end{bmatrix} = 0. \qquad (21)$$

The other group of constraints represents the pointwise constraints, which are connected with the fact, that the multistage system is considered. In the other words it means, that the differential state trajectories are continuous across the process stages. The vector of pointwise algebraic constraints denoted as $c_{cont}$ is defined as follows

$$c_{cont}(\{s_y^{i+1}, s_z^{i+1}, u^i, p, t_i\}_{i=1}^{NS-1}) =$$

$$= \begin{bmatrix} y^1(t_1) - s_y^2 \\ y^2(t_2) - s_y^3 \\ \vdots \\ y^{NS-1}(t_{NS-1}) - s_y^{NS} \end{bmatrix} = 0. \qquad (22)$$

In some numerical optimization algorithms, e.g. Sequential Quadratic Programming, it is suggested to treat each constraint function in the same way. Therefore, in each iteration, a large-scale and possibly structured matrix of constraint functions derivatives can be obtained. In general case, when the structure of the process constraints is unknown, the matrix of partial derivatives of the constraints might be time and memory consuming.

If, in general, $c(x)$ denotes

$$c(x) = \begin{bmatrix} c_1(x) \\ c_2(x) \\ \vdots \\ c_{NS}(x) \end{bmatrix} = 0, \qquad (23)$$

then the vector of the constraint functions can be replaced using one of the following real-valued aggregated functions

$$\tilde{c}(x) = \|c(x)\|_1, \qquad (24)$$

$$\tilde{c}(x) = \|c(x)\|_2 \qquad (25)$$

or

$$\tilde{c}(x) = \|c(x)\|_\infty. \qquad (26)$$

One cannot determine, which one measure of the constraints infeasibility has the best features. The constraints aggregation function can be chosen depending on the considered process, optimization methods and, first of all, available computing resources.

The other question is, how to treat the process performance index. Parametrization of the control function, as well as differential and algebraic states, enables us to rewrite the process performance index as the objective function in nonlinear optimization problem. In some applications the penalty function approach and many variations of it can be observed. Advanced optimization methods focus on methods without penalty function [12], [22]. For these reasons, in presented methodology, the objective function $\hat{Q}(x)$, which represents control quality index, was remained unchanged.

## IV. Stretching function

In the previous section it has been indicated, how to apply the direct shooting method to transform the optimal control problem into the nonlinear optimization task. In general, the presented methodology enables us to consider the optimization problem over the feasible region

$$\min_x f(x), \qquad \forall x \in \mathcal{A}, \tag{27}$$

where $f(x)$ is a real-valued objective function

$$f : \mathcal{A} \to \mathcal{R} \tag{28}$$

where $\mathcal{A} \subseteq \mathcal{R}^D$ is the D-dimensional compact set.

The important question is, how to treat the nonlinear optimization problem with the large number of equality, as well as inequality constraints.

To solve the optimization problem on the compact set, the ”stretching” technique for the objective function can be applied [19].

Let us consider the objective function $f(x)$ defined on the compact set. Let the objective function $f$ get the local minimum in a point $x^\star$. Therefore, the point $x^\star$ is a local minimizer of the function $f$. As a consequence, it means, that in the neighborhood $\mathcal{NB}$ of the point $x^\star$, the inequality

$$f(x^\star) \leq f(x) \tag{29}$$

is satisfied for all $x \in \mathcal{NB}$.

In real-life optimization and control problems, it is highly undesirable, that one of the low quality local minima can be treated as the global solution. The ”stretching” function method was designed as a remedy for described situation. This method enables us to alleviate the local minimum using the following two-step transformation.

Let $x^{\hat{\star}}$ be the local minimum of the function $f$, then in the step 1, the function $f(x)$ elevates and all the local minima, which are above the point $f(x^{\hat{\star}})$, disappear

$$\mathcal{G}(x) = f(x) + \frac{\gamma_1}{2}\|x - x^{\hat{\star}}\|\Big(sign(f(x) - f(x^{\hat{\star}})) + 1\Big) \tag{30}$$

In the step 2, the neighborhood of the point $x^{\hat{\star}}$ is stretched upwards, in this way, that it assigns higher values of those points

$$\mathcal{S}\Big(f(x)\Big) = f^{\mathcal{S}}(x) = \mathcal{H}(x) =$$

$$= \mathcal{G}(x) + \frac{\gamma_2\Big(sign(f(x) - f(x^{\hat{\star}})) + 1\Big)}{2\tanh\Big(\mu(\mathcal{G}(x) - \mathcal{G}(x^{\hat{\star}}))\Big)}. \tag{31}$$

The parameters $\gamma_1$, $\gamma_2$ and $\mu$ are the positive constants with arbitrary chosen value. The authors, who proposed the ”stretching” function technique as the global optimization method, suggested the values of the mentioned parameters as $\gamma_1 = 10000$, $\gamma_2 = 1$, $\mu = 10^{-10}$. It was suggested, that the choice of the considered parameters seems not to be critical for the success of the optimization, but they can influence the convergence rate. For these reasons, the parameter tuning was suggested [20].

It is an important question, how the ”stretching” function technique can be applied for optimization problems, when a feasible region is no more a compact set? The mentioned situation can be meet in all cases, when optimization problems with equality constraint are considered [21].

To find the answer for the question posed above, the constraints aggregation methodology is very helpful.

## V. Simulated annealing with constraints

The main ideas connected with simulated annealing algorithm have their origin in thermodynamics. The physical laws, which governs cooling of molten metal in annealing process, have been used to design new optimization methods. After slow cooling in annealing process, the metal tends to reach a state with a low energy. In general, the state with the minimum energy is desired. In analogy to annealing process, the energy represents the possible solution of the optimization problem. Therefore, the state with the minimal energy represents the solution in the optimization algorithm [15], [25].

The simulated annealing algorithm is constructed as follows.

**The simulated annealing algorithm**

```
Initialization $(x_0, T_0)$
Calculation of $f(x_0)$
Until convergence
  Generation of new state $x_1$
  if $f(x_1) < f(x_0)$
    $x_0 = x_1$
  else
    if $\exp\left(\frac{f(x_0)-f(x_1)}{T}\right) > Rand(0,1)$
      Accept new solution $x_0 = x_1$
    else
      Reject new solution
      Decrease the temperature
end (Until)
```

At this moment we would like to introduce new simulated annealing method for optimal control of the multistage differential-algebraic systems. We start the presentation with some remarks.

**Remark 1.** *The optimal control problem with index-1 differential-algebraic constraints using the direct shooting approach has been transformed into nonlinear optimization* *problem with equality constraints of the following form*

$$\begin{array}{c} \min_x \hat{Q}(x) \\ c_{cons}(x) = 0 \\ c_{cont}(x) = 0. \end{array} \tag{32}$$

**Remark 2.** *The quality $\mathcal{Q}$ of any proposed solution $x$ can be specified using a triple*

$$\mathcal{Q}(x) = \begin{bmatrix} q_1(x) \\ q_2(x) \\ q_3(x) \end{bmatrix} = \begin{bmatrix} \hat{Q}(x) \\ c_{cons}(x) \\ c_{cont}(x) \end{bmatrix}. \tag{33}$$

**Remark 3.** *The "stretching" function technique applied for some proposed solution transforms the vector, which parametrized the quality of the proposed solution*

$$\mathcal{S}\Big(\mathcal{Q}(x)\Big) = \mathcal{Q}^S(x) = \begin{bmatrix} q_1^S(x) \\ q_2^S(x) \\ q_3^S(x) \end{bmatrix} = \begin{bmatrix} \hat{Q}^S(x) \\ c_{cons}^S(x) \\ c_{cont}^S(x) \end{bmatrix}. \tag{34}$$

The decisions of the new solution acceptance can be made by quality comparison of the considered solutions

**Definition 4.** *Any two values $\mathcal{Q}_a$ and $\mathcal{Q}_b$ are in the relation $\oplus$ if*

$$\begin{array}{c} q_1(a) \oplus q_1(b) \\ q_2(a) \oplus q_2(b) \\ q_3(a) \oplus q_3(b) \end{array} \tag{35}$$

**Theorem 5.** *Let $x_0$ and $x_1$ be any two solutions. The solution $x_0$ is unconditionally better than solution $x_1$ if and only if*

$$\mathcal{Q}(x_0) < \mathcal{Q}(x_1). \tag{36}$$

Let us transform the triple, which defines the quality of the solution, using the "stretching" function technique.

**Theorem 6.** *Let $x_0$ and $x_1$ be any two solutions. If*

$$\begin{array}{c} q_1^S(x_0) < q_1^S(x_1) \\ q_2^S(x_0) < q_2^S(x_1) \\ q_3^S(x_0) < q_3^S(x_1) \end{array} \tag{37}$$

*then*

$$\mathcal{Q}(x_0) < \mathcal{Q}(x_1). \tag{38}$$

The question is, how to compare two solutions, which are not in the relation presented in Theorem 6.

One of the most important features of simulated annealing algorithm is a possibility of accepting a step, which does not improve the current solution. Therefore, to compare any two solution, which are not in clear relation, a conditional acceptance mode can be applied. In simulated annealing methodology, it is known as *an acceptance with probability.* In the presented work, because of three-dimensional quality solution index $\mathcal{Q}(x)$ under considerations, a new extended method of acceptance with probability was proposed.

**Proposition 7.** *Let $x_0$ be the current solution and $x_1$ be any other solution generated by simulated annealing algorithm from the neighborhood of the point $x_0$, and $T$ denote the value of the temperature in the current iteration. If the point $x_1$ cannot be unconditionally accepted as a new solution, then it can be accepted with the probability $p \in \mathcal{N}(0,1)$. The acceptance probability is dependent on the index quality in the following way*

$$\exp\Big( - \frac{\hat{Q}^S(x_1) - \hat{Q}(x_0)}{T} \Big) > p$$

*or*

$$\exp\Big( - \frac{c^S_{cons}(x_1) - c_{cons}(x_0)}{T} \Big) > p$$

*or*

$$\exp\Big( - \frac{c^S_{cont}(x_1) - c_{cont}(x_0)}{T} \Big) > p.$$

There are two important remarks connected with the presented acceptance condition

**Remark 8.** *The multiple conditions increase the general acceptance probability of the worse solution.*

**Remark 9.** *Application of the "stretching" function technique transforms the bad solution in the worse one. Moreover, "stretching" function technique supports searching for better solutions and alleviate the local minima.*

The presented simulated annealing "stretching" function algorithm with new conditional solution acceptance approach was applied to solve the optimal control problem of the multistage rectors system.

## VI. Numerical examples

The presented methodology was tested on the optimization problem of the three-stage technological process. The presented system consists of two chemical reactors with mixing [24]. The sketch of the process was presented in the Fig. 1.

At the beginning, the first reactor was loaded with the substrate $A$ with the volume 0.1 m$^3$ and concentration 2000 mol/m$^3$. Due to reactions taking place in the system, the products $B$ and $C$ are obtained according to the following scheme

$$2A \to B \to C. \tag{39}$$

Additionally, the first reactor was equipped with a heating exchanger, which can be used to control the process temperature and in this way - to influence the trajectories of the process variables. The concentrations of the substrate and products were changing in the following way

$$\dot{C}_A = -2k_1(T)C_A^2 \tag{40}$$

$$\dot{C}_B = 2k_1(T)C_A^2 - k_2(T)C_B \tag{41}$$

$$\dot{C}_C = k_2(T)C_B \tag{42}$$

with the kinetics constraints

$$k_1(T) = 0.0444\exp(-2500/T) \tag{43}$$

and

$$k_2(T) = 6889.0\exp(-5000/T). \tag{44}$$

Then, in the mixing stage at the time $t_1$, the component $B$ of concentration $C_B^0 = 600$mol/m$^3$ and some volume $S$ was added. Therefore, the volume and consecrations of the substrates were changing, so the following relations were satisfied

$$V_2C_A(t_2^0) = V_1C_A(t_1) \tag{45}$$

$$V_2C_B(t_2^0) = V_1C_B(t_1) + SC_B^0 \tag{46}$$

$$V_2 C_C(t_2^0) = V_1 C_C(t_1) \tag{47}$$

where $V_1$ is the volume of substrates loaded at the beginning of the first reactor. Therefore, the volume $V_2$ in the second reactor was given by

$$V_2 = V_1 + S \tag{48}$$

The volume $S$ is a decision parameter with

$$0 \le S \le 0.1(m^3) \tag{49}$$

After the mixing stage, the substrates were loaded into the last reactor, where three reactions were taking a place

$$B \to D \tag{50}$$

$$B \to E \tag{51}$$

$$2B \to F \tag{52}$$

In the $2nd$ reactor, the reactions take place under isothermal conditions. The state variables are changing in the following way

$$\dot{C}_A = 0 \tag{53}$$

$$\dot{C}_B = -0.02 C_B - 0.05 C_B - 2 \times 4.0 \times 10^{-5} C_B^2 \tag{54}$$

$$\dot{C}_C = 0 \tag{55}$$

$$\dot{C}_D = 0.02 C_B \tag{56}$$

$$\dot{C}_E = 0.05 C_B \tag{57}$$

$$\dot{C}_F = 4.0 \times 10^{-5} C_B^2 \tag{58}$$

The combined processing time for both reactors is equal to 180 min

$$t_1 + t_2 = 180, \tag{59}$$

$$t_1 > 0, \qquad t_2 > 0. \tag{60}$$

The decision variables are the profile of the temperature $T(t)$, the duration time of the reactions in each stage, and the amount $S$ of component $B$, which is added at the mixing step.

The process is aimed to maximize the amount of the product $D$ at the output of the $2nd$ reactor

$$\max_{t_1, t_2, S, T(t)} V_2 C_D(t_2) \tag{61}$$

subject to the constraints on the temperature profile

$$298 \le T(t) \le 398(K), \qquad t \in [0 \quad t_1]. \tag{62}$$

The direct shooting method enables us to transform the multistage optimal control problem as an nonlinear optimization problem with both continuous and pointwise constraints. In classical approaches, the nonlinear optimization problem with constraints can be solved using penalty function approach or NLP algorithms, like Sequential Quadratic Programming or Barrier methods. In this way, the process in first reactor was divided into 10 equidistance intervals. Additionally, the initial conditions for the second reactor should be consistent with the results of the mixing stage. Therefore, the considered NLP problem consisted of 27 continuity constraints, 3 consistency constraints and 42 decision variables.

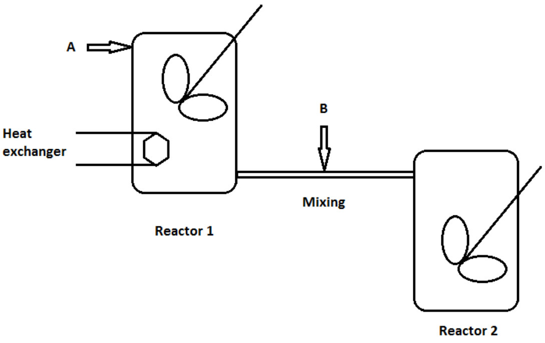


Fig. 1. System of the two-reactors with mixing.

The aggregated model was defined using 3 scalar functions

$$\hat{Q}(x) = Q(x) \tag{63}$$

$$c_{cont}(x) = \sum_{i=1}^{27} \|c_{cont,1}(x)\|_1 \tag{64}$$

$$c_{cons}(x) = \sum_{i=1}^{3} \|c_{cons,1}(x)\|_1 \tag{65}$$

In the next step, all three function were transformed using "stretching" function technique

$$\mathcal{S}\Big(\hat{Q}(x)\Big) = \hat{Q}^{\mathcal{S}}(x) \tag{66}$$

$$\mathcal{S}\Big(c_{cont}(x)\Big) = c_{cont}^{\mathcal{S}}(x) \tag{67}$$

$$\mathcal{S}\Big(c_{cons}(x)\Big) = c_{cons}^{\mathcal{S}}(x) \tag{68}$$

and optimized using presented simulated annealing algorithm with new acceptance method.

The parameters of the methods were adjusted automatically. Therefore, the simulations were performed with the following parameters

- number iterations in the outer loop $MaxIter = 500$,
- inner iterations in cooling loop $MaxIterCool = 10$,
- diameter of the considered neighborhood $\alpha = 0.4$
- initial temperature $T_0 = 100$,
- temperature update factor $T_{i+1} = \beta T_i$ is $\beta = 0.6$.

The obtained results, which were obtained for the aggregated model, were applied to the original model of the considered process. After 702 model evaluations, the final value of the original process performance index was equal to 22.7233 mol. The obtained state trajectories were presented in the Fig. 2.

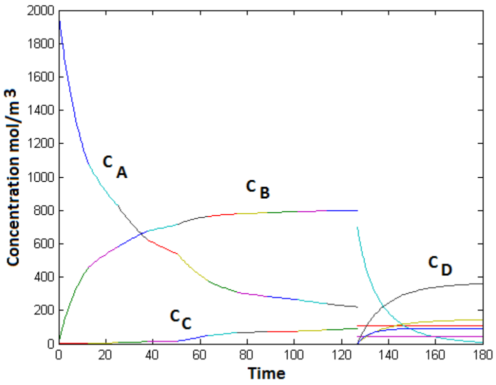


Fig. 2. Trajectories of the state variables.

## VII. Conclusion

In the article a problem of solving complex multistage technological processes was considered. The new methodology, which was presented, applies aggregation-disaggregation approach. The optimal control problem was transformed into the nonlinear optimization problem using the direct shooting approach. Then, the NLP problem has been defined using only three functions, which reflects the characteristics features of the process. The first one denotes the value of the process performance index. Then, the second one is connected with the continuity of the differential state trajectories, and the last one denotes the consistent initial conditions of the process. The quality of the proposed solution was defined using only values of these three functions. To compare any two proposed solutions, the new decision approach was proposed. The presented methodology was applied to the two-stage reactor system.

The future research will be devoted to error analysis of the aggregation-disaggregation procedures for control and optimization of the multistage technological processes with nonlinear differential-algebraic constraints.

## Acknowledgment

Many thanks are due to two anonymous referees, whose constructive suggestions led to improvement of the original article.

The project was supported by the grant Młoda Kadra B40099/I6 at Wrocław University of Technology.

## References

[1] J.R. Banga, E. Balsa-Canto, C.G. Moles, A.A. Alonso. 2005. Dynamic optimization of bioprocesses: Efficient and robust numerical strategies. Journal of Biotechnology. 117:407-419, http://dx.doi.org/10.1016/j.jbiotec.2005.02.013

[2] J.T. Betts. 2010. Practical Methods for Optimal Control and Estimation Using Nonlinear Programming, Second Edition. SIAM, Philadelphia, http://dx.doi.org/10.1137/1.9780898718577

[3] L.T. Biegler. 2010. Nonlinear Programming. Concepts, Algorithms and Applications to Chemical Processes. SIAM, Philadelphia, http://dx.doi.org/10.1137/1.9780898719383

[4] L.T. Bielger, S. Campbell, V. Mehrmann. 2012. DAEs, Control, and Optimization. Control and Optimization with Differential-Algebraic Constraints. SIAM, Philadelphia, http://dx.doi.org/10.1137/9781611972252.ch1

[5] K.F. Bloss, L.T. Biegler, W.E. Schiesser. 1999. Dynamic process optimization through adjoint formulations and constraint aggregation. Ind. Eng. Chem. Res. 38:421-432, http://dx.doi.org/10.1021/ie9804733

[6] K.E. Brenan, S.L. Campbell, L.R. Petzold. 1996. Numerical Solution of Initial- Value Problems in Differential-Algebraic Equations. SIAM, Philadelphia, http://dx.doi.org/10.1137/1.9781611971224

[7] M. Diehl, H.G. Bock, J.P. Schlöder, R. Findeisen, Z. Nagy, F. Allgöwer. 2002. Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations. Journal of Process Control. 12:577-585, http://dx.doi.org/10.1016/S0959-1524(01)00023-3

[8] P. Drąg, K. Styczeń. 2012. A Two-Step Approach for Optimal Control of Kinetic Batch Reactor with electroneutrality condition. Przeglad Elektrotechniczny. 6:176-180.

[9] S. El Moumen, R. Ellaia, R. Aboulaich. 2011. A new hybrid method for solving global optimization problem. Applied Mathematics and Computation. 218:3265-3276, http://dx.doi.org/10.1016/j.amc.2011.08.066

[10] R. Faber, T. Jockenhövel, G. Tsatsaronis. 2005. Dynamic optimization with simulated annealing. Computers and Chemical Engineering. 29:273-290, http://dx.doi.org/10.1016/j.compchemeng.2004.08.020

[11] S. Fidanova, M. Paprzycki, O. Roeva. 2014. Hybrid GA-ACO Algorithm for a Model Parameters Identification Problem. 2014 Federated Conference on Computer Science and Information Systems. 413-420, http://dx.doi.org/10.15439/2014F373

[12] R. Fletcher, S. Leyffer. 2002. Nonlinear programming without a penalty function. Mathematical Programming. 91:239-269, http://dx.doi.org/10.1007/s101070100244

[13] M. Gerdts. 2003. Direct Shooting Method for the Numerical Solution of Higher-Index DAE Optimal Control Problems. Journal of Optimization Theory and Applications. 117:267-294, http://dx.doi.org/10.1023/A:1023679622905

[14] M. Jeon. 2005. Parallel optimal control with multiple shooting, constraints aggregation and adjoint methods. J. Appl. Math. and Computing. 19:215-229, http://dx.doi.org/10.1007/BF02935800

[15] M. Ji, Z. Jin, H. Tang. 2006. An improved simulated annealing for solving the linear constrained optimization problems. Applied Mathematics and Computation. 183:251-259, doi:10.1016/j.amc.2006.05.070

[16] M. Kwiatkowska. 2015. DAEs method for time-varying indoor air parameters evaluation. In: A. Kotowski, K. Piekarska, B. Kaźmierczak (eds.) Interdyscyplinarne zagadnienia w inżynierii i ochronie środowiska T. 6. Wrocław 2015, pp. 214-220.

[17] D.B. Leineweber, I. Bauer, H.G. Bock, J.P. Schlöder. 2003. An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part 1: theoretical aspects. Computers and Chemical Engineering. 27:157-166, http://dx.doi.org/10.1016/S0098-1354(02)00158-8

[18] J. Nocedal, S.J. Wright. 2006. Numerical Optimization. Second Edition. Springer, New York, http://dx.doi.org/10.1007/978-0-387-40065-5

[19] K.E. Parsopoulos, V.P. Plagianakos, G.D. Mogoulas, M.N. Vrahatis. 2001. Improving the particle swarm optimizer by function ”stretching”, in: N. Hadjisavvas, P.M. Pardalos (eds.), Advances in Convex Analysis and Global Optimization, Kluwer Academic Publishers. 445-457, http://dx.doi.org/10.1007/978-1-4613-0279-7_28

[20] K.E. Parsopoulos, V.P. Plagianakos, G.D. Mogoulas, M.N. Vrahatis. 2001. Objective function ”stretching” to alleviate convergence to local minima. Nonlinear Analysis. 47:3419-3424, http://dx.doi.org/10.1016/S0362-546X(01)00457-6

[21] K.E. Parsopoulos, M.N. Vrahatis. 2002. Particle Swarm Optimization Method for Constrained Optimization Problems, in: P. Sincak et al.(Eds.), Intelligent Technologies - Theory and Applications: New Trends in Intelligent Technologies, IOS Press, pp. 214-220.

[22] E. Rafajłowicz, K. Styczeń, W. Rafajłowicz. 2012. A modified filter SQP method as a tool for optimal control of nonlinear systems with spatio-temporal dynamics. Int. J. Appl. Math. Comput. Sci. 22:313-326, http://dx.doi.org/10.2478/v10006-012-0023-8

[23] D.F. Rogers, R.D. Plante, R.T. Wong, J.R. Evans. 1991. Aggregation and disaggregation techniques and methodology in optimization. Operations Research. 39:553-582, http://dx.doi.org/10.1287/opre.39.4.553

[24] V.S Vassiliadis, R.W.H. Sargent, C.C. Pantelides. 1994. Solution of a Class of Multistage Dynamic Optimization Problems. 1. Prob-

lems without Path Constraints. Ind. Eng. Chem. Res. 33:2111-2122, http://dx.doi.org/10.1021/ie00033a014

[25] Y.-J. Wang, J.-S. Zhang. 2007. An efficient algorithm for large scale global optimization of continuous functions. Journal of Computational and Applied Mathematics. 206:1015-1026, http://dx.doi.org/10.1016/j.cam.2006.09.006

[26] K.F.C. Yiu, Y. Liu, K.L. Teo. 2004. A hybrid descent method for global optimization. Journal of Global Optimization. 28:229-238, http://dx.doi.org/10.1023/B:JOGO.0000015313.93974.b0

DOI: 10.15439/2015F148
ACSIS, Vol. 5

# Flow design in photonic data transport network

Mateusz Dzida
Optimax
ul. Wolbromska 19/A, 03-680 Warszawa, Poland
Email: mdzida[at]onet.eu

Andrzej Bąk
Optimax
ul. Wolbromska 19/A, 03-680 Warszawa, Poland
Email: abak[at]poczta.pl
Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warszawa, Poland
Email: bak[at]tele.pw.edu.pl

*Abstract*—**Development of sophisticated photonic transmission systems enabled evolution of photonic data transport networks towards cost-efficient and energy-efficient platforms capable to carry enormous traffic. Given access to technologically advanced equipment, network operator faces a series of decision problems related to how to efficiently use this technology. In this paper, we propose a functional model of modern photonic network with wavelength division multiplexing (WDM). Proposed functional model is a basis for formulating corresponding flow design problem in terms of mathematical programming.**

## I. Introduction

RECENT advances in the photonic networking enabled rapid growth of the transmission rates in the modern photonic data transport networks. Thus, photonic data transport networks became considerable alternative for traditional electric-based transmission systems, and are more and more widely deployed in the Autonomous Systems composing the Internet.

An important area of research in the domain of photonic data transport networks is associated with development of functional models of photonic networks and mathematical models of the decision problems associated with designing such networks.

Modeling telecommunication network is not a trivial task. On one hand, development of functional models requires detailed knowledge of transmission technology and networking protocols. Thus, such functional models must be sufficiently detailed to accurately represent costs of equipment. On the other hand, it is necessary to consider a number of specific aspects that may further turn into design constraints. Most important of these include:

- architecture of switching equipment and transmission,
- mechanisms for traffic grooming and consolidation,
- network reconfiguration in case of failure,
- physical effects associated with signal propagation.

In the balance of this paper we propose mathematical (optimization) model of the flow design problem related to photonic data transport network. Developed model is expressed in terms of integer programming. It refers to generic input data, including: network topology, infrastructure, and considered products. In particular, on one hand, input data must determine full cost characteristics of the considered equipment, usually in form of cost of particular expansion cards. On the other hand, input data are supposed to include locations of client devices and their demand for data transport services. It is therefore assumed that knowledge possessed by a network operator about demand structure is certain. In practice, knowing exact demand for transport services can be difficult, and sometimes even impossible. Still, we assume that through appropriate statistical methodology, it is possible to determine demand value with reasonable degree of confidence.

Paper is organized as follows. In Section II we present general principles of photonic data transport networks, including: standardized interfaces (Section II-A), architecture (Section II-B) and configuration (Section II-B) of photonic devices, signal impairments (Section III-F), performance aspects (Section II-E), and cost model (Section II-F). All elements presented throughout Section II define functional model of photonic network. Further, in Section III we define corresponding mathematical model associated with designing flows in photonic data transport network. Assumptions and construction of network graph are discussed in Sections III-A and III-B, respectively. Considered flow design problem is formulated as mixed-integer programme in Section III-C. Further, we investigate modeling specific aspects of the photonic data transport networks, related to: consistency of client flow at technology level (Section III-D), redundancy (Section III-E), impairments (Section III-F), and network cost (Section III-G). Paper is summarized with estimation of formulation complexity in Section IV, and conclusions in Section V.

## II. Photonic data transport networks

Photonic data transmission exploits optical tracks (fibers) to send photonic signals between transceivers (lasers) and receivers (photo-diodes). Photonic signal is modulated to represent values of consecutive bits, composing transmitted piece of data. Low interference with other electromagnetic signals makes photonic signals stable and robust to distortions. Thus, photonic signals can be successfully sent over long distances, reaching thousands of kilometers, and transmitted messages can be still reproduced with small error rate.

This work presents results of realization of first phase of ongoing project "Optymalizacja przepływów w wielowarstwowych sieciach dostępu do Internetu w technologii DWDM/IP/MPLS" (ODIN). Project is co-founded by The National Centre for Research and Development, Poland.

978-83-60810-66-8/$25.00©2015, IEEE

Photonic signals can be easily multiplexed in the frequency domain. Frequency in the case of photonic networking is called wavelength, and the corresponding multiplexing is called Wavelength Division Multiplexing or WDM. WDM is also a term describing the related channel-oriented transport technology, exploiting regular channel widths and spacing (so-called optical grid [1]). Each wavelength constitutes thus an isolated communication channel, in the following referred to as $\lambda$.

### A. Optical Transport Network

G.709 [2] is an ITU-T recommendation defining interfaces of Optical Transport Network (OTN), defined in another ITU-T recommendation G.872 [3]. It characterizes so-called Optical Transport Hierarchy (OTH), i.e., structured hierarchy of interfaces that can be treated as an extension of the SDH hierarchy, defined in ITU-T recommendation G.707 [4]. Interfaces defined in G.709 are considered at two levels: user-to-network (UNI) and network-to-network (NNI). UNI interface is exposed towards an OTN client, and NNI interface is exposed towards an OTN network.

According to recommendation G.709, the basic OTH information structure is called Optical Channel (OCh). OCh spans between line ports of each pair of adjacent transponders, responsible for converting electrical data signal into photonic signal, modulated as single wavelength (and vice versa). Each OCh is associated with out-of-band signaling channel used for management and supervision of optical channel. OCh payload is filled with lower layer information structure called Optical Transport Unit (OTU). Similarly as OCh, OTU spans between two adjacent transponders, but it is associated with data connection in the electric domain. OTU signaling creates a control channel to exchange information related to supervision and conditioning signal for transport (including Forward Error Correction). Payload of OTU is further filled with information structure called Optical channel Data Unit (ODU). ODU is a function of termination point of digital end-to-end data path between two optical devices, crossing amplification and regeneration sections. ODU signaling is used to exchange control information related to end-to-end path supervision and monitoring so-called tandem connections.

Finally, client signal is mapped into information structure called Optical Payload Unit (OPU). OPU structure is responsible for the adaptation of the client signals and lower layer ODU signals. OPU overhead comprises information required to perform rate adaptation between the client signal rate and the OPU payload rate, and other OPU overheads supporting the client signal transport. OPU structure is filled into the ODU payload.

A number of the OCh structures is multiplexed into structure called Optical Multiplex Section (OMS). OMS is a function of termination point of the (de-)multiplexing section. OMS is further filled into the payload of structure called Optical Transport Section (OTS), responsible for transporting multiplexed optical channels between two adjacent amplification points through single fiber span. OTS is a function of termination point of optical line which transports multiplexed optical channels from one node to another. OTN layers are presented in Figure 1. In the following, the introduced notation is used to map particular OTN functions into generic types of photonic cards needed to build OTN compatible networks.

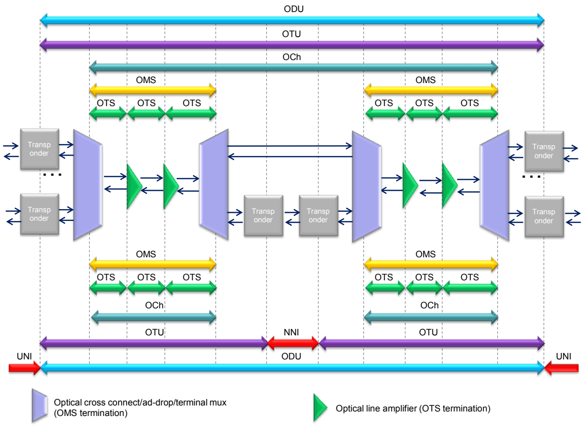


Fig. 1. OTN layers

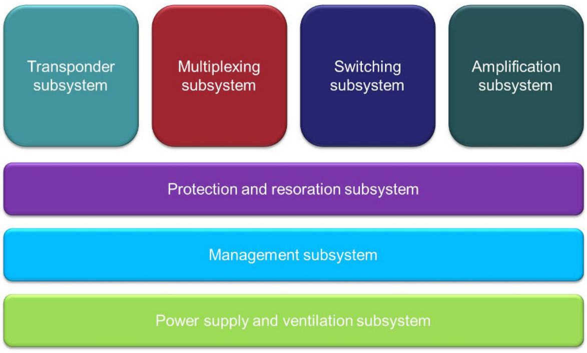


Fig. 2. ONE architecture

### B. Architecture of optical network element

Architecture of Optical Network Element (ONE) is composed of seven functional blocks presented in Figure 2. Each of these functional blocks is briefly described in the following.

*1) Transponder subsystem:* Transponder subsystem is a functional block responsible for adapting client black&white signals, typically in $1310nm$ band, to colorful $\lambda$ signals within $1550nm$ band, according to used optical grid definition. Adaptation concerns both signal frequency and signal rate. Transponder is typically available in the form of expansion card, required for each add-drop channel. It provides UNI interface, and optionally NNI interface. Transponders are sometimes combined with Forward Error Correction blocks, that allow to correct some transmission errors, and decrease bit error rate.

Some vendors combine transponders with concatenation function, which allows to concentrate a number of low order speed signals into high order speed signal. For instance, four STM-16 ($2.5Gbps$) signals can be concentrated into one

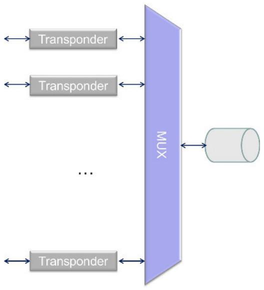


Fig. 3. Multiplexing subsystem: 1-stage multiplexer

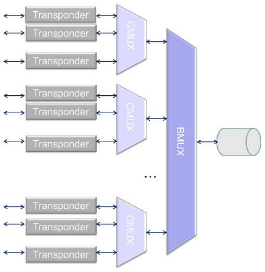


Fig. 4. Multiplexing subsystem: 2-stage multiplexer

ODU2 signal ($10.709 Gbps$). Sometimes, concentrators are provided as stand-alone cards, further connected to transponder tributary interfaces. For instance, two GigabitEthernet signals ($1 Gbps$ each) using Generic Framing Procedure (GFP) can be concentrated into one STM-16 signal ($2.5 Gbps$), further combined with mentioned $4 \times 2.5 Gbps$ transponder card. Such concentrating stand-alone card is called concentrator. Using concentrating function allows network operator to maintain fine grain granularity of transmitted signals, and optimized usage of the WDM network resources.

*2) Multiplexing subsystem:* Multiplexing subsystem is responsible for multiplexing multiple colorful add $\lambda$ signals produced by transponders into single multi-channel signal, according to definition of used optical grid. Similarly, demultiplexer separates particular drop channels from single multi-channel WDM signal. Thus, tributary side of multiplexing subsystem expose a number of physical interfaces towards transponder subsystem and single line interface towards switching subsystem.

For the sake of flexibility, vendors of optical devices often use multi-stage multiplexers. as those presented in figures 3 and 4. In 2-stage multiplexer, presented in Figure 4, multiplexing is decomposed in two stages: band multiplexing (BMUX) and channel multiplexing (CMUX). For instance, $N = 80 \times \lambda$ loading plan, constructed according to 50GHz optical grid, can be decomposed into ten bands, each containing eight channels. Cost of 2-stage multiplexing with expandable CMUX cards can be more granular than 1-stage multiplexing. In particular, network operator may add new CMUX cards as its demand for capacity grows.

*3) Switching subsystem:* Switching subsystem is responsible for directing selected channels to the add-drop multiplexing block and creating pass-through connections (between different ONE directions) for the other channels. In the pass-through mode, channel in one adjacent fiber is switched to channel in another adjacent fiber. Example 4-direction switching module is presented in Figure 5. Presented module is based on using Wavelength Selective Switches (WSS) and Optical SPlitters (SPL).

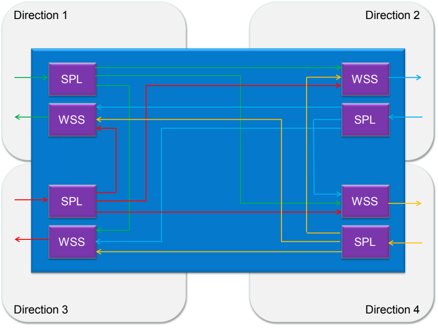


Fig. 5. Switching subsystem

*4) Amplification subsystem:* Although optical signals do not interfere with other electromagnetic signals, power of optical signals is dispersed due to impurity of transmission medium. Thus, power of transmitted signal in the source node is much higher than power of received signal in the destination node. To assure right power level received by photo-detectors, optical signal may need to be amplified. Today, there are three types of optical amplifiers in common use:

- Semiconductor optical amplifier,
- Raman fiber amplifier,
- Erbium doped fiber amplifier.

Semiconductor Optical Amplifier (SOA), in a similar manner as Fabry-Perot laser diodes, uses a semiconductor to provide the gain medium. Signal amplification is realized in SOA by inducing energetic level of medium material. SOA amplifiers need anti-reflection components at the end faces, so amplifier not turns into a laser diode. SOA amplifiers in practice operate at signal wavelengths between $0.4 \mu m$ and $2.0 \mu m$, and generate gain up to $30 dB$. Compared with other amplifier types (mainly EDFA) SOA may generate higher noise, lower gain, moderate polarization dependence, and high non-linearity with fast transient time. Still, SOA amplifiers may have compact design, easily plug-able into fiber-pigtailed components. Despite relatively high noise level, high amplification and wide amplification spectrum make SOA amplifiers attractive option in certain optical applications. In particular, SOA amplifiers are widely used as wavelength convertors.

Raman Fiber Amplifier (RFA) exploits Raman distortion effect, i.e., nonlinear interaction between signal and pump laser within an optical fiber. In RFA amplifier, pump laser generates optical signal in lower range of optical spectrum ($1535 nm$) that is further coupled with operational signal in higher optical spectrum ($1540 nm$ – $1580 nm$). In other words, RFA amplifier induces energy transfer from optical signal generated by pump laser to operational signal. RFA amplifier in practice generates low noise level, has high amplification level, and can produce high power signal. Construction of RFA amplfier may be relatively complex to provide smooth and wide gain characteristics, i.e., multiple pumping signals with different frequencies may be required.

Erbium Doped Fiber Amplifier (EDFA) is likely most widely used type of amplifiers. EDFA amplifier is composed of a short Erbium doped fiber span (typically several kilometers) and pump laser. Optical signal generated by the pump laser is absorbed by Erbium atoms, that in turn release absorbed energy to the operational signal transmitted through doped fiber. EDFA amplifiers in practical applications use $980nm$ – $1480nm$ pumping signal to amplify operational signals transmitted at wavelength $1550nm$. EDFA amplifiers in a process of Amplified Spontaneous Emission (ASE) generate noise that further decrease signal-to-noise ratio.

*5) Protection subsystem:* Typically, data stream, transmitted through WDM network, may be protected by parallel optical channel spreading between the same type of transponders in the same end devices as the nominal channel. For this purpose, vendors provide special protection cards responsible for activating protection channel in case of unavailability of the nominal channel. Such cards can work in one of two commonly used modes:

- source switched,
- destination switched.

Source switch mode is realized by turning off laser in the nominal channel, and immediate activation of laser in the protection channel, as only failure is detected. Source switch is thus completely transparent to carried data, and directly switches optical signals. Besides, source switch mode allows to use simple optical couplers at receiver side, because only one signal: nominal or protection is transmitted at time.

Destination switch is realized as data switch at receiver side. Switch continuously compares quality of nominal and protection signals, and according to result of this comparison passes one of them to client interface. It means that lasers of both signals must be active all the time, and data stream must be replicated to both channels at source.

*6) Management subsystem:* According to OTH hierarchy, overhead information can be sent in-band only to certain layer defined by OTU. Below OTU layer, signaling is sent out-of-band through so-called Optical Supervisory Channel (OSC). OSC organization is proprietary.

### *C. ONE configuration*

Basic element of the WDM network is device called Optical Add-Drop Multiplexer (OADM). In general, OADM is responsible for multiplexing and concatenating tributary signals into line signals transmitted towards other devices. Depending on its configuration and design, optical devices can be used in different fashions. Typical optical device configurations include:

1) Line terminal (LT),
2) Back-to-back (BtB),
3) Band optical add-drop multiplexer (BOADM) – 2-stage multiplexer only,
4) Reconfigurable optical add-drop multiplexer (ROADM),
5) Tunable and reconfigurable optical add-drop multiplexer (T&ROADM).

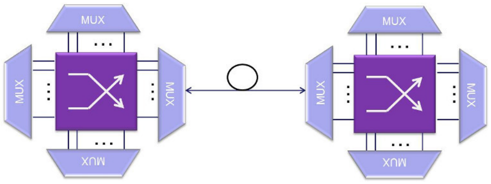


Fig. 6. Opaque mode

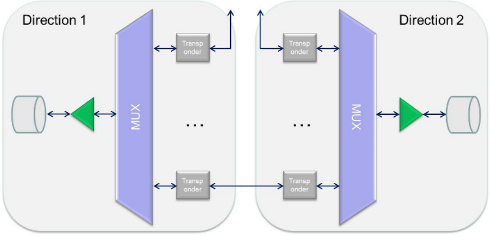


Fig. 7. Back-to-back configuration

Below, we briefly characterize each of these configurations.

1. Line terminal is the most basic configuration of an optical device. Device configured as a line terminal is responsible for adopting and modulating tributary signals as colorful signals and multiplexing resulting signals into $\lambda$ channels according to selected optical grid definition. Line terminals are commonly used in so-called opaque mode (see Figure 6), to build simple point-to-point optical spans. In opaque mode, device receiving optical signal transforms it to the electric domain and switches obtained electrical signals. Opaque mode is common configuration of the SDH/Sonet devices. Line terminal can be equipped with transponders, concentrators, multiplexers, and amplifiers.

2. Back-to-back configuration allows to create simple multi-point optical networks, by static coupling tributary ports in transponders. Back-to-back configuration is presented in Figure 7.

3. Band optical add-drop multiplexer is an extension of back-to-back configuration which additionally allows subset of band signals to be transmitted in pass-through mode directly between band multiplexers (BOADM configuration, presented in Figure 8, refers to 2-stage add-drop multiplexers.)

4. Reconfigurable optical add-drop multiplexer is a configuration of optical device capable to add, drop, and pass-through wavelengths (switch between adjacent fibers). It can be (remotely) configured (through management system) to put each channel within line fiber into one of two states: pass-through or add-drop (see Figure 9).

5. Tunable and reconfigurable optical add-drop multiplexer is the most advanced configuration that extends remote configuration of ROADM with wavelength conversion function.

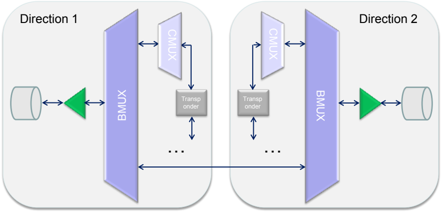


Fig. 8. Band optical add-drop multiplexer configuration

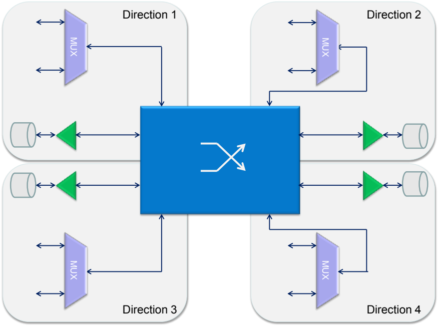


Fig. 9. Remote optical add-drop multiplexer configuration

### D. Optical impairments

Fiber attenuation is the most fundamental impairment that affects optical signal propagation. Attenuation is a fiber property resulting from using various material, structural, and modular impairments. Still, fiber attenuation is an effect with intensity linearly proportional to fiber length and number of optical elements: connectors, splitters, etc. Thus, optical signal power lost due to fiber attenuation can be recovered by using so-called Linearized Optical Fiber Amplifier (LOFA) cards, containing amplification modules described in Section II-B4. Beyond amplifiers itself, LOFA cards may also contain a module called Variable Optical Attenuator (VOA), responsible for enforced attenuation of optical signal, so volume of power received by photo-diode is contained in strictly defined working window. Attenuation of VOA modules is typically dynamically set in a closed-loop feedback between adjacent ONE's. Unfortunately, amplifiers, as all active optical elements, introduce into transmitted signal some portion of noise. Fiber attenuation is well standardized for particular fiber types (cf. [5]).

Further, shape of optical signal can be distorted due to dispersion. Dispersion cause that optical impulse is widened in time. In extreme case, two consecutive impulses can overlap, leading to potential reception errors. Two types of dispersion ca be categorized: chromatic dispersion and polarization mode dispersion.

Chromatic dispersion is related to velocity difference between different wavelengths in particular medium. Due to linear character of chromatic dispersion, impairments introduced by this effect can be eliminated by using fiber spans with reverse dispersion characteristics. Compensating modules, called Dispersion Compensating Fiber (DCF), are typically attached after each fixed-length section of long-haul connection.

Polarization mode dispersion is caused by asymmetry of the fiber-optic strand. Polarization mode dispersion has thus completely random character. According to ITU-T recommendation G.652 [5], unit dispersion coefficient of optical fiber G.652B should not cross $17ps/(nm \times km)$ for chromatic dispersion and $0.20ps/(\sqrt{km})$ for polarization mode dispersion.

Another serious source of impairments is light scattering. Light scattering results from localized non-uniformity in the fiber medium. It can be seen as a deflection of a ray from a straight path. Deviations from the law of reflection due to irregularities on a surface of optical connectors are also usually considered to be a form of scattering. Among light scattering effects, several effects can imply serious impairments in photonic networks: stimulated Brillouin scattering and stimulated Raman scattering. Scattering effects are non-linear, and they tend to manifest themselves when optical signal power is high.

Other serious non-linear impairments in photonic networks are related to: Four-Wave Mixing, Self-Phase Modulation, and Cross-Phase Modulation. These effects result from transmitting multiple wavelengths over single fiber. Non-linear effects increase level of noise in the optical signal.

### E. Network performance

During transmission in medium, optical signals are distorted by certain optical effects. Resulting transmission errors affects OTN service quality. Some of effects, having linear characteristics, like fiber attenuation or chromatic dispersion, can be compensated by using dedicated modules. Other effects introduce distortions that cannot be recovered without translation to electric domain.

Major performance indicator of OTN service quality, called Bit Error Rate or simply BER, is thus related to the number of erroneously transmitted bits in digital client signal. In relation to noise introduced by non-linear optical effects and active optical elements, BER can be treated as a function of Optical Signal to Noise Ratio (OSNR), representing gap between power of optical signal and power of noise. In relation to effects not changing power level, but its distribution in time, BER can be treated as a function of dispersion coefficients.

OSNR related to sequence of active elements can be calculated according to formula (1).

$$1/OSNR = \textstyle\sum_i 1/OSNR_i, \qquad (1)$$

where $OSNR_i$ determines partial OSNR of $i$-th element in the sequence.

### F. Network cost model

Adjacent optical devices are connected by optical fibers, attached to their line interfaces. Multiple fibers are further combined into optical cables, connecting network sites and wells. Cables are attached to optical distribution frames, where incoming fibers are interconnected in order to establish end-to-end optical spans between devices. Signaling between WDM devices is realized via proprietary out-of-band channel, outside optical grid.

Optical device is a complex device, equipped with variety of specialized functions required for photonic signal processing. Typically, those specialized functions are realized by expansion cards, fitted into device slots. Cards can be further interconnected by back-plane device wiring or through external patch-cords, fitted manually into their front panel interfaces. In practice, to transport client signals, optical device must be equipped with a minimal set of supervisory cards, responsible for device control and management. Those cards implement proprietary vendor-specific algorithms to enable proper transmission and reception of photonic signals between adjacent devices. In the following, cost related to those cards is treated as a part of device fixed cost, being in turn a part of CAPital EXpenditure (CAPEX) related to network deployment.

In practice, optical devices are sold in the form of racks (for example, one rack per direction), equipped with certain number of slots for shelves. Each shelf in turn may be equipped with a number of slots to host expansion cards. Tributary and linear cards may be combined with embedded or external Network Interface Controllers (NICs). In the latter case, external NICs are fitted into appropriate slots on cards. ONE slots must thus be filled with a set of required expansion cards providing specialized functionality. Expansion cards can be classified into two categories:

- fixed cards: switching matrices, band multiplexers, supervision cards, fans, power suppliers, etc.,
- elastic cards: channel multiplexers, transponders, and muxponders.

Number and type of fixed cards is predefined by system vendor. Cost of fixed cards, together with cost of racks and shelves, is accounted into system installation cost. In other words, cost of the fixed cards is independent of traffic amount handled by ONE. Contrary to fixed cards, number and type of elastic cards can vary according to specific usage of the WDM system. In particular, network operator may choose between different types of transponders providing different modulation types, capacity, and forward error correction methods. In some ONE designs, channel multiplexers can be not expandable, and full range of multiplexers must be installed in the form of fixed cards. However, in the following, we consider elastic channel multiplexer cards as more general case. Band multiplexers are typically fixed cards. Still, due to particular composition of the switching matrices, some ONE designs may require installation of one rack per each direction (representing long-haul connection with neighboring ONE). In that case, predefined set of fixed cards, including band multiplexers, must be installed in each rack.

According to ITU-T recommendation G.709 [2], signals transmitted through OTN network compose the OTH hierarchy. OTH also defines structure and bandwidth of tributary signals crossing the UNI interface of OTN network. As bandwidth of single WDM channel may be greater than bandwidth of typically used Layer 2 signals, multiple client signals may be concatenated into larger signals, better fitted to bandwidth of WDM channels. Signals are typically concatenated in the time domain through dedicated concentrator cards.

Tributary signals need to be further framed according to ODU definition, coupled with error correction overhead, and transformed into colored optical signals to be transmitted through WDM network. Card responsible for these functions is called transponder. Optical signal transmitted by source transponder is again converted to the digital domain in the destination transponder. Between transponders, signal remains in the optical domain. Still, certain control information related to the transmitted signal is carried along signal path through dedicated administration channel.

## III. Flow design

In this section we consider flow design problem related to OTN/WDM photonic networks. Considered flow design problem is formulated in terms of mathematical programming. Having given basic WDM network topology and set of traffic demands to be realized, OTN/WDM flow design problem is aimed at identifying a flow distribution and composition of expandable WDM components (e.g., muxponders, transponders, multiplexers, etc.) leading to optimized value of certain objective function. In particular, feasible solution of the considered problem identifies design of the ONE nodes in terms of number, type, and configuration of expansion cards necessary to realize traffic demands, and associated cost.

Considered problem is described in the literature as Routing Wavelength Assignment (RWA) problem. It is commonly considered in combination with objective function maximizing the number of concurrent connections. Example integer linear programming formulation of this problem can be found in [6]. Independently in [7] and [8] it was proved that RWA problem is $\mathcal{NP}$-complete. Formulations proposed in the literature ([9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]) differ from formulation proposed in the following in terms of graph construction. Namely, in this paper it is assumed that each $\lambda$ channel constitutes separate edge in the network graph. Such assumption is not common in other works, but it allows to simplify formulation, and increase problem flexibility through graph construction. Moreover, classical RWA problem is concerned with routing and $\lambda$ selection only. Here, problem is extended with consideration of the access side. This extension is motivated by usage of objective function related to cost of elastic expansion cards.

### A. Assumptions

Traffic demands are assumed to be known in advance, e.g., they can be sourced from some external business forecast and measurement tool. As demand variation is out of the scope at considered network design problem, demand volumes may be additionally adjusted with some security margin. Each traffic demand is defined by triple: source, destination, and bandwidth volume. Demand source and destination are external clients connected to local ONE nodes through intra-office or short-haul black&white fibers.

Client devices and ONE devices are installed within Points of Presence (PoPs) of a network operator, and each client device is connected to uniquely defined ONE, usually in the same PoP. Even if in some PoP, ONE device is not installed, client localized in such PoP must be unambiguously assigned and connected to one ONE in one of the other PoPs.

After installing full suite of channel multiplexers, ONE device is capable to handle $N$ channels, equal to its maximum capacity. Each channel has precisely defined central frequency and width. Central frequencies of consecutive channels are supposed to be compatible with one of optical grids defined by ITU-T.

### B. Network graph

Network topology at the simplest level defines locations, configuration, and type of network elements, and arrangement of long-haul fibers connecting network elements. Depending on required level of granularity, network topology can be more or less detailed. At level of details required by flow optimization, this simple topology needs to be extended with deeper insight into composition of network elements. For this purpose, we define a directed graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ composed of set of nodes $\mathcal{V}$ and set of edges $\mathcal{E}$. Graph composition is used in the following as a basic modeling methodology. It allows to formulate considered flow design problem in terms of multi-commodity flow optimization.

*1) Graph nodes:* In general, node set $\mathcal{V}$ can refer to four types of physical elements (cards or whole devices):

- optical network element - device responsible for multiplexing, switching, and (optionally) converting colorful $\lambda$ signals ($\mathcal{O}$),
- transponder - expansion card responsible for adopting colorless tributary signals and modulating them as colorful $\lambda$ signals ($\mathcal{T}$),
- muxponders - expansion card responsible for concatenating multiple colorless signals into higher-order colorful signals ($\mathcal{M}$),
- client - non-WDM device, consuming OTN services ($\mathcal{C}$).

In order to model switching and converting colorful $\lambda$ signals, each ONE is represented in the network graph $\mathcal{G}$ by a set of graph nodes (referred to as *colorful nodes*), each associated with exactly one $\lambda$ and one direction towards adjacent ONE. Accordingly, number of colorful graph nodes associated with single ONE is equal to $N \times D$, where $D$ is the number of ONE neighbors. Similarly, basic graph of long-haul fiber connections is replicated, so there exists $N$ (equal to number of $\lambda$'s) parallel subgraphs, each topologically isomorphic with original network graph. If ONEs are capable to convert $\lambda$ frequencies, all colorful nodes associated with single ONE needs to be interconnected. For example, if in the feasible solution, such artificial link is crossed on path between colorful nodes associated with $\lambda_1$ and $\lambda_2$, it means that signal incoming to the related ONE at $\lambda_1$ is transmitted out through $\lambda_2$. Example subgraph of colorful nodes associated with two neighboring 3-direction ONEs in ROADM configuration is presented in Figure 10. T&ROADM counterpart extends the ROADM subgraph with full mesh connections between colorful nodes inside ONE, as presented in Figure 11.

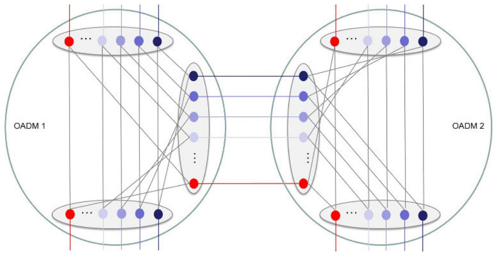


Fig. 10. ROADM subgraph

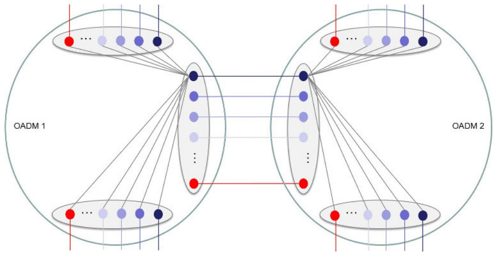


Fig. 11. T&ROADM subgraph

Transponder and concentrator are sometimes combined as one expansion card – muxponder. If not combined, clients can be connected to transponders two-fold: through direct connections or indirectly through hierarchy of compatible concentrators. In the network graph, stand-alone transponders are associated with a subset of graph nodes $\mathcal{T}$, where each graph node is associated with one transponder type and one ONE location. Associated subgraph is presented in Figure 12. In the figure, there are three transponder types (say 10Gbps, 40Gbps, and 100Gbps) and two clients. Transponder graph nodes representing each transponder type in one location are connected to all colorful nodes.

Muxponders form subset of graph nodes $\mathcal{M}$, where each muxponder type is replicated $N$ times, so each colorful node

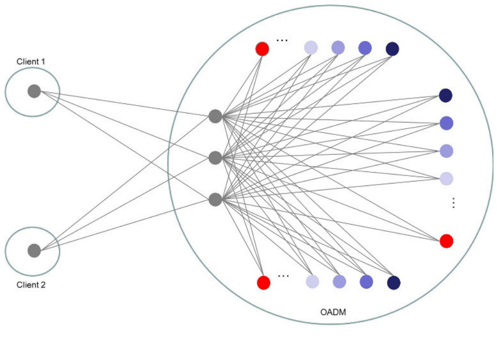


Fig. 12. Transponder subgraph

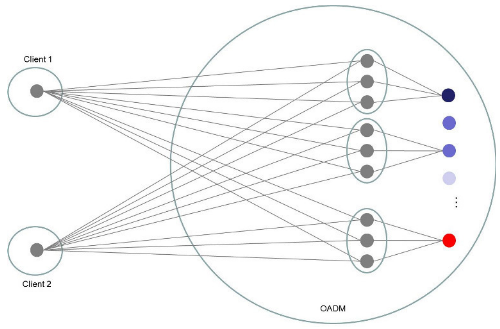


Fig. 13. Muxponder subgraph

can be connected to its own unique suite of muxponders. Despite graph contains all potential muxponder cards, only some subset of cards may be required in the optimal solution. All muxponder graph nodes are connected with graph nodes representing compatible interfaces in the client devices, as it is presented in Figure 13. Whole muxponder hierarchy is represented by single graph node, which means that multiple different connection types (different transport technology modules) are represented as parallel graph links, each associated with one transport technology type.

Client devices are not replicated in the network graph, and there exists exactly one graph node associated with each physical client device.

*2) Graph edges:* Nodes $\mathcal{V}$ are connected by set of edges $\mathcal{F}$, referring to physical connections:

- long-haul fibers connecting ONEs ($\mathcal{H}$),
- intra-office fibers connecting line ports in client devices and tributary ports in muxponders/transponders,
- patch-cords and back-plane wiring connecting line ports in transponders and tributary ports in ONE multiplexers.

All enumerated types of connections: fibers, patch-cords, and wiring are further described by common term link. Set $\mathcal{F}$ is assumed to be a superset of the link set defined by the basic network topology. In particular, it contains replicated edges between adjacent colorful nodes. Finally, not all edges contained in $\mathcal{F}$ will be deployed, because some graph nodes represent non-existent components and link deployment will depend on card installation. A subset of these potential links will be selected for deployment or activation.

## C. Mathematical formulation

Based on introduced network graph definition, the WDM flow design problem can be formulated as below mixed integer programme.

**object sets**

| | |
|---|---|
| $\mathcal{E}$ | (directed) demands (e.g., IP links) |
| $\mathcal{V}$ | nodes |
| $\mathcal{O} \subset \mathcal{V}$ | colorful ONE nodes |
| $\mathcal{T} \subset \mathcal{V}$ | transponders |
| $\mathcal{M} \subset \mathcal{V}$ | muxponders |
| $\mathcal{C} \subset \mathcal{V}$ | clients (e.g., IP routers) |
| $\mathcal{F} = \mathcal{H} \cup \mathcal{L}$ | (directed) edges (WDM links) |
| $\mathcal{G} \subset \mathcal{F}$ | edges associated with transponder links |
| $\mathcal{H} \subset \mathcal{F}$ | edges associated with long-haul links |
| $\mathcal{L} \subset \mathcal{F}$ | edges associated with intra-office links |
| $\mathcal{A}_v \subset \mathcal{F}$ | edges outgoing from node $v \in \mathcal{V}$ |
| $\mathcal{B}_v \subset \mathcal{F}$ | edges incoming to node $v \in \mathcal{V}$ |
| $\mathcal{P}_v \subset \mathcal{F}$ | edges associated with add-drop (tributary)<br>ports in colorful ONE nodes $v \in \mathcal{O}$ |
| $\mathcal{Q}$ | data transmission technologies |

**predefined objects**

| | |
|---|---|
| $a(e) \in \mathcal{C}$ | originating client node (source) of demand $e \in \mathcal{E}$ |
| $b(e) \in \mathcal{C}$ | terminating client node (sink) of demand $e \in \mathcal{E}$ |
| $a(f) \in \mathcal{V}$ | originating client node (source) of edge $f \in \mathcal{F}$ |
| $b(f) \in \mathcal{V}$ | terminating client node (sink) of edge $f \in \mathcal{F}$ |
| $\alpha(fe) \in \mathcal{L}$ | terminating line related to originating link $f \in \mathcal{L}$ with regard to demand $e \in \mathcal{E}$ |

**constants**

| | |
|---|---|
| $c_e$ | volume of demand $e \in \mathcal{E}$ |
| $l_f$ | capacity module of link $f \in \mathcal{F}$ |
| $t_v$ | equal to the maximum number of active tributary links of muxponder, if $v \in \mathcal{M}$;<br>equal to 1, if $v \in \mathcal{O}$ |
| $n_f$ | equal to $N$, if $f \in \mathcal{G}$;<br>equal to 1, if $f \in \mathcal{F} \setminus \mathcal{G}$ |

**variables**

| | |
|---|---|
| $s_{fe} \in \{0,1\}$ | variable equal to 1 if demand $e \in \mathcal{E}$ is realized on link $f \in \mathcal{F}$, and 0 otherwise |
| $z_f \in \mathbb{Z}$ | variable equal to the number of transport modules on link $f \in \mathcal{F}$ |

**constraints**

$$\sum_{f \in \mathcal{A}_v} l_f s_{fe} = c_e \qquad e \in \mathcal{E}, v = a(e) \in \mathcal{C} \tag{2a}$$
$$\sum_{f \in \mathcal{B}_v} l_f s_{fe} = c_e \qquad e \in \mathcal{E}, v = b(e) \in \mathcal{C} \tag{2b}$$
$$\sum_{f \in \mathcal{A}_v} s_{fe} = \sum_{f \in \mathcal{B}_v} s_{fe} \qquad e \in \mathcal{E}, v \in \mathcal{V} \setminus \{a(e), b(e)\} \tag{2c}$$
$$\sum_{e \in \mathcal{E}} s_{fe} \leq z_f \qquad f \in \mathcal{L} \tag{2d}$$
$$\sum_{e \in \mathcal{E}} s_{fe} \leq M z_f \qquad f \in \mathcal{H} \tag{2e}$$
$$\sum_{f \in \mathcal{A}_v} z_f = \sum_{f \in \mathcal{B}_v} z_f \qquad v \in \mathcal{O} \tag{2f}$$
$$\sum_{f \in \mathcal{B}_v} z_f \leq t_v \qquad v \in \mathcal{M} \cup \mathcal{O} \tag{2g}$$

$$\textstyle\sum_{f\in\mathcal{A}_v} z_f \le 1 \qquad v\in\mathcal{M}\cup\mathcal{O} \qquad (2h)$$
$$z_f \le n_f \qquad f\in\mathcal{F}. \qquad (2i)$$

Presented formulation is a modified form of classical formulation of multi-commodity flow optimization problem (see [21] and [22]). In this formulation, flow distribution is described by values of binary variables $\boldsymbol{s}$ representing flows on particular network links. Having given feasible values of $\boldsymbol{s}$ one can easily reconstruct particular paths selected to carry traffic.

Integer variables $\boldsymbol{z}$ determine in general number of transmission modules on particular links. However, in case of links associated with nodes $v\in\mathcal{M}\cup\mathcal{O}$ this number is strictly binary (due to constraints (2h) and (2i)). For the rest, variable $\boldsymbol{z}$ is integer (due to constraints (2g) and (2i)).

Due to classical flow conservation constraints (see [21]), in relation to specific demand, in all nodes, except end nodes of this demand, the total volume of incoming flows must be balanced by total volume of outgoing flows. Formulation (2) involves two groups of flow conservation constrains: constraints (2a)-(2c) related to variables $\boldsymbol{s}$ and constraints (2f) related to variables $\boldsymbol{z}$.

Usage of particular network links, including all types of inter-card patch-cords and back-plane wiring, by flows determines consumption of transport modules (their number is expressed by variables $\boldsymbol{z}$), according to constraints (2d) and (2e).

Constraints (2g) assure that only one transponder or muxponder can be coupled with each channel tributary port in ONE multiplexer. Similarly, number of active tributary and line links connected to muxponder ports are limited by constraints (2g) and (2h), respectively.

Formulation (2) gathers constraints related to using WDM transport to carry client traffic. Based on this formulation, in the following we consider a number of its extensions and composition of objective function related to the overall cost associated with WDM transport.

### D. L2 technology

To express that each demand can be realized using homogenous L2 technology, like GigabitEthernet, FC800, STM64, binary variable vector $\boldsymbol{k}$ was introduced. Non-zero value of variable $k_{ge}$ enforces through constraints (3a) that demand $e\in\mathcal{E}$ can be realized using only links compliant with technology $g\in\mathcal{Q}$. If one technology (say $g\in\mathcal{Q}$) is selected (value $k_{ge}$ is 1), links associated with other technologies cannot be used, what is assured by constrains (3b).

$$\textstyle\sum_{f\in\mathcal{R}_g} s_{fe} \le |\mathcal{R}_g| k_{ge} \qquad g\in\mathcal{Q}, e\in\mathcal{E} \qquad (3a)$$
$$\textstyle\sum_{g\in\mathcal{Q}} k_{ge} \le 1 \qquad e\in\mathcal{E}. \qquad (3b)$$

Above, $\mathcal{R}_g \subset \mathcal{F}$ denotes set of edges associated with technology $g\in\mathcal{Q}$.

### E. Redundancy

To provide uninterrupted services, able to survive failures of optical network elements and fiber connections, client devices need additional bandwidth, allocated along paths not affected by considered failures. Additional bandwidth, required by protection, is associated with certain level of resource redundancy. Redundant resources are either not used in the nominal network state or can be used for transmitting low priority traffic, preempted in case of failure occurrence.

In case of the WDM networks, redundant resources can be provided either at client digital signal level (called client protection) or photonic signal level (called photonic protection). In the former case, client device is responsible for activating redundant resources. Redundant resources cover optical channels allocated along protection path, and transponder/muxponder cards and ports. In the latter case, specialized protection cards are required. Such protection cards split power of the protected optical signal between multiple (usually two) ports connected to different add-drop ports within multiplexer subsystem. In both cases, the nominal and protection paths should be topologically disjoint with regard to failure occurrence, so under any failure at least one of the paths survives.

Let set $\mathcal{F}_i, i\in\mathcal{I}$ represents an arbitrary set of links that share risk of failure. Such group is described in the literature as Shared Risk Link Group (SRLG) or in general Shared Risk Resource Group (SRRG) [23]. Each SRLG associated with single link $f\in\mathcal{F}$ failure contains exactly one element. Each SRLG associated with node $v\in\mathcal{O}$ failure contains all adjacent links, i.e., $\mathcal{A}_v\cup\mathcal{B}_v$. SRLG should be constructed case by case in relation to specific needs and requirements of a network operator. Specific composition of SRLG thus remains out of scope of this paper.

In order to determine capacity $c_{ei}$ allocated to demand $e$ and available during failure state $i$, constraints (4a)–(4b) should be added to the problem formulation (2):

$$0 \le c_e - c_{ei} \le M r_{ei} \qquad e\in\mathcal{E}, i\in\mathcal{I} \qquad (4a)$$
$$0 \le c_{ei} \le M(1-r_{ei}) \qquad e\in\mathcal{E}, i\in\mathcal{I} \qquad (4b)$$
$$0 \le r_{ei} \le \textstyle\sum_{f\in\mathcal{F}_i} s_{fe} \le M r_{ei} \qquad e\in\mathcal{E}, i\in\mathcal{I}. \qquad (4c)$$

Above, each variable $c_{ei}$ expresses volume of link flow associated with demand $e$ in failure state $i$. Value of variable $r_{ei}$ indicates if link $e$ is available throughout failure state $i$. Consequently, if for some pair $(e,i)$ $r_{ei}=1$ then associated $c_{ei}$ is equal to $c_e$, and $c_{ei}$ is zero otherwise. In according to constraints (4c), value of $r_{ei}$ is positive if and only if at least one link $f$ realizing demand $e$ is affected by failure $i$, i.e., when $\sum_{f\in\mathcal{F}_i} s_{fe} \ge 0$.

Redundancy required by protection mechanisms can be also modeled through multiplication of demand volume to be realized by the transport WDM/OTN network, and additional constraints assuring that only fraction of demand volume is transmitted through specific resources (network element or link). Protection method associated with described resource redundancy requirement is commonly described in the literature as path diversity [24]. To assure introduced requirement constraints (2a)–(2b) must be rewritten as:

$$\textstyle\sum_{f\in\mathcal{A}_v} l_f s_{fe} = 2c_e \qquad e\in\mathcal{E}, v=a(e) \quad (5a)$$
$$\textstyle\sum_{f\in\mathcal{B}_v} l_f s_{fe} = 2c_e \qquad e\in\mathcal{E}, v=b(e) \quad (5b)$$
$$\textstyle\sum_{f\in\mathcal{F}_i} l_f s_{fe} \le c_e \qquad e\in\mathcal{E}, i\in\mathcal{I}. \quad (5c)$$

Constraints (5c) assure that any link flow do not exceed demand volume. In result, each demand must be realized on at least two disjoint paths.

### F. Optical impairments

Optical impairments (described in Section III-F) affecting optical signals, transmitted through photonic network, can be modeled in the form of three limitations:

- fiber length limit,
- hop-count limit,
- noise accumulation limit.

Chromatic dispersion is responsible for spreading duration of optical signal peaks. As a linear effect, proportional to the total length of fiber, chromatic dispersion can be eliminated by using DCM modules. DCM modules are supposed to introduced chromatic dispersion in reverse direction than dispersion introduced by regular fiber. DCM modules compensate thus chromatic dispersion related to central frequency of the optical signal. Consequently, dispersion affecting frequencies far from the central frequency are not compensated completely. Amount of uncompensated chromatic dispersion is called residual dispersion. Residual dispersion is an important factor that limits the total length of fiber traversed by optical signal. Maximum admissible fiber length depends on transponder type and characteristics.

Optical signal propagating through fiber medium is attenuated. Fiber attenuation is proportional to the total length of fiber spans crossed by the signal. To restore signal power to the level required by photo-detector, optical signal is amplified by LOFA cards localized in selected points along the path. However, LOFA cards, beside signal amplification, introduce some portion of noise. To keep signal quality at high level, network operator should control the total amount of introduced noise. On one hand, low noise level can be assured by hop-count constraints. On the other hand, noise characteristic of active optical elements, as mentioned LOFA cards, can be expressed in terms of ONSR value. As described in Section II-E, total value of OSNR is proportional to partial OSNR of particular elements in the path.

All introduced in this section limitations can be associated with additive metrics: length, hop-count, and inverse OSNR. Accordingly, all can be modeled similarly by set of so-called shortest path constraints. Formulation of shortest path constraints is based on path length variables $\boldsymbol{p} = (p_v : v\in\mathcal{O})$. Each $p_v$ represents the length of the shortest path from $v$ with respect to weight system $\boldsymbol{q}$. Then, for each link $f$ outgoing from node $v$ ($a(f)=v$) contained in the shortest path crossed by edge the following shortest path condition must hold.

$$p_{a(f)} + q_f = p_{b(f)}. \qquad (6)$$

Condition (6) is commonly used by the shortest path algorithms to validate if the path traversing edge $f$ is shorter than the shortest path found so far. For our purposes we adopt condition (6) to formulate shortest path constraints:

$$p_{a(f)} + q_f - p_{b(f)} = 0 \text{ if value of } z_f \text{ is } 1 \quad f\in\mathcal{E} \quad (7a)$$
$$p_{a(f)} + q_f - p_{b(f)} \ge m \text{ if value of } z_f \text{ is } 0 \quad f\in\mathcal{E}. \quad (7b)$$

Conditions (7a)-(7b) state that if and only if edge $f$ is contained in the shortest path to $b(f)$, length of this path must be equal to sum of the length of a shortest path to $a(f)$ and weight $q_f$. Otherwise; the value of the expression $p_{a(f)} + q_f - p_{b(f)}$ must be greater or equal to $m$, which is the smallest difference between lengths of two paths. Accordingly, length limitation constraints can be formulated as follows:

$$m(1-z_f) \le p_{a(f)} + q_f - p_{b(f)} \le Mz_f \qquad f\in\mathcal{E} \quad (8a)$$
$$p_v \le p^* \qquad v\in\mathcal{V}. \quad (8b)$$

Considered limitations require additional constraints (8b) to enforce that values of required parameters remain under maximum admissible level $p^*$. Weight system $\boldsymbol{q}$ is constant, and is supposed to express value of required parameter:

- link length,
- number of active elements associated with link (usually one),
- inverse OSNR associated with link.

### G. Network cost

Cost of WDM transport is mostly related to the number and type of used elastic expansion cards: channel multiplexers, transponders, and muxponders. Cost related to installation of transponder and muxponder cards can be expressed as follows:

$$\textstyle\sum_{v\in\mathcal{O}}\sum_{f\in\mathcal{P}_v} \frac{1}{2} g_v z_f. \qquad (9)$$

Above, unitary cost related to card associated with node $v\in\mathcal{M}\cup\mathcal{T}$ is given by constant $g_v$.

To calculate cost related to installation of multi-stage multiplexer expansion cards we need to introduce additional variables and constrains. Binary variable $m_j$ associated with multiplexer $j\in\mathcal{J}$ states if card is installed or not. Variable is positive if at least one channel associated with this particular multiplexer is used. This relation is expressed by constrains (10). Set of colorful channel links and cost associated with multiplexer $j\in\mathcal{J}$ are given by $\mathcal{S}_j$ and $h_j$, respectively.

$$\textstyle\sum_{f\in\mathcal{S}_j} z_f \le |\mathcal{S}_j| m_j \qquad j\in\mathcal{J}. \quad (10)$$

Finally, with respect to (10) and other constraints defined above, the objective function can be formulated as:

$$\boldsymbol{min}\ F(\boldsymbol{z},\boldsymbol{m}) = \textstyle\sum_{v\in\mathcal{O}}\sum_{f\in\mathcal{P}_v} \frac{1}{2} g_v z_f + \sum_{j\in\mathcal{J}} h_j m_j. \qquad (11)$$

Objective function (11) is related to minimization of number of expansion cards.

TABLE I
CHARACTERISTICS OF THE SELECTED NETWORK INSTANCES

| Network instance | Nodes | Links | Demands |
|---|---|---|---|
| abilene | 12 | 15 | 132 |
| atlanta | 15 | 22 | 210 |
| brain | 161 | 332 | 14311 |
| cost266 | 37 | 57 | 1332 |
| geant | 22 | 36 | 462 |
| germany50 | 50 | 88 | 662 |
| giul39 | 39 | 172 | 1471 |
| france | 25 | 45 | 300 |
| janos-us | 26 | 84 | 650 |
| janos-us-ca | 39 | 122 | 1482 |

TABLE II
CHARACTERISTICS OF THE NETWORK GRAPHS

| Network instance | $\lvert\mathcal{O}\rvert$ | $\lvert\mathcal{C}\rvert$ | $\lvert\mathcal{L}\rvert$ | $\lvert\mathcal{H}\rvert$ | $\lvert\mathcal{F}\rvert$ |
|---|---|---|---|---|---|
| abilene | 2400 | 12 | 1200 | 36 | 1236 |
| atlanta | 3520 | 15 | 1760 | 45 | 1805 |
| brain | 53120 | 161 | 26560 | 483 | 27043 |
| cost266 | 9120 | 37 | 4560 | 111 | 4671 |
| geant | 5760 | 22 | 2880 | 66 | 2946 |
| germany50 | 14080 | 50 | 7040 | 150 | 7190 |
| giul39 | 27520 | 39 | 13760 | 117 | 13877 |
| france | 7200 | 25 | 3600 | 75 | 3675 |
| janos-us | 13440 | 26 | 6720 | 78 | 6798 |
| janos-us-ca | 19520 | 39 | 9760 | 117 | 9877 |

## IV. Complexity

Throughout this section we estimate complexity of the considered formulation of the WDM flow design problem. Complexity estimation is based on calculation of the numbers of variables and constrains necessary to formulate the considered WDM flow design problem in relation to the selected instances of network instances defined in the SNDLib library [25]. Referenced network instances are characterized in Table I, where particular columns contain the numbers of network nodes, network links, and traffic demands, respectively.

Further, assuming 80-channel WDM technology and three types of transponders (10Gbps, 40Gbps, 100Gbps) we calculate the numbers of particular types of graph elements related to the considered network instances. Calculation results are presented in Table II.

Finally, Table III contains the numbers of constrains and variables necessary to formulate the considered WDM flow design problem in relation to the selected SNDLib network instances. Number of constraints is contained in a range from 19 thousands to 26 millions. Number of variables is even larger and is contained in a range from 164 thousands to 387 millions. Such enormous numbers of constraints and variables make in practice the considered formulations numerically intractable for resolving with exact optimization methods.

TABLE III
CHARACTERISTICS OF THE FORMULATIONS

| Network instance | Constraints | Variables |
|---|---|---|
| abilene | 19704 | 164388 |
| atlanta | 38335 | 380855 |
| brain | 26151647 | 387039416 |
| cost266 | 484005 | 6226443 |
| geant | 114852 | 1363998 |
| germany50 | 372608 | 4766970 |
| giul39 | 1263603 | 20426944 |
| france | 96825 | 1106175 |
| janos-us | 307484 | 4425498 |
| janos-us-ca | 956135 | 14647591 |

## V. Conclusion

Paper investigates mathematical modeling of photonic networks applying wavelength division multiplexing. Based on standardization efforts, research work, and commercial offerings, a functional model of WDM network is proposed. Proposed network model involves a series of specific aspects of photonic transmission, like fiber attenuation, dispersion, and noise accumulation. Developed functional model is a basis for development of mathematical models of multicommodity flows in WDM network.

Number of integer variables used in a mathematical model in high degree determines computational complexity of optimization formulations based on this model. In case of the proposed model, this number is proportional to the squared number of WDM devices and number of WDM channels. For even small network instances (composed of several devices), this number can be at level of thousands. Thus, in the future work, in order to reduce complexity of the proposed model to numerically tractable level, authors will try to decompose it. In particular, future work will focus on adopting general decompositions methods proposed in context of large-scale linear programming, like Dantzig-Wolfe decomposition, Lagrangean relaxation, and Benders decomposition, for the case of proposed model.

## References

[1] *Spectral grids for WDM applications: DWDM frequency grid.* Recommendation ITU-T G.694.

[2] *Interfaces for the optical transport network.* Recommendation ITU-T G.709.

[3] *Architecture of optical transport networks.* Recommendation ITU-T G.872.

[4] *Network node interface for the synchronous digital hierarchy (SDH).* Recommendation ITU-T G.707.

[5] *Characteristics of a single-mode optical fibre and cable.* Recommendation ITU-T G.652.

[6] H. Zang, J. Jue, and B. Mukherjee, "A review of routing and wavelength assignment approaches for wavelength routed optical WDM networks," *Optical Networks Magazine*, January 2000.

[7] I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: an approach to high bandwidth optical WAN's," *IEEE Transactions on Communications*, vol. 40, no. 7, pp. 1171–1182, July 1992. doi: 10.1109/26.153361. [Online]. Available: http://dx.doi.org/10.1109/26.153361

[8] S. Evan, A. Itai, and A. Shamir, “On the complexity of timetable and multicommodity flow problems,” *SIAM Journal of Computing*, vol. 5, pp. 691–703, 1976. doi: 10.1137/0205048. [Online]. Available: http://dx.doi.org/10.1137/0205048

[9] E. Varvarigos, K. Manousakis, and K. Christodoulopoulos, “Cross layer optimization of static lightpath demands in transparent WDM optical networks,” in *IEEE Information Theory Workshop on Networking and Information Theory*, 2009. doi: 10.1109/ITWNIT.2009.5158553. [Online]. Available: http://dx.doi.org/10.1109/ITWNIT.2009.5158553

[10] W. Zhang, J. Tang, K. Nygard, and C. Wang, “REPARE: Regenerator placement and routing establishment in translucent networks,” in *IEEE Global Telecommunications Conference GLOBECOM*, 2009. doi: 10.1109/GLOCOM.2009.5425649. [Online]. Available: http://dx.doi.org/10.1109/GLOCOM.2009.5425649

[11] K. Christodoulopoulos, K. Manousakis, and E. Varvarigos, “Considering physical layer impairments in offline RWA,” *IEEE Network*, vol. 23, June 2009. doi: 10.1109/MNET.2009.4939260. [Online]. Available: http://dx.doi.org/10.1109/MNET.2009.4939260

[12] K. Manousakis, K. Christodoulopoulos, E. Kamitsas, I. Tomkos, and E. Varvarigos, “Offline impairment-aware routing and wavelength assignment algorithms in translucent WDM optical networks,” *Journal of Lightwave Technology*, vol. 27,12, 2009. doi: 10.1109/JLT.2009.2021534. [Online]. Available: http://dx.doi.org/10.1109/JLT.2009.2021534

[13] P. Pavon-Marino, S. Azodolmolky, R. Aparicio-Pardo, B. Garcia-Manrubia, Y. Pointurier, M. Angelou, J. Sole-Pareta, J. Garcia-Haro, and I. Tomkos, “Offline impairment aware RWA algorithms for cross-layer planning of optical networks,” *Journal of Lightwave Technology*, vol. 27,12, 2009. doi: 10.1109/JLT.2009.2018291. [Online]. Available: http://dx.doi.org/10.1109/JLT.2009.2018291

[14] C. Saradhi and S. Subramaniam, “Physical layer impairment aware routing (PLIAR) in WDM optical networks: issues and challenges,” *Communications Surveys & Tutorials, IEEE*, vol. 11, 4, 2009. doi: 10.1109/SURV.2009.090407. [Online]. Available: http://dx.doi.org/10.1109/SURV.2009.090407

[15] N. Sengezer and E. Karasan, “Static lightpath establishment in multilayer traffic engineering under physical layer impairments,” *IEEE/OSA Journal of Optical Communications and Networking*, vol. 2, 9, 2010. doi: 10.1364/JOCN.2.000662. [Online]. Available: http://dx.doi.org/10.1364/JOCN.2.000662

[16] E. Varvarigos, K. Manousakis, and K. Christodoulopoulos, “Offline routing and wavelength assignment in transparent WDM networks,” *IEEE/ACM Trans. on Networks*, vol. 18,5, 2010. doi: 10.1109/TNET.2010.2044585. [Online]. Available: http://dx.doi.org/10.1109/TNET.2010.2044585

[17] Y. Zhai, A. Askarian, S. Subramaniam, Y. Pointurier, and M. Brandt-pearce, “Cross-layer approach to survivable DWDM network design,” *IEEE/OSA Journal of Optical Communications and Networking*, vol. 2,6, 2010. doi: 10.1364/JOCN.2.000319. [Online]. Available: http://dx.doi.org/10.1364/JOCN.2.000319

[18] R. Aparicio-Pardo, M. Klinkowski, B. Garcia-Manrubia, P. Pavon-Marino, and D. Careglio, “Offline impairment-aware rwa and regenerator placement in translucent optical networks,” *Journal of Lightwave Technology*, vol. 29,3, 2011. doi: 10.1109/JLT.2010.2098393. [Online]. Available: http://dx.doi.org/10.1109/JLT.2010.2098393

[19] N. Sengezer and E. Karasan, “Multi-layer virtual topology design in optical networks under physical layer impairments and multi-hour traffic demand,” *EEE/OSA Journal Journal of Optical Communications and Networking*, vol. 4,2, 2012. doi: 10.1364/JOCN.4.000078. [Online]. Available: http://dx.doi.org/10.1364/JOCN.4.000078

[20] J. Sole, S. Subramaniam, D. Careglio, and S. Spadaro, “Cross-layer approaches for planning and operating impairment-aware optical networks,” in *Proc. the IEEE 100*, 2012. doi: 10.1109/JPROC.2012.2185669. [Online]. Available: http://dx.doi.org/10.1109/JPROC.2012.2185669

[21] M. Minoux, *Mathematical Programming: Theory and Algorithms*. John Wiley & Sons, 1986.

[22] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[23] J. Strand, A. Chiu, and R. Tkach, “Issues for routing in the optical layer,” *IEEE Communications Magazine*, 2001. doi: 10.1109/35.900635. [Online]. Available: http://dx.doi.org/10.1109/35.900635

[24] M. Dzida, n. T. Sliwi M. Zagozdzon, W. Ogryczak, and M. Pioro, “Path generation for a class survivable network design problems,” in *NGI 2008 Conference on Next Generation Internet Networks, Cracow, Poland*, 2008. doi: 10.1109/NGI.2008.11. [Online]. Available: http://dx.doi.org/10.1109/NGI.2008.11

[25] “SNDlib 1.0 – Survivable network design data library,” 2005. [Online]. Available: http://sndlib.zib.de

DOI: 10.15439/2015F340
ACSIS, Vol. 5

# On the generalized Wiener polarity index for some classes of graphs

Halina Bielak
Maria Curie – Skłodowska University
Pl. Marii Curie – Skłodowskiej 5
20-031 Lublin, Poland
Email: hbiel@hektor.umcs.lublin.pl

Kinga Dąbrowska, Katarzyna Wolska
Maria Curie – Skłodowska University
Pl. Marii Curie – Skłodowskiej 5
20-031 Lublin, Poland
Email: kinga.wiktoria.dabrowska@gmail.com,
katarzyna.anna.wolska@gmail.com

***Abstract*—The generalized Wiener polarity index $W_k(G)$ of a graph $G = (V, E)$ is defined as a number of unordered pairs $\{u, v\}$ of $G$ such that the shortest distance between $u$ and $v$ is equal to $k$:**

$$W_k(G) = |\{\{u, v\}, d(u, v) = k, u, v \in V(G)\}|$$

**In this paper we give some results for 2-trees in case of mentioned index. We present an infinite family of 2-trees with maximum value of generalized Wiener polarity index.**

## I. Introduction

Let $G = (V(G), E(G))$ be a connected, simple graph with $V(G)$ the vertex set and $E(G)$ the edge set. Let $n$ be the number of vertices and $m$ the number of edges. By $d(u, v)$ we denote the distance between two vertices $u$ and $v$ in the graph $G$. What we call a diameter $diam(G)$ is the longest distance between two vertices of $G$. The degree of the vertex $u$ in the graph $G$ is denoted by $deg(u)$. Other definitions, not mentioned here can be found in [1].

The Wiener polarity index of a graph $G = (V(G), E(G))$ is defined as

$$WP(G) = |\{\{u, v\} : d(u, v) = 3; u, v \in V(G)\}|$$

which is a number of unordered pairs of vertices $\{u, v\}$ of $G$ such that $d(u, v) = 3$. Authors of [4, 5, 7, 13] studied this index for trees with different parameters such that number of pendant vertices, diameter or maximum degree. Additionally, in [12] there are described algorithms for counting $W_k(T)$ for trees.

The generalized Wiener polarity index of a graph $G = (V(G), E(G))$ is defined as

$$W_k(G) = |\{\{u, v\}, d(u, v) = k, u, v \in V(G)\}|$$

which is a number of unordered pairs of vertices $\{u, v\}$ of $G$ such that the distance between $u$ and $v$ is equal to $k$.

Let us now remind the definition of the Wiener index $W(G)$

$$W(G) = \sum_{\{u,v\} \subseteq V(G)} d(u, v) = \frac{1}{2} \sum_{v \in V(G)} D(v),$$

where $D(v) = \sum_{u \in V(G)} d(u, v)$ is the sum of all distances from the vertex $v$. As we can see $W(G)$ is defined as the sum of the distances between all pairs of vertices in the graph $G$. Note that: $W(G) = \sum_{k=1}^{diam(G)} kW_k(G)$. The Hosoya polynomial (Wiener polynomial) of $G$ in $x$ is defined as follows

$$W(G, x) = \sum_{u,v \in V(G)} x^{d(u,v)} = \sum_{k=1}^{diam(G)} W_k(G) \cdot x^k$$

More information about Hosoya polynomial the reader can find in [9].

The applications of mentioned indices are described in the papers [2, 3] and also in [9, 10]. Probably the best known topological index is the Wiener index and this is the one described by many authors, for example [2, 8].

## II. Generalized Wiener polarity index

In case of generalized Wiener polarity index for trees there are some known results presented in [12]. Let $T$ be a tree. If $k = 1$ then $W_1(T) = m$, where $m$ is the number of edges. If $k = 2$ then

$$\begin{aligned} W_2(T) &= \sum_{v \in V(T)} \binom{deg(v)}{2} = \frac{\sum_{v \in V(T)} deg^2(v)}{2} - m \\ &= \frac{M_1(G)}{2} - m \end{aligned}$$

where $M_1(G)$ is the first Zagreb index of a graph. For detailed information on Zagreb indices the reader is referred to [11].

If $k = 3$ we have

$$\begin{aligned} W_3(T) &= \sum_{uv \in E(T)} (deg(v) - 1)(deg(u) - 1) \\ &= \sum_{uv \in E(T)} deg(u)deg(v) - \sum_{v \in V(T)} deg^2(v) + m \\ &= M_2(T) - M_1(T) + m \end{aligned}$$

where $M_2(T)$ is the second Zagreb index of a graph.

Let us now assume that $k \geq 3$. In a situation when diameter of $T$ is less than $k$ we have $W_k(T) = 0$ and that is why the minimum value of $W_k(T)$ is equal to zero. This is

978-83-60810-66-8/$25.00©2015, IEEE

achieved for all trees for which $diam(T) < k$. Actually, this is simple fact for each graph.

Now we will study the generalized Wiener polarity index for 2-trees. Let us define a 2-tree first. The smallest 2-tree is a complete graph $K_3$ of order $n = 3$. A 2-tree of order $n$ is obtained from a 2-tree $G$ of order $n-1$ by attaching a new vertex $v$ and two edges $\{vx, vy\}$ such that $\{x, y\} \in E(G)$. Concerning 2-trees with $diam(G) \geq k$ is more difficult than for trees.

Let $G$ be a 2-tree of order $n$ and size $m$. A pendant vertex in a 2-tree is a vertex with degree equal to 2. Now, for $k = 1$ the value of $W_1(G)$ stays the same as for trees. For $k = 2$ we have

$$W_2(G) = \sum_{v \in V(G)} \left( \binom{deg(v)}{2} - m \right)$$

But let us move on to what will be considered now and this are the maximum values of $W_k(G)$ where $G$ is a 2-tree.

What we are going to do is to decompose all vertices $v$ in $G$ with $deg(v) = 2$ into some number of groups. Each group has the following property

$$A_i = \{v \in V(G) : deg(v) = 2 \wedge \exists_{e_i = \{u_i, w_i\}}; vu_i, vw_i \in E(G)\}$$

for $i = 1, 2, ...$

We have at least two such groups. Let us say that the distance between two arbitrary pendant vertices from different groups is not equal to $k$. Distances between vertices in each group are equal to 2.

Let $p_1$ and $p_2$ be the numbers of vertices on distance $k$ from an arbitrary pendant vertex from $A_1$ and $A_2$, respectively. We can ssume that $p_1 \geq p_2$ with no loss of generality. After removal of all pendant vertices from $A_2$ and addition to the group $A_1$ we get the transformed 2-tree $G'$

$$\begin{aligned} & W_k(G') - W_k(G) \geq \\ & = (|A_1|p_1 + |A_2|p_1) - (|A_1|p_1 + |A_2|p_2) = \\ & = |A_2|(p_1 - p_2) \geq 0 \end{aligned} \tag{1}$$

Note this is true for two groups. If there are more of them inequality in (1) may not hold.

By repetition of this transformation we will get a new 2-tree with possibly greater generalized Wiener polarity index. The diameter of $G'$ after each transformation is less or stays the same as the one for $G$. Each transformation gives us also one new pendant vertex. If we will choose the most distant groups of pendant vertices we will get a 2-tree with diameter equal to $k$. After that we can apply the transformation finitely many times until all pendant vertices are on distance $k$ and no other vertex of the final 2-tree has eccentricity equal to $k$. During this process the $W_k(G)$ may be changing by decreasing or increasing. Some example is presented in $Fig.1$.

Let us assume we have $p$ groups of pendant vertices with sizes: $a_1, a_2, ..., a_p$ and $a_1 + a_2 + ... + a_p = q$. We consider a 2-tree with $diam(G) = k$. We have then $n - 2(k-1) \geq q \geq 2$. Assume that the distance between any two pendant vertices not from the same group is equal to $k$ and that is why

$$W_k(G) = \frac{1}{2} \sum_{i=1}^{p} a_i(q - a_i) = \frac{1}{2}\left( q^2 - \sum_{i=1}^{p} a_i^2 \right) \tag{2}$$

In the case when the distance between the group $A_i$ and $A_j$ for $i \neq j$ is less than $k$ the generalized Wiener polarity index is less than the one presented above. If $p = 2$ we have $W_k(G) = a_1 a_2$. This value is maximum for $a_1 + a_2 = n - 2(k-1)$, $a_1 = \left\lfloor \frac{n-2(k-1)}{2} \right\rfloor$ and $a_2 = \left\lceil \frac{n-2(k-1)}{2} \right\rceil$.

$$\begin{aligned} W_k(G) &= \left\lfloor \frac{n-2(k-1)}{2} \right\rfloor \left\lceil \frac{n-2(k-1)}{2} \right\rceil = \\ &= \left( \left\lfloor \frac{n}{2} \right\rfloor - (k-1) \right) \left( \left\lceil \frac{n}{2} \right\rceil - (k-1) \right) = \\ &= \left\lfloor \frac{n}{2} \right\rfloor \left\lceil \frac{n}{2} \right\rceil - (k-1) \left( \left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{n}{2} \right\rceil \right) + (k-1)^2 \end{aligned}$$

so

$$W_k(G) = \left\lfloor \frac{n}{2} \right\rfloor \left\lceil \frac{n}{2} \right\rceil - (k-1)(n-(k-1)) \tag{3}$$

Let $p > 2$ and $k > 2$. First we consider the even $k$. By $n \geq 2 + p(k-2) + q$ and $2 < p \leq q$ we have $p \leq \frac{n-2-q}{k-2}$, so

$$p < \frac{n-2}{k-2}. \tag{4}$$

We have the following

$$q = \sum_{i=1}^{p} a_i,$$

$$n = 2 + 2p\left( \frac{k}{2} - 1 \right) + \sum_{i=1}^{p} a_i = 2 + p(k-2) + \sum_{i=1}^{p} a_i.$$

Hence

$$n \geq 2 + p(k-2) + q. \tag{5}$$

We apply Cauchy - Schwarz ineqality to the formula (2) with $q \leq n - 2 - p(k-2)$

$$W_k(G) = \frac{1}{2}\left( q^2 - \sum_{i=1}^{p} a_i^2 \right) \leq \frac{1}{2}\left( q^2 - \frac{q^2}{p} \right) \leq \frac{1}{2} f(p) \tag{6}$$

where

$$f(p) = (n - 2 - p(k-2))^2 \left( 1 - \frac{1}{p} \right).$$

The extremal generalized Wiener polarity index $W_k(G)$ is obtained for the case when we have equality in (6). We are going to study this case. We will give some examples of extremal 2-trees and then we will state the final result in Theorem 1.

So for real variable $p$ we study

$$\begin{aligned} f(p) = & \left( (n-2)^2 + p^2(k-2)^2 \right) \left( 1 - \frac{1}{p} \right) \\ & - 2(k-2)(n-2)(p-1). \end{aligned} \tag{7}$$

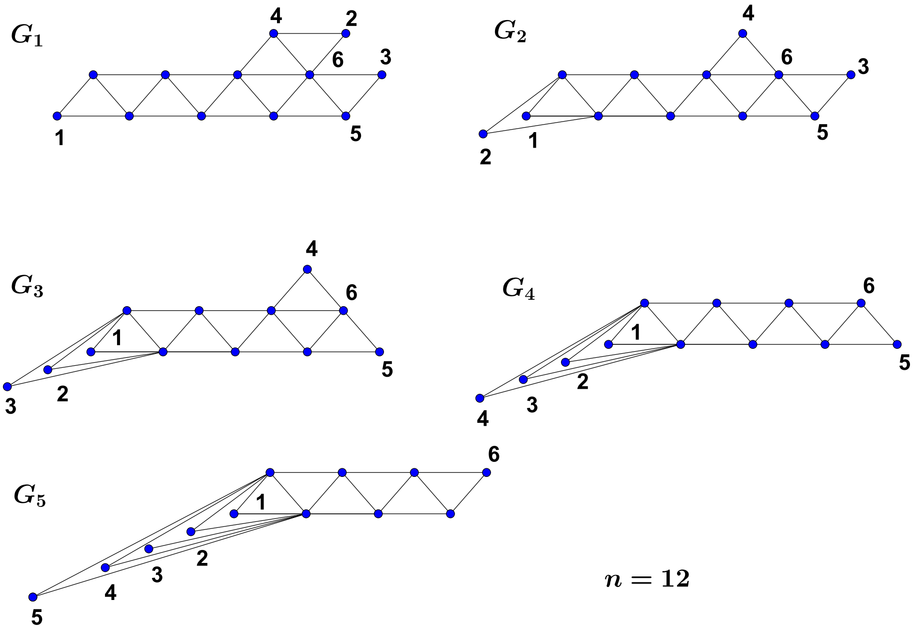


Fig. 1. A process of moving pendant vertices $G_1 \to G_2 \to G_3 \to G_4 \to G_5$. $W_4(G_1) = 8$, $W_4(G_2) = 9$, $W_4(G_3) = 10$, $W_4(G_4) = 9$, $W_4(G_5) = 5$.

Let $h(p) = 2(2-k)(1-\frac{1}{p}) + \frac{n-2-p(k-2)}{p^2}$. Then the first derivative equals

$$f'(p) = (n-2-p(k-2))h(p)$$

By (4) we have $f'(p) = 0$ if and only if $p = \hat{p}$, where

$$\hat{p} = \frac{1}{4} + \frac{1}{4}\sqrt{\frac{8n+k-18}{k-2}} \tag{8}$$

Similarly $f'(p) > 0$ if and only if $h(p) > 0$. This is equivalent to the inequality

$$g(p) = 2p^2 - p - \frac{n-2}{k-2} < 0.$$

So $g(p) < 0$ if and only if $p < \hat{p}$ .
Let

$$s = 4\hat{p} = 1 + \sqrt{\frac{8n+k-18}{k-2}}.$$

Then

$$\frac{1}{\hat{p}} = \frac{\sqrt{(k-2)(8n+k-18)} - (k-2)}{2n-4} = \frac{(k-2)(s-2)}{2n-4}.$$

By (6) we can write

$$f(\hat{p}) = \left(n - 2 - \frac{1}{4}(k-2)s\right)^2 \left(1 - \frac{(k-2)(s-2)}{2n-4}\right).$$

Then

$$f(\hat{p}) = \left((n-2)^2 - \frac{(n-2)(k-2)}{2}s + \frac{(k-2)^2}{16}s^2\right) \cdot \left(1 - \frac{(k-2)(s-2)}{2n-4}\right). \tag{9}$$

We are interested in the case with $\hat{p} \geq 3$. By (8) we get $n \geq 15k - 28$.

**Example 1:**
By the formula (9) for $k = 6$ we get

$$f(\hat{p}) = \left((n-2)^2 - 2(n-2)\left(1+\sqrt{2n-3}\right) + \left(1+\sqrt{2n-3}\right)^2\right) \cdot \left(1 - \frac{2\sqrt{2n-3}-2}{n-2}\right).$$

Let us set

$$n = 2t^2 + 2 \geq 15k - 28 = 62. \tag{10}$$

By (8) for even $t$ we have

$$\lfloor \hat{p} \rfloor = \left\lfloor \frac{1}{4} + \frac{1}{4}\sqrt{4t^2+1} \right\rfloor = \left\lfloor \frac{1}{4} + \frac{t}{2} \right\rfloor = \frac{t}{2}. \tag{11}$$

Note that by (7)

$$f(\lfloor \hat{p} \rfloor) = 4t(t-1)^2(t-2), \tag{12}$$

and

$$f(\lceil \hat{p} \rceil) = 4t(t^2-t-2)^2 \frac{1}{t+2} > f(\lfloor \hat{p} \rfloor),$$

for $t > 2$.

We can note that by the formula (6) we get extremal 2-trees for the case

$$W_6(G) = \frac{1}{2} f(\lfloor \hat{p} \rfloor).$$

Now we compare this value with $W_k(G)$ for $p > 2$. By the formula (3) for $p = 2$ we get

$$W_6(G) \le (t^2+1)^2 - 5(2t^2+2) + 25 = t^4 - 8t^2 + 16.$$

So we get the following inequality

$$\frac{1}{2} f(\lfloor \hat{p} \rfloor) > t^4 - 8t^2 + 16.$$

By the formula (12) we get

$$2t(t-2)(t-1)^2 > t^4 - 8t^2 + 16. \tag{13}$$

The inequality (13) is equivalent to the following one

$$2t(t-1)^2 > (t-2)^2(t+2). \tag{14}$$

Suppose now that $t = 6$, then by (10) we have $n = 2 \cdot 6^2 + 2 = 74$ and by (9) we have $\lfloor \hat{p} \rfloor = 3$. The inequality (14) holds in this case. So we obtained the maximum $W_6(G) = 3 \cdot 20^2$ for the 2-tree $G$ with parameters $n = 74,\ k = 6,\ p = 3$ and $|A_i| = (n-14)/3 = 20$.

An extremal 2-tree is presented below in Fig. 2.

**Example 2:**

By the formula (9) for $k = 4$ we get

$$f(\hat{p}) = \left( (n-2)^2 - (n-2)\left(1 + \sqrt{4n-7}\right) + \frac{1}{4}\left(1 + \sqrt{4n-7}\right)^2 \right) \cdot \left(1 - \frac{\sqrt{4n-7} - 1}{n-2}\right).$$

We have:

$$n = t^2 + 2 \ge 15k - 28 = 15 \cdot 4 - 28 = 32. \tag{15}$$

By (8) for even $t$ we have

$$\lfloor \hat{p} \rfloor = \left\lfloor \frac{1}{4} + \frac{1}{4}\sqrt{4t^2+1} \right\rfloor = \left\lfloor \frac{1}{4} + \frac{t}{2} \right\rfloor = \frac{t}{2}. \tag{16}$$

and then by (7)

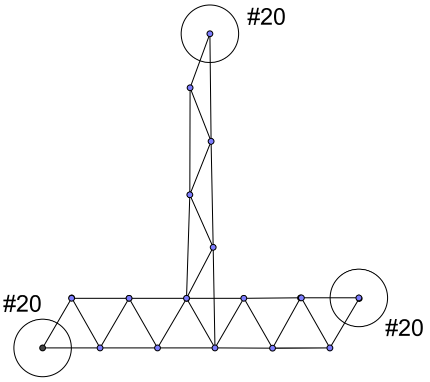


Fig. 2. An extremal graph of order $n = 74$ with $k = 6$ and three groups $|A_i| = 20,\ i = 1, 2, 3$.

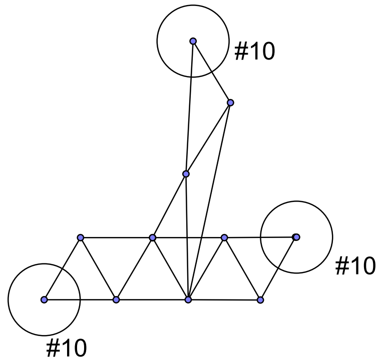


Fig. 3. An example of extremal graph for $k = 4$.

$$f(\lfloor \hat{p} \rfloor) = t(t-2)(t-1)^2,$$

and

$$f(\lceil \hat{p} \rceil) = t(t^2 - t - 2)^2 \frac{1}{t+2},$$

We can note that by the formula (6) we get

$$W_4(G) = \frac{1}{2} f(\lfloor \hat{p} \rfloor).$$

By the formula (3) for $p = 2$ and $k = 4$ we get $W_4(G) = \frac{1}{4}t^4 - 2t^2 + 4$.

Now we get the following inequality

$$f(\lfloor \hat{p} \rfloor) = t(t-2)(t-1)^2 > t(t^2 - t - 2)^2 \frac{1}{t+2} = f(\lceil \hat{p} \rceil).$$

The above inequality is equivalent to the following one

$$(t+2)(t-2)(t-1)^2 > (t^2 - t - 2)^2. \tag{17}$$

Suppose now that $t = 6$, then by (15) we have $n = 6^2 + 2 = 38 > 32$ and $\lfloor \hat{p} \rfloor = 3$. So the inequality (17) holds in this case and we have the maximum $W_4(G) = 3 \cdot 10^2$ for the 2-tree $G$ with parameters $n = 38,\ k = 4,\ \hat{p} = 3$ and $|A_i| = (n-8)/3 = 10, i = 1, 2, 3$.

An extremal 2-tree is presented in Fig. 3.

In general case we have the following result.

Let $p_- = \lfloor \hat{p} \rfloor$ and $p_+ = \lceil \hat{p} \rceil$ where $\hat{p}$ is defined in (8) . We present a theorem for 2-trees of order $n$ equal to $g(k)$, where $g(k)$ is some function defined in the proof.

**Theorem 1.** *Let $n$ and $k$ be integers. For each even integer $k \ge 4$ there exists a 2-tree $G$ of order $n$ with extremal generalized Wiener polarity index $W_k(G)$ and with $p_- \ge 3$ or $p_+ \ge 3$ groups of pendant vertices for $n = g(k)$ where $g(k)$ is some function in variable $k$. Then we have an infinite family of such 2-trees.*

**Proof:** By (7) and (8) we have $p_- \le \hat{p} \le p_+$ and $W_k(G) = \frac{1}{2}\max\{f(p_+), f(p_-)\}$, where

$$f(p_-) = \left((n-2)^2 + p_-^2(k-2)^2\right)\left(1 - \frac{1}{p_-}\right) - 2(k-2)(n-2)(p_- - 1).$$

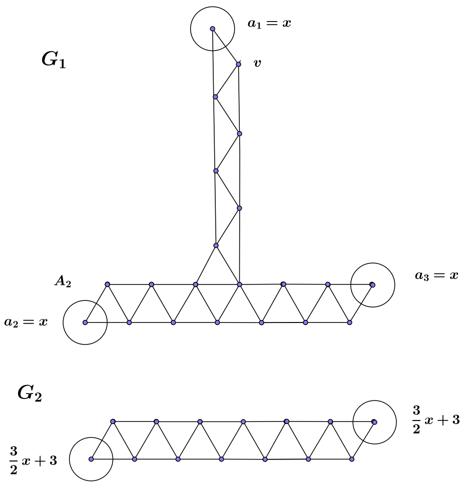


Fig. 4. Examples of 2-trees of the same order with diameter $k = 7$, where $W_7(G_1) > W_7(G_2)$ for even $x \geq 12$.

We get the inequality

$$\frac{1}{2}f(p_-) > \left\lfloor\frac{n}{2}\right\rfloor\left\lceil\frac{n}{2}\right\rceil - (k-1)(n-(k-1)).$$

Hence

$$p_-^3(k-2)^2 - p_-^2(k-2)(2n+k-6) + p_-\left((n-2)^2 - 2\left\lfloor\frac{n}{2}\right\rfloor\left\lceil\frac{n}{2}\right\rceil + 4kn - 6n - 2k^2 + 6\right) - (n-2)^2 > 0.$$

Similarly we can compare

$$\frac{1}{2}f(p_+) > \left\lfloor\frac{n}{2}\right\rfloor\left\lceil\frac{n}{2}\right\rceil - (k-1)(n-(k-1)).$$

This two above inequalities are equivalent to the following one:

$$n^2 - n(12((p-2)k - 2p) + 52) + 52 - 12k^2 + 6p((p-2)(k-2)^2 + (k-2)(k+2)) > 0$$

where $p = p_+$ or $p = p_-$.

By solving this inequality we can construct 2-trees $G$ with $p_- \geq 3$ or $p_+ \geq 3$ groups of pendant vertices with extremal generalized Wiener polarity index $W_k(G)$. It is enough to take $g(k) = 2 + p(k - 2 + a)$, where $a = |A_i|$ for each integer $a \geq (k-2)\max\{11, 2(p-1)\}$ and $i = 1, ..., p$ with $p = p_-$ or $p = p_+$. It follows by formula (8). This is the end of the proof.

In the theorem we are presenting the results for even $k$. Note that for odd $k$ the generalized Wiener polarity index $W_k(G)$ for 2-trees of order $n$ with two groups of pendant vertices in general case is not greater than such index for 2-trees of order $n$ with $p = 3$ groups of pendant vertices.
An infinite number of such examples of 2-trees is presented in Fig. 4.

In this paper we proved Theorem 1 for 2-trees of order $n$ with an extremal $W_k(G)$ for given $n$ and $k$. In the future work we wish to find an efficient algorithm for counting $W_k(G)$ for the considered family of graphs.

## REFERENCES

[1] Bondy, J. A., Murty, U. S. R., Graph Theory with Applications, Macmillan London and Elsevier, New York, 1976

[2] Chepoi, V., Klavžar, S.: The Wiener index and the Szeged index of benzenoid systems in linear time. J. Chem. Inf. Comput. Sci. 37, 752-755 (1997); DOI: 10.1021/ci9700079

[3] Deng, H.: On the extremal Wiener polarity index of chemical trees.MATCH Commun. Math. Comput. Chem. 60, 305-314 (2011)

[4] Deng, H., Xiao, H., Tang, F.: The maximum Wiener polarity index of trees with k pendants. Appl. Math. Lett. 23, 710-715 (2010); DOI: 10.1016/j.aml.2010.02.013

[5] Deng, H., Xiao, H., Tang, F.: On the extremal Wiener polarity index of trees with a given diameter. MATCH Commun. Math. Comput. Chem. 63, 257-264 (2010) 123

[6] Deng, X., Zhang, J.: Equiseparability on terminal Wiener index. In: Goldberg, A.V., Zhou, Y. (eds.) Algorithmic Aspects in Information and Management, pp. 166-174. Springer, Berlin (2009); DOI: 10.1007/978-3-642-02158-9 15

[7] Du,W., Li, X., Shi, Y.: Algorithms and extremal problem on Wiener polarity index.MATCH Commun. Math. Comput. Chem. 62, 235-244 (2009)

[8] Dobrynin, A.A., Entringer, R.C., Gutman, I.: Wiener index of trees: theory and applications. Acta Appl. Math. 66, 211-249 (2001); DOI: 10.1023/A:1010767517079

[9] Gutman, I., Zhang, Y., Dehmer, M., Ilić, A.: Altenburg,Wiener, and Hosoya polynomials. In: Gutman, I., Furtula, B. (eds.) Distance in Molecular Graphs-Theory, pp. 49-70. Univerity of Kragujevac, Kragujevac (2012)

[10] Hosoya, H.: Mathematical and chemical analysis of Wiener's polarity number. In: Rouvray, D.H., King, R.B. (eds.) Topology in Chemistry-Discrete Mathematics of Molecules. Horwood, Chichester (2002); DOI: 10.1533/9780857099617.38

[11] Ilić, A., Stevanović, D.: On Comparing Zagreb Indices. MATCH Commun. Math. Comput. Chem. 62, 681-687 (2009)

[12] Ilić, A., Ilić, M.: Generalizations of Wiener Polarity Index and Terminal Wiener Index, Graphs and Combinatorics 29, 1403-1416; 2013; DOI 10.1007/s00373-012-1215-6

[13] Liu, B., Hou, H., Huang, Y.: On the Wiener polarity index of trees with maximum degree or given number of leaves. Comp. Math. Appl. 60, 2053-2057 (2010); DOI:10.1016/j.camwa.2010.07.045

DOI: 10.15439/2015F242
ACSIS, Vol. 5

# New proposed implementation of ABC Method to Optimization of Water Capsule Flight

Jacek Czerniak
Kazimierz Wielki University in Bydgoszcz,
Institute of Technology
ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland
Email: jczerniak@ukw.edu.pl

Grzegorz Śmigielski,
Kazimierz Wielki University in Bydgoszcz,
Institute of Mechanics and Applied Computer Science
Bydgoszcz, Poland
Email: gsmigielski@ukw.edu.pl

Dawid Ewald
Kazimierz Wielki University in Bydgoszcz,
Institute of Technology
ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland
Email: dawidewald@ukw.edu.pl

Marcin Paprzycki
Systems Research Institute of the Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: marcin.paprzycki@ibspan.waw.pl

Wojciech Dobrosielski
Kazimierz Wielki University in Bydgoszcz,
Institute of Technology
ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland
Email: wdobrosielski@ukw.edu.pl

***Abstract*—The physical model of Water Capsule Flight is relatively simple but analytically unsolvable. The input data includes the mass of the capsule, velocity, altitude, aerodynamic coefficients of the capsule, and horizontal and vertical winds. The ABC optimization is focused on those attributes. This article is a part of the series dedicated to Inspired by Nature Methods of AI and their implementation in the mechatronic systems. A bag filled with water is an excellent source of explosion-produced water spray which can be used for extinguishing large fires or for other purposes. The paper presents theoretical models of flight of a bag filled with water, dropped from an aircraft moving horizontally. Results of numerical computations based on this model are compared with results of measurements for the trajectory of a bag dropped from a helicopter. A description of the experimental and numerical setup for this experiment are also discussed.**

## I. Introduction

BEHAVIOR of many animal species in nature is similar to the swarm behavior. Shoals of fish, flocks of birds and flocks of land animals are created as a result of the biological drive to live in a group. Specific individuals belonging to a flock or a shoal are characterized by higher survival probability because predators or raptors usually attack only one individual. Group movement is characteristic for flocks of birds and other animals as well as shoals of fish. Flocks of land animals react quickly to changes of movement direction and velocity of neighboring individuals. Herd behavior is also one of the main characteristic features of insects living in colonies (bees, wasps, ants, termites) Communication between individual insects of the swarm of social insects has already been thoroughly studied and is still subject of studies. The systems of communication between individual insects contribute to creation of "collective intelligence" of swarms of social insects [4], [11]. Thus the term "Swarm intelligence" emerged, meaning the above mentioned "collective intelligence" [6], [10], [9], [15], [16]. The swarm intelligence is part of the Artificial Intelligence as per examination of activities performed by separate individuals in decentralized systems [14]. The Artificial Bee Colony (ABC) metaheuristics has been introduced quite recently as a new trend in the Swarm intelligence domain [17][13][2]. Artificial bees represent agents solving complex combinatorial optimization problems. This article presents proposed optimization of water capsule flight using ABC method [10][9]. Data obtained from real water capsule flights developed for firefighting was used here. A very efficient way of water spray formation is explosion method consisting in detonation of an explosive placed in a water container [19]. Water sufficiently eliminates undesired consequences of detonation, which provides potential possibility for applications of that method. Water spray can be used, e.g. to extinguish fire and to neutralize contaminated areas [12][5]. Water capsule suspended under a helicopter or another aircraft enables fast transport of water to the area of airdrop. Described system allows automatic release of the water capsule at such a distance from the target so that, after some time of its free fall, it is located over the target at the specified altitude and then detonated to generate spray which covers specified area of the ground [5][18].

## II. Physical methodology of the water capsule flight analysis

In principle the problem of delivering a water capsule to a given point on the ground is very similar to the problem of hitting a surface target by a bomber with an unguided bomb. There are, however, two problems that make difficult a direct

978-83-60810-66-8/$25.00©2015, IEEE

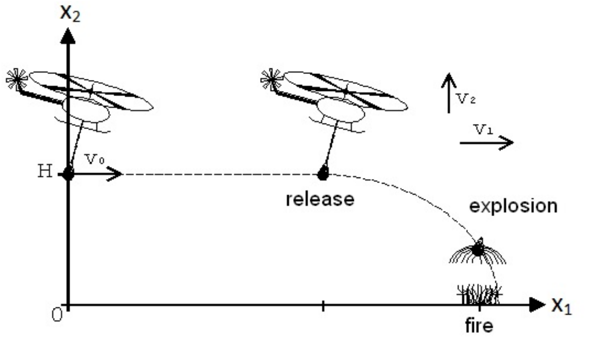


Fig. 1. Schematic view of the procedure of delivering water-capsule to a designed point.

application of the procedures used by military aviation. The first follows from the fact that such procedures, as majority of procedures used by the military are either classified as a whole or comprise classified crucial components [1] [7]. The second problem is connected with much higher safety standards that must be observed in the case of placing water-capsule "in target". It seems then more reasonable to develop procedures from the very beginning than to try to adopt non-classified components of similar military procedures. The ultimate objective consists in developing a high precision system of delivering by an aircraft (presumably a helicopter) a water-capsule to a defined point where it should be exploded in order to generate cloud of water-spray playing role of the fire-extinguishing agent. A scheme of such procedure is shown in Fig. 1.

Designing a suitable system must be based on theoretical models that can serve as a foundation of numerical programs. The models are founded on the assumption that the water-capsule moves in the air under the influence of a constant and vertical gravitational force and of the Bernoulli drag (pressure drag) that acts against its motion with respect to the air and is proportional to the square of the velocity of this motion. After denoting the velocity by $\overrightarrow{v}$ one can write the following formula the drag force.

$$\overrightarrow{O} = \frac{c\rho A}{2} v \overrightarrow{v}, \tag{1}$$

where $v = |\overrightarrow{v}| = \sqrt{v_1^2 + v_2^2}$,
$c$ is the drag coefficient depending on the shape of the moving body, $\rho$ denotes density of the air, and A is the frontal cross-section of the body.

### A. Equations describing flight of a water capsule

A water capsule dropped from a horizontally moving aircraft (e.g. helicopter) falls down under composite action of the drag force that has both vertical and horizontal components and the gravitational force that acts all time vertically. Introducing Cartesian coordinates: the horizontal one $x_1$ and the vertical one $x_2$, one can write equations of motion in the form

$$\dot{v_1} = -\frac{c_1\rho A_1}{2M}\sqrt{v_1^2 + v_2^2}v_1, \dot{v_2} = -\frac{c_2\rho A_2}{2M}\sqrt{v_1^2 + v_2^2}v_2 - g \tag{2}$$

where $v_1$ and $v_2$ are the horizontal and vertical coordinates of the capsule's velocity respectively, $M$ is its mass and $g$ denotes gravitational acceleration. One has to do with a set of ordinary, first order, nonlinear differential equation with respect to the Cartesian coordinates of capsules velocity. Having these equations solved, one can obtain coordinates of the capsule by simple integration coordinates of velocity with respect to time. Unfortunately, the equations (2) cannot be solved analytically without far going simplifications. It is so due to the coupling square root term. As such, one has to apply numerical methods for solving the equations.

### B. Numerical solutions

In this case the standard fourth order Runge-Kutta method was used, and numerical computations were performed inside the MATLAB environment. In practice some additional work aimed, e.g., on optimization of the length of the step of integration, had to be done, but we will not go into technical details [14].

The solution is obtained for standard initial conditions given by the equations

$$v_1(0) = v_0, v_2(0) = v_0 \tag{3}$$

which corresponds to horizontal motion of the water-capsule at the moment of release. Provided the value of the drag coefficient $c$ is known, one can obtain both components of capsule's velocity as functions of time. Since the main objective consists in computing trajectory of the capsule, one has to compute its horizontal and vertical component using integrals

$$\begin{aligned} x_1(t) &= \int_0^t v_1(\tau)d\tau + x_1(0), \\ x_2(t) &= \int_0^t v_2(\tau)d\tau + x_2(0) \end{aligned} \tag{4}$$

that, in general, have to be computed numerically since the functional form of $v_1$ and $v_2$ with respect to time are not known. The numerical solution of equations for the components $v_1$ and $v_2$ of the capsule's velocity has one more advantage. After some modifications such a procedure can be applied to the problem of flight in the air moving with respect to the ground. In fact, equations (2) describe velocity of the capsule with respect to the ground under the assumption that the air is still. If, however, velocities of wind and that of ascending or descending current are considerable, the equations have to be modified

$$\dot{v_1} = -\frac{c_1 \rho A_1}{2M} \sqrt{\tilde{v_1}^2 + \tilde{v_2}^2} \tilde{v_1}$$
$$\dot{v_2} = -\frac{c_2 \rho A_2}{2M} \sqrt{\tilde{v_1}^2 + \tilde{v_2}^2} \tilde{v_2} - g \tag{5}$$

where

$$\tilde{v_i} = \tilde{v_i} - \tilde{V_i}, i = 1, 2 \tag{6}$$

are coordinates of the capsule's velocity with respect to the air; $V_1$ denotes velocity of wind and $V_2$ velocity of vertical current (a further generalization, we will not discuss here, would be taking into account the fact that strong and random winds make the problem 3-dimensional instead of 2-dimensional planar problem of a flight in the still air).

Numerical solution of equation of motion requires inserting numerical data from the very beginning. Some of them like the mass $M$ of the capsule or the density of the air are at hand, but the drag coefficients $k_1 = cA_1$and $k_2 = cA_2$, appearing in (2) and (5) have to be determined from experimental data.

## III. ABC APPLICATION TO OPTIMIZATION OF FUEL CONSUMPTION OF A HELICOPTER

### A. Numerical solutions

Artificial bee colony (ABC) is a model proposed in 2005 by a Turkish scientist Dervis Karaboga [10][3][9]. Like other algorithms described herein, it is also based on herd behavior of honey bees. It differs from other algorithms in the application of higher number bee types in a swarm. After the initialization phase, the algorithm consists of the following four stages repeated by iteration until the number of repetitions specified by the used is competed:

- Employed Bees stage,
- Onlooker Bees stage,
- Scout Bees stage,
- storage of the best solution so far.

The algorithm starts with initialization of the food source vectors $x_m$, where $m = 1...SN$, while $SN$, is the population size. Each of those vectors stores $n$ values $x_m, i = 1...n$, that shall be optimized during execution of that method. The vectors are initialized using the following formula:

$$x_{mi} = l_i + rand(0, 1)x(u_i - l_i) \tag{7}$$

where:

$l_i$-lower limit of the searched range,

$u_i$- upper limit of the searched range,

Bees adapted to different tasks participate in each stage of the algorithm operation. In case of ABC, there are 3 types of objects involved in searching:

- Employed Bees - bees searching points near points already stored in memory,
- Onlooker Bees - objects responsible for searching neighborhood of points deemed the most attractive,
- Scout Bees - (also referred to as scouts) this kind of bees explores random points not related in any way to those discovered earlier.

Once initialization is completed, Employed Bees start their work. They are sent to places in the neighborhood of already known food sources to determine the amount of nectar available there. Results of the Employed Bees work are used by Onlooker Bees. Onlooker Bees randomly select a potential food source using the following relationship:

$$v_i = x_{mi} + \varphi_{mi}(x_{mi} + x_{ki}) \tag{8}$$

where:

$v_i$-vector of potential food sources,

$x_k$- randomly selected food source,

$\varphi_{mi}$- random number from the range [-a,a] Once the vector is determined its fitting is calculated based on the formula dependent on the problem being solved and the fitting $v_m$ is compared with $x_m$. If the new vector fits better than the former one, then the new replaces the old one. Another phase of the algorithm operation is the Onlooker Bees stage. Those bees are sent to food sources classified as the best ones and in those very points the amount of available nectar is determined. The probability of the $x_m$ source selection is expressed with the formula:

$$p_m = \frac{fit_m(x_m)}{\sum_{k=1}^{SN} fit_k(x_k)} \tag{9}$$

where:

$fit_m(x_m)$- value of fitting functions for a given source.

Obviously, when onlooker bees gather information on the amount of nectar, such data is compared with results obtained so far and if the new food sources are better, they replace the old ones in the memory. The last phase of this algorithm operation is exploration by scouts. Bees of that type select random points from the search space and then check nectar volumes available there. If newly found volumes are higher than the volumes stored so far, they replace the old volumes. The activity of those bees makes it possible to explore the space unavailable for the remaining types of bees thus allowing to omit any extremes.

### B. Application of ABC

The reach of the capsule flight is calculated so that the distance from the helicopter does not exceed 140 m and then the initial velocity V0 and the altitude Z are optimized. The fuel consumption at the power of 2225KM - 292 g /KMh (i.e. 292 g of fuel per horse power per hour) is assumed as the cost. At the moment the program estimates the results only approximately, but author believes that he shall be able to make the results more real in the near future.

```
Random selection of the initial altitude and velocity of the helicopter;
REPEAT
The selected altitude is put into the water capsule flight reach formula;
The selected velocity is put into the water capsule flight reach formula;
```

*Then we calculate the function of the cost of rising the*
*helicopter to the specified altitude and accelerating it to*
*the specified velocity so that the capsule is dropped not further than 140 m*
*away from the target;*
*Sources;*
*The verified velocity and altitude are replaced by new values;*
*The best velocity and altitude are stored in the memory;*
*UNTIL (the conditions are met)*

The main underlying idea of the optimization is to select such altitude and velocity of the helicopter that shall enable the capsule to reach the target. The path covered by the capsule falling from the helicopter to the vicinity of fire depends on the altitude from which the capsule was dropped. It is obvious that increase of altitude or velocity depends directly on helicopter rotor power. The power to be generated influences specific fuel consumption. As illustrated in the graph, vertical climb of the helicopter at zero horizontal velocity generates huge power demand. One can significantly reduce power needed to climb the helicopter to the specified altitude by increasing its horizontal velocity. This relationship results from the way of generating aerodynamic lift by the helicopter [8]. However, too high horizontal velocity can significantly increase power demand causing increased specific fuel consumption.

## IV. Conclusion

Obviously, the power required during forward flight will also be the function of GTOW (Gross Takeoff Weight). Representative results illustrating the effect of GTOW on the rotor power required are provided in Fig. 2 for a sample helicopter at sea-level (SL) conditions. It should be noted that with increasing GTOW, the excess power available decreases gradually, this phenomenon applies in particular at lower airspeed where the induced power requirement is a higher percentage of the total power. In the subject case, the power available at SL is 2800 hp and for a gas turbine this remains relatively constant versus airspeed. The airspeed value at the intersection of the power required curve and the power available curve indicates the maximum level flight speed. However, the maximum velocity is limited by probable onset of rotor stall and compressibility effects before this point is reached. Multi-objective optimization of a helicopter flight carrying a water capsule is a non-trivial problem. ABC algorithm application enables efficient optimization of fuel costs. Due to high complexity of that problem, one must bear in mind that optimization results may deviate from real results. These can be caused by the fact that wind drag and direction were skipped. The distance at which the helicopter must approach the fire can have significant impact on the final result while the air temperature can significantly influence the fuel demand of the helicopter engine. There is also an issue of the angle at which the capsule is dropped. That aspect can also be taken into account in further studies on ABC application to fuel consumption optimization and as a consequence, on reduction fire extinguishing cost using that method. Summing up, that problem is very complex, which makes it a good example of ABC application.

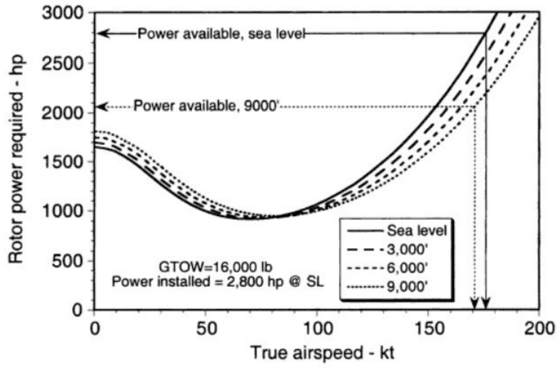


Fig. 2. The graph of fuel consumption versus the velocity and altitude of the helicopter flight.

## References

[1] Angryk, R., Czerniak, J.: Heuristic algorithm for interpretation of multi-valued attributes in similarity-based fuzzy relational databases. International Journal of Approximate Reasoning 51(8), 895–911 (oct 2010)

[2] Czerniak, J.: Evolutionary approach to data discretization for rough sets theory. Fundamenta Informaticae 1-2, 43–61 (2009)

[3] Czerniak, J., Ewald, D., Macko, M., Śmigielski, G., Tyszczuk, K.: Approach to the monitoring of energy consumption in eco-grinder based on abc optimization. Beyond Databases, Architectures and Structures pp. 516–529 (2015)

[4] Czerniak, J., Apiecionek, L., Zarzycki, H.: Application of ordered fuzzy numbers in a new ofnant algorithm based on ant colony optimization. Communications in Computer and Information Science 424, 259–270 (2014)

[5] Dygdała, R., Stefański, K., Śmigielski, G., Lewandowski, D., Kaczorowski, M.: Aerosol produced by explosive detonation. Measurement Automation and Monitoring 53(9), 357–360 (2007)

[6] Ewald, D., Czerniak, J., Zarzycki, H.: Approach to solve a criteria problem of the abc algorithm used to the wbdp multicriteria optimization. Intelligent Systems' 2014 pp. 129–137 (2014)

[7] Ganesan, P.K., Angryk, R., Banda, J., Wylie, T., Schuh, M.: Spatiotemporal co-occurrence rules, new trends in databases and information systems pp. 27–35

[8] Gordon Leishman, J.: Principles of Helicopter Aerodynamics. Cambridge University Press (2002)

[9] Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (abc) algorithm. Journal of Global Optimization 39, 459–171 (2007)

[10] Karaboga, D., Gorkemli, B.: A quick artificial bee colony (QABC) algorithm and its performance on optimization problems. Applied Soft Computing 23, 227–238 (2014)

[11] Kowalski, P., Łukasik, S.: Experimental study of selected parameters of the krill herd algorithm. Intelligent Systems'2014: Proceedings of the 7th IEEE International Conference Intelligent Systems ISŠ2014 1, 473–477 (2014)

[12] Liu, Z., Kim, A.K., Carpenter, D.: Extinguishment of large cooking oil pool fires by the use of water mist system. Combustion Institute/Canada Section, Spring Technical Meeting pp. 1–6 (may 2004)

[13] Marbac-Lourdelle, M.: Model-based clustering for categorical and mixed data sets (2014)

[14] Plucński, M.: Mini-models-local regression models for the function approximation learning artificial intelligence and soft computing. Lecture Notes in Artificial Intelligence 7268 edited by L. Rutkowski et al pp. 160–167 (2012)

[15] Reina, M.D., Trianni., V.: Towards a cognitive design pattern for collective decision-making. in swarm intelligence - proceedings of ants 2014 - ninth international conference. Lecture Notes in Computer Science 8667, 194–205 (2014)

[16] Roeva, O., Slavov, T.: Firefly algorithm tuning of pid controller for glucose concentration control during e. coli fed-batch cultivation process. Proceedings of the Federated Conference on Computer Science and Information Systems p. 455Ű462 (2012)

[17] Sameon, D., Shamsuddin, S., Sallehuddin, R., Zainal, A.: Compact classification of optimized boolean, reasoning with particle swarm optimization. Intelligent Data Analysis 16 IOS Press pp. 915–931 (2012)

[18] Śmigielski, G., Dygdała, R.S., Lewandowski, D., Kunz, M., Stefański, K.: High precision delivery of a water capsule. theoretical model, numerical description, and control system. IMEKO XIX World Congress, Fundamental and Applied Metrology pp. 2208–2213 (2009)

[19] Stebnovskii, S.V.: Pulsed dispersion as the critical regime of destruction of a liquid volume. Combustion, Explosion, and Shock Waves 44(2), 228–238 (2008)

DOI: 10.15439/2015F33
ACSIS, Vol. 5

# Ant Colony Optimization with environment changes: an application to GPS surveying

Antonio Mucherino*, Stefka Fidanova†, Maria Ganzha‡

*IRISA, University of Rennes 1, Rennes, France.
antonio.mucherino@irisa.fr

†BAS, University of Sofia, Sofia, Bulgaria.
stefka@parallel.bas.bg

‡SRI, Polish Academy of Science, Warsaw, Poland.
maria.ganzha@ibspan.waw.pl

***Abstract*—We propose a variant on the well-known Ant Colony Optimization (ACO) general framework where we introduce the environment to play an important role during the optimization process. Together with diversification and intensification, the environment is introduced with the aim of avoiding the search to get stuck at local optima. In this work, the environment is simulated by means of the Logistic map, that is used in ACO for perturbing the update of the pheromone trails. Our preliminary experiments show that our environmental ACO (*e*ACO), with variable environment, outperforms the standard ACO on a set of instances of the GPS Surveying Problem (GSP).**

## I. Introduction

GLOBAL optimization consists in finding the global optimum of a given objective function which is generally subject to a given number of constraints. Many real-life problems can be formulated as a global optimization problem where the objective function often contains several local optima. A major issue in solving such problems is to "escape" from local optima for converging towards the global optimum of the objective function [12].

Meta-heuristics are general-purpose methods for global optimization. They are usually based on the simulation of animal or natural behaviors. They consist in a list of actions (generally repetitive) to be performed for finding an approximation of the global optimum of a given (and generally hard) optimization problem [23]. In recent years, several meta-heuristic approaches have been proposed in the scientific literature. A classical example is the Simulating Annealing (SA) proposed in the 80s, which is still used in some applications, such as the ones where it is necessary to deal with biological molecules [16]. Other well-known examples of meta-heuristics are the Genetic Algorithms (GAs), Differential Evolution (DE), the Tabu Search (TS), the Variable Neighborhood Search (VNS), etc. In this paper, we consider the Ant Colony Optimization (ACO) approach (see Section II-A). For a quick survey on meta-heuristics and the main references for the meta-heuristics mentioned above, the reader can refer, for example, to [17].

Every meta-heuristic is developed in order to find the best trade-off between two main concepts: *diversification* and *intensification* [23]. While the former tries to widely extend the search in the domain of the optimization problem, the latter improves candidate solutions by focusing on local neighbors of the current best known solutions. In ACO, diversification is guaranteed by the simulation of the typical ant behavior, while intensification is performed by applying a local search to a set of candidate solutions.

In this work, we introduce the *environment* in our ACO implementation. Pheromone updates are not supposed to be performed only on the basis of found solutions, but also on the basis of the current environment "surrounding" the ants. As an example, particular real-life environment conditions, such as strong wind, may alter the perception of the deposited pheromone. We will simulate environment changes by employing the Logistic map [25]. This simulation will help our artificial ants to escape from local optima, which may otherwise focus on firstly discovered paths and completely skip better ones. We warn the reader that the Logistic map has already been employed in optimization for performing chaotic searches [3], [26]. However, its use is different from the one considered in this work for the simulation of environment changes.

This paper is organized as follows. In Section II, we will extend the classic ACO approach for managing environment changes, which will affect the fitness values used in the pheromone updating rule during the execution of ACO. We will refer to our extended ACO as "environmental ACO" (*e*ACO). In Section III, we will describe the problem we consider in our preliminary computational experiments: the GPS Surveying Problem (GSP). Computational experiments will be presented in Section IV: they show that our *e*ACO, with environment changes, outperforms the standard ACO on the set of considered instances. Finally, conclusions will be given in Section V.

978-83-60810-66-8/$25.00©2015, IEEE

## II. ACO WITH ENVIRONMENT CHANGES

We propose an extension of ACO for managing variable environments during the execution of the meta-heuristic. In Section II-A, we give a brief overview of this meta-heuristic search, and the reader who is interested in additional details is referred to the references given in the same section. In our ACO implementation, the pheromone updating rule is modified in order to take into consideration the current environment, and not only the objective function values of obtained solutions. Section II-B shows how to simulate environment changes in ACO. The environment is simulated by means of the Logistic map.

### *A. Ant Colony Optimization*

Ants foraging for food deposit a substance named pheromone on the paths they follow. An isolated ant would just move randomly. Ants encountering previously laid pheromone marks are however stimulated to follow the same paths. This way, the pheromone trails are reinforced around the optimal ones, so that the probability for the other ants to follow optimal paths increases with time. The repetition of this mechanism represents the auto-catalytic behavior of ant colonies in nature [1], [6].

Ant Colony Optimization (ACO) is inspired by this ant behavior. A colony of artificial ants working into a mathematical space is simulated. These ants search for candidate solutions of a given optimization problem, while possible paths are marked by artificial pheromone for guiding other ants in the regions of the search space where good-quality solutions were already found. In ACO, therefore, the artificial ants generally create a sort of environment by themselves, by depositing the pheromone on marked paths. As explained in details in Section II-B, we will perturb this environment by means of the Logistic map.

The ants' search space is represented by the so-called *construction graph*, which is a weighted graph $G_C = (S_C, E_C, \eta)$ where vertices in $S_C$ are solution components and edges in $E_C$ indicate the possibility to combine the two connected components for obtaining a partial solution. The weight $\eta : (u,v) \in E_C \longrightarrow \Re$ associated to each edge $(u,v)$ is named heuristic information, whose equation is generally tailored to the problem at hand. A path on $G_C$ allows to combine several partial solutions and to construct one complete solution.

Alg. 1 is a sketch of the ACO-based meta-heuristic. The transition probability $p_{uv}$, necessary in the algorithm when the ants need to decide on which edge $(u,v)$ to walk, is based on the heuristic information $\eta_{uv}$ and on the current pheromone level $\tau_{uv}$:

$$p_{uv} = \tau_{uv}^{\alpha}\eta_{uv}^{\beta}\left[\sum_{(u,w)\in E_S : w \not\subset X}\left(\tau_{uw}^{\alpha}\eta_{uw}^{\beta}\right)^{-1}\right], \tag{1}$$

where $\alpha$ and $\beta$ are transition probability parameters. When the algorithm starts, small positive values are given to every $\tau_{uv}$, which represent the current pheromone values on the edges $(u,v) \in E_C$. Then, every time a new solution is identified, the edges considered by the ants are marked with a new level of pheromone. In ACO, one possible updating rule (for a minimization problem) is the following:

$$\tau_{uv} = \tau_{uv} + \frac{1}{f(X)}, \tag{2}$$

where $X$ is the current solution, and $f$ is the objective function of the considered problem.

**Algorithm 1** Ant Colony Optimization

```
1: ACO  (in: $N$, $G_C$, $\alpha$, $\beta$;  out: $X_{best}$)
2: let $X_{best} = \emptyset$;
3: while (stopping criteria not satisfied) do
4:    for ($k = 1$, $N$) do
5:       place $k^{th}$ ant on a random vertex $u \in S_C$;
6:       let $X = \{u\}$;
7:       while ($X$ is incomplete) do
8:          select the vertex $v$ in the star of $u$ having higher
            probability $p_{uv}$ (see equ. (1));
9:          let $X = X \cup \{v\}$;
10:         let $u = v$;
11:      end while
12:      update pheromone (see equ. (2));
13:      apply local search starting from $X$ (optional);
14:      if ($X$ is better than $X_{best}$) then
15:         let $X_{best} = X$;
16:      end if
17:   end for
18: end while
```

In ACO, the general ant behavior allows to perform a wide search on the search domain (diversification), while the local search (see line 13 in Alg. 1) from constructed solutions $X$ allows to focus on promising neighbors (intensification). Our implementation of ACO makes use of MaxMin Ant System (MMAS). The reader is referred to the paper [22] for a wider explanation of the ACO implementation considered in this work.

### *B. Simulating the environment*

The Logistic map is a quadratic dynamical equation proposed in 1938 as a demographic model [25]. It is a rather simple quadratic polynomial

$$x_{n+1} = rx_n(1 - x_n), \qquad n > 0, \tag{3}$$

where $x_n$ represents the population size at time $n$ and $r$ is a constant, named growth coefficient. Given $x_0 \in [0,1]$ and a value for $r \in [0,4]$, this dynamical equation can either converge or be chaotic. In the first case, given any $x_0 \in [0,1]$, $x_n$ tends to the so-called "attraction domain". In the second case, $x_n$ never converges, but it can rather take, in an apparent random way, any possible value in the range $[0,1]$.

Fig 1 shows the behavior of the Logistic map for different values of $r$ in the range $[2,4]$ (its behavior is linear in the range $[0,2]$). On the $x$-axis, we consider a discrete subset of 3000 equidistant values for $r$ between 2 and 4; on the $y$-axis,

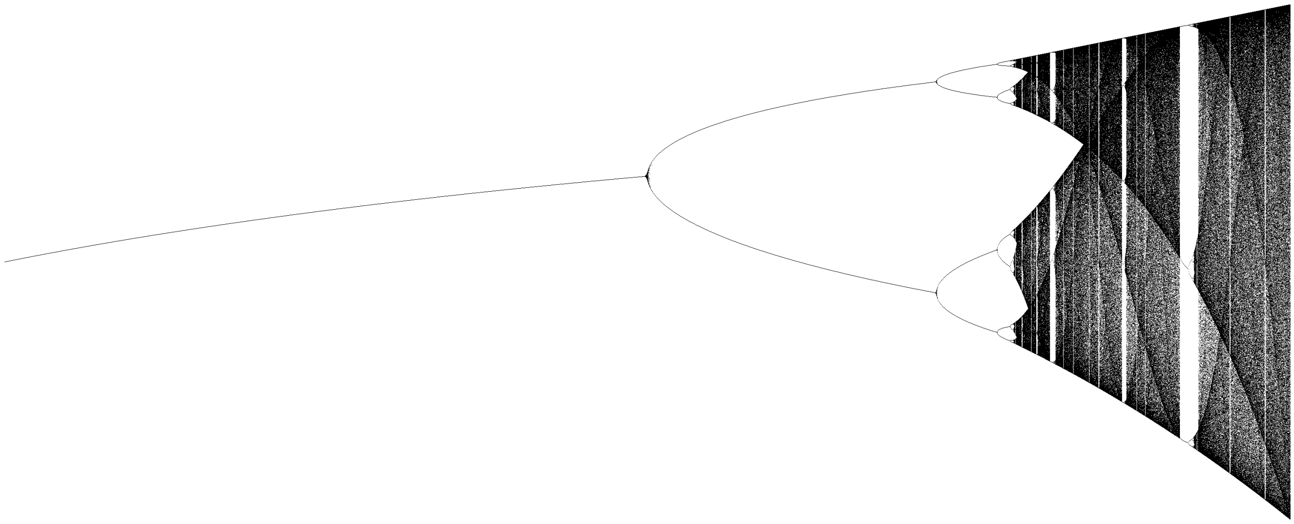


Fig. 1. The behavior of the Logistic map for different values of $r$ (on the $x$-axis, for $r = 2$ to 4 in the figure). The attraction domain can be either regular or chaotic.

for every considered value for $r$, we report the corresponding attraction domain. In order to identify the attraction domains, we take 1500 equidistant points in the interval $[0,1]$ and we apply equation (3) 1000 times for each of them. For small values of $r$, the Logistic map always converges to one single point, i.e. the attraction domain consists of one point only. The first bifurcation appears when $r = 3$, where the attraction domain consists of 2 points; then there is another bifurcation when $r = 1 + \sqrt{6}$, where the attraction domain consists of 4 points. For larger values for $r$, the Logistic map experiences other bifurcations, and it can be chaotic for some subintervals of $r$. However, in these chaotic regions, it is still possible to identify regular attraction domains. For example, in Fig 2, the same graphic reported in Fig. 1 is zoomed in the region $r = [3.901, 3.908]$, where this phenomenon is clearly shown. Regular regions, that can be glimpsed in Fig. 1, still contain bifurcations. Moreover, we can notice that the whole graphic in Fig. 1 reappears in our zoomed region. Other regular attraction domains can be identified by looking at tighter subintervals of $r$, as well as other copies of the entire graphic. The graphic in Fig. 1 is in fact a fractal, because of its self-similarity [7], [19].

We simulate regular and chaotic changes of environment in ACO by introducing the Logistic map in equation (2), which is used in ACO for updating the pheromone trails. In the hypothesis the objective function of the considered problem is positive and greater than 1, the term $1/f(X)$ in equation (2) has always values ranging between 0 and 1. It can therefore take the place of $x_0$ in the Logistic map, so that a perturbed value $x_1$ can be computed, for a given value of $r$ in $[0,4]$. The equation for updating the pheromone therefore becomes:

$$\tau_{uv} = \tau_{uv} + r \cdot \frac{1}{f(X)} \cdot \left(1 - \frac{1}{f(X)}\right). \tag{4}$$

With this simple change in the rule for updating the pheromone, we artificially perturb the environment of the ants, which would otherwise only depend on the solution fitness values. Different values for $r$ can produce different environment changes, depending on the behavior of the Logistic map. For values of $r$ for which the Logistic map converges, the pheromone levels added to $\tau_{uv}$ tend to be constant, reducing in this way the effects of good-quality solutions, that might mislead the ants towards a local optimum. For values of $r$ for which the Logistic map behaves instead chaotically, the environment is dominant on the choices of the ants, as the pheromone update mostly depend on the simulated environment, rather than on the actual fitness value.

We refer to ACO with environment changes as *environmental* ACO (*e*ACO). In this work, we present some preliminary experiments (see Section IV) where *e*ACO is employed for solving the GSP (see next section).

## III. GPS SURVEYING

The Global Positioning System (GPS) was originally developed in the US for military purposes, even if it was soon after used as well for civil applications [11]. It consists of a certain number of satellites that constantly orbit around earth and that are provided with sensors able to communicate with machines located on earth. The power that is necessary for establishing a satellite-earth communication allows for estimating the distance between the two communicating machines. Since the machine located on earth lies over a sphere that does not contain the satellite, a very precise information about the distance between the earth surface and the satellite would allow for determining the precise location of the machine on earth [15]. Moreover, the precision in locating sensor machines on earth can still be high when the distance information is not very precise, but the communication with more than one satellite can be established [14].

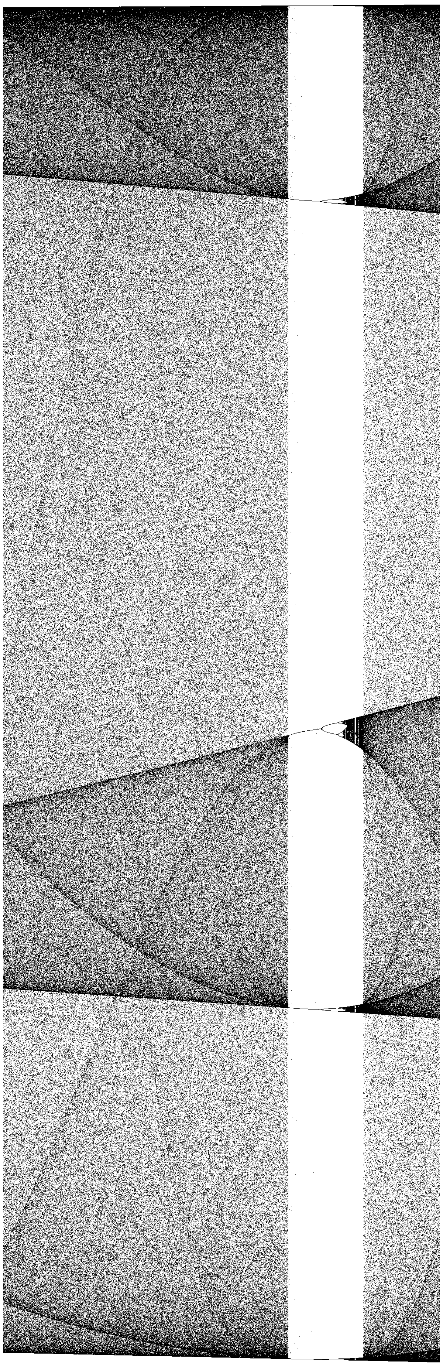


Fig. 2. The behavior of the Logistic map for values of $r$ in the interval $[3.901, 3.908]$. This region of the Logistic map is generally chaotic, but regular attraction domains (with the typical bifurcations) can still be identified.

GPS technology can in fact provide very accurate locations for all sensors forming a given sensor network. The related costs can however be too high when it is necessary to deal with large networks. For this reason, over the last years, researchers have been trying to design and install local ground networks having the task of recording satellite signals with the aim of decreasing the overall network functioning cost [5], [20]. A network is composed by a certain number of *receivers* working at different *stations* in different moments. Therefore, given a certain number of *sessions*, representing the temporarily assignment of a given number of receivers to a set of distinct stations, the problem is to find a suitable *order* for such sessions for reducing the overall cost. This cost is in fact strictly related to the order of the sessions, because receivers need to be moved from one station to another when stepping from one session to another. Therefore, the distance between two involved stations is important for the computation of the costs. However, there are also additional costs that we might need to consider: if the number of working days necessary to perform the operation is more than one, then the need of planning an over-night stop at a company office can make the cost of the operation increase. The session order is also named *session schedule*. Generally, in order to alleviate the impact of measurement errors in the data, at least two receivers per session are considered [24].

The GPS Surveying Problem (GSP) can be formalized as follows. Let

$$S = \{s_1, s_2, \ldots, s_n\}$$

be a set of stations, and let

$$R = \{r_1, r_2, \ldots, r_m\}$$

be a set of receivers, with $m < n$. Sessions can be defined by a function $\sigma : R \longrightarrow S$ that associates one receiver to one station. Considering that no more than one receiver should be assigned to the same station, $\sigma$ can be represented by an $m$-vector $(\varsigma_1, \varsigma_2, \ldots, \varsigma_m)$ containing, for each of the $m$ receivers, the labels of the chosen stations. Since $m < n$ (and generally fixed to 2 or 3 in the applications), the number of permutations of $m$ objects from $n$ distinguishable ones is $n!/(n-m)!$, which can be huge when the network is large. Notice, however, that not all permutations may actually be possible, depending on the problem at hand.

Let $C$ be an $n \times n$ matrix providing the costs $c(\varsigma_u, \varsigma_v)$ for moving one receiver from the station $\varsigma_u$ to the station $\varsigma_v$. This matrix can be symmetric when moving between $\varsigma_u$ and $\varsigma_v$ is independent from the directionality; the non-symmetric case is however more realistic.

An instance of the GSP can be represented by a weighted undirected multigraph $G = (V, E, c)$ where vertices represent sessions $\sigma_v$ and arcs $(\sigma_u, \sigma_v)$ indicate the possibility to switch from session $\sigma_u$ to session $\sigma_v$. The upper bound on the cardinality of $V$ is $n!/(n-m)!$, which corresponds to the maximum number of possible sessions. The weight associated

| | | ACO | *e*ACO | | | |
|---|---|---|---|---|---|---|
| *instances* | $\|V\|$ | | $r = 1$ | $r = 2$ | $r = 3$ | $r = 4$ |
| Malta | 38 | 899.50 | **897.00** | **897.00** | **897.00** | 900.33 |
| Seyshels | 71 | 922.06 | 905.73 | 905.60 | **887.33** | 906.73 |
| kro124p | 100 | 40910.60 | **40725.40** | 40799.30 | 40753.00 | 40803.76 |
| ftv170 | 171 | 3341.93 | 3314.20 | **3313.76** | 3319.83 | 3338.53 |
| rgb323 | 323 | 1665.90 | 1654.40 | **1648.66** | 1649.43 | 1649.53 |
| rgb358 | 358 | 1692.66 | **1679.63** | 1689.00 | 1682.80 | 1685.95 |
| rgb403 | 403 | 3428.56 | 3413.63 | 3392.23 | 3393.76 | **3386.10** |
| rgb443 | 443 | 3765.80 | 3749.93 | 3742.86 | **3742.43** | 3754.50 |

TABLE I
COMPARISON BETWEEN ACO AND *e*ACO ON A SET OF GSP INSTANCES.

to the arcs provides the cost $c(\sigma_u, \sigma_v)$ for moving every receiver from the station $\varsigma_{u,i}$ to the station $\varsigma_{v,i}$, for each $i$:

$$c(\sigma_u, \sigma_v) = \sum_{i=1}^{m} c(\varsigma_{u,i}, \varsigma_{v,i}).$$

The graph $G$ is not simple in general, because it might be feasible to switch from session $\sigma_u$ to session $\sigma_v$, as well as from $\sigma_v$ to $\sigma_u$, but with a different total cost. The problem consists in finding an optimal path on $G$, i.e. a path for which all selected arcs give the minimal total cost, while covering the entire vertex set $V$ [5].

The GSP can be seen as the classic Traveling Salesman Problem (TSP), which asks for determining the optimal route for a salesman to visit a given number of cities while minimizing the traveled distance [13]. If the two cities are replaced with sessions $\sigma_v$ and the distances are replaced with the weights associated to the arcs $(\sigma_u, \sigma_v)$ of our graph $G$, the equivalence between the two problems becomes evident. More precisely, since the weights on the arcs $(\sigma_u, \sigma_v)$ and $(\sigma_v, \sigma_u)$ are generally different, GSP better fits with the Asymmetric TSP (ATSP), where distances between the cities depend upon the order in which the cities are reached. Finally, we also remark that the necessity of an over-night office stop can be formalized by adding a fictive session in $G$ where receivers come back to base offices. This special session might need to be "traveled" more than once: we can therefore say that, in general, the GSP can be seen as a Multiple Asymmetric TPS (MATSP) [2]. The TSP and these variants are NP-hard [18]. Therefore, the GSP is NP-hard as well.

First attempts for solving the GSP were based on the idea of transforming GSP instances into instances of the class of TSP-like problems, and to employ existing methods and algorithms. In [4], a branch-and-bound approach was employed, which is actually able to find the optimal solutions for small GSP instances. As the size of the networks increases, the complexity grows and the time necessary for a branch-and-bound to converge becomes prohibitive. On the other side, in real-life applications, there is generally the need to obtain GSP solutions as fast as possible, even if only approximated ones. Therefore, heuristic approaches particularly developed for this application have been proposed over the last years for identifying optimal or near-optimal session orders [8], [9], [20]. In this work, we consider a set of instances of the GSP for testing our ACO approach with environment changes described in Section II.

## IV. COMPUTATIONAL EXPERIMENTS

We apply our *e*ACO (see Section II) for solving instances of the GSP (see Section III). In this application, the construction graph $G_C$ corresponds to the weighted undirected multigraph $G$ where vertices are sessions $\sigma_u$ and edges indicate the possibility to switch from one session to another. As test cases, we consider data from two real networks: Malta [20], composed by 38 sessions, and Seychelles [21], composed by 71 sessions. We also consider larger instances designed for testing the ATSP, which are freely available on the Internet.[1]

Our *e*ACO implementation is based on Alg. 1, where the transition probability $p_{uv}$ is computed by equation (1), and the pheromone update is performed by applying equation (4). The heuristic information $\eta_{uv}$ is given by the formula:

$$\eta_{uv} = \frac{1}{c(\sigma_u, \sigma_v)},$$

where $c(\sigma_u, \sigma_v)$ is the total cost for switching from session $\sigma_u$ to session $\sigma_v$ (see Section III). Finally, the values for the transition probability parameters $\alpha$ and $\beta$ are fixed to 1 and 2, respectively. These values were identified in previous works as the optimal ones for ACO when solving instances of the GSP [10]. In the following experiments, we will focus on the quality of the found solutions, rather than on the algorithms' performances. In fact, the increase in complexity for using equation (4), rather than equation (2), can be neglected when considering the overall algorithms' complexity. We do not compare our results to the best results currently known for the considered instances. In the present work, our aim is not to improve those results, but only to give a preliminary validation to the usefulness of the presented idea.

Table I shows some computational experiments for different values of $r$. Average values over 30 runs are reported in the table. *e*ACO is able to identify better quality solutions in all experiments and for almost all used values for $r$. This shows that, in fact, a variable environment for the ants, instead of a constant one, gives benefits to the search. For values of $r$ equal to 1, 2 and 3, the Logistic map converges to one unique value; it is instead chaotic for $r = 4$. It seems therefore that

[1] http://www.informatik.uni-heidelberg.de/groups/comopt/software/TSLIB95/ATSP.html

the best results can be achieved when the environment tends to homogenize the pheromone trails. Notice, however, that only one experiment with $r = 4$, when the environment is chaotic, was not able to provide a better solution (w.r.t the one provided by the standard ACO).

## V. Conclusions

We introduced a variant of ACO where the ant environment is not constant but it is rather subject to change over time. This way, in our ACO implementation, there are less chances for the meta-heuristic search to be trapped at a local optimum. In this work, we have tested this idea on instances of the GSP. In order to simulate the environment in our *e*ACO, we employed the well-known Logistic map, which can have either a regular or a chaotic behavior, depending on the values assigned to its parameter $r$.

We believe this is the first research contribution where an "environment" is introduced in a meta-heuristic search. In this work, we showed that this novel idea seems to be promising when working with ACO. Evidently, the idea still needs to be studied in details from a theoretical point of view, as well as from a practical one. The effect of a suitable environment should be studied in conjunction with other meta-heuristic frameworks. Moreover, a much wider experimental analysis should be performed, on a larger set of instances of the GSP, as well as on instances of other known hard problems. This will be our main direction for future works.

## Acknowledgments

This work was partially supported by two grants of the Bulgarian National Scientific Fund: "Efficient Parallel Algorithms for Large Scale Computational Problems" and "InterCriteria Analysis. A New Approach to Decision Making".

## References

[1] V. Atanassova, S. Fidanova, I. Popchev, P. Chountas, *Generalized Nets, ACO-Algorithms and Genetic Algorithm*. In: "Monte Carlo Methods and Applications", K.K. Sabelfeld, I. Dimov, De Gruyter, 39–46, 2012.

[2] T. Bektas, *The Multiple Traveling Salesman Problem: an Overview of Formulations and Solution Procedures*, Omega **34**(3), 209–219, 2006.

[3] L. Chen, K. Aihara, *Chaotic Simulated Annealing by a Neural Network Model with Transient Chaos*, Neural Networks **8**(6), 915–930, 1995.

[4] P. Dare, *Optimal Design of GPS Networks: Operational Procedures*, PhD Thesis, School of Surveying, University of East London, UK, 1995.

[5] P. Dare, H.A. Saleh, *GPS Network Design: Logistics Solution using Optimal and Near-Optimal Methods*, Journal of Geodesy **74**, 467–478, 2000.

[6] M. Dorigo, M. Birattari, *Ant Colony Optimization*. In: "Encyclopedia of Machine Learning", C. Sammut, G.I. Webb (Eds.), Springer, 36–39, 2010.

[7] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, Wiley, 400 pages, 2013.

[8] S. Fidanova, *Hybrid Heuristics Algorithms for GPS Surveying Problem*, Lecture Notes in Computer Science **4310**, Proceedings of the $6^{th}$ International Conference on Numerical Methods and Applications, T. Boyanov, S. Dimova, K. Georgiev, G. Nikolov (Eds.), 239–248, 2007.

[9] S. Fidanova E. Alba, G. Molina, *Memetic Simulated Annealing for GPS Surveying Problem*, Lecture Notes in Computer Science **5434**, Proceedings of the $4^{th}$ International Conference on Numerical Analysis and Its Applications, S. Margenov, L.G. Vulkov, J. Waśniewski (Eds.), 281–288, 2009.

[10] S. Fidanova, E. Alba, G. Molina, *Hybrid ACO Algorithm for the GPS Surveying Problem*, Lecture Notes in Computer Science **5910**, Proceedings of Large Scale Scientific Computing, I. Lirkov, S. Margenov, J. Waśniewski (Eds.), 318–325, 2010.

[11] B. Hofmann-Wellenhof, H. Lichtenegger, J. Collins, *Global Positioning System: Theory and Practice*, Springer, 326 pages, 1993.

[12] R. Horst, P.M. Pardalos, *Handbook of Global Optimization*, Springer, 879 pages, 1995.

[13] J.B. Kruskal, *On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem*, Proceedings of the American Mathematical Society **7**(1), 48–50, 1956.

[14] A. Leick, *GPS Satellite Surveying*, $3^{rd}$ edition, Wirley, 464 pages, 2004.

[15] L. Liberti, C. Lavor, N. Maculan, A. Mucherino, *Euclidean Distance Geometry and Applications*, SIAM Review **56**(1), 3–69, 2014.

[16] T.E. Malliavin, A. Mucherino, M. Nilges, *Distance Geometry in Structural Biology: New Perspectives*. In: "Distance Geometry: Theory, Methods and Applications", A. Mucherino, C. Lavor, L. Liberti, N. Maculan (Eds.), Springer, 329–350, 2013.

[17] A. Mucherino, O. Seref, *Modeling and Solving Real Life Global Optimization Problems with Meta-Heuristic Methods*. In: "Advances in Modeling Agricultural Systems", Springer Optimization and Its Applications **25**, P.J. Papajorgji, P.M. Pardalos (Eds.), 403–420, 2008.

[18] C.H. Papadimitriou, *The Euclidean Travelling Salesman Problem is NP-complete*, Theoretical Computer Science **4**(3), 237–244, 1977.

[19] M. Rani, R. Agarwal, *Generation of Fractals from Complex Logistic Map*, Chaos, Solitions and Fractals **42**, 447–452, 2009.

[20] H.A. Saleh, P. Dare, *Effective Heuristics for the GPS Survey Network of Malta: Simulated Annealing and Tabu Search Techniques*, Journal of Heuristics **7**, 533–549, 2001.

[21] H.A. Saleh, P. Dare, *Heuristic Methods for Designing a Global Positioning System Surveying Network in the Republic of Seychelles*, The Arabian Journal for Science and Engineering **26**(1B), 74–93, 2002.

[22] T. Stutzle, H.H. Hoos, *MAX-MIN Ant System*; In: "Future Generation Computer Systems", vol. 16, M. Dorigo, T. Stutzle, G. Di Caro (Eds.), 889–914, 2000.

[23] E-G. Talbi, *Metaheuristics: From Design to Implementation*, Wiley, 624 pages, 2009.

[24] P. Teunissen, A. Kleusberg, *GPS for Geodesy*, $2^{nd}$ edition, Springer, 650 pages, 1998.

[25] P-F. Verhulst, *A Note on the Law of Population Growth*, Correspondence Mathematiques et Physiques **10**, 113–121, 1938 (in French).

[26] D. Yang, G. Li, G. Cheng, *On the Efficiency of Chaos Optimization Algorithms for Global Optimization*, Chaos, Solitions and Fractals **34**, 1366–1375, 2007.

DOI: 10.15439/2015F223
ACSIS, Vol. 5

# InterCriteria Analysis of a Model Parameters Identification using Genetic Algorithm

Olympia Roeva
Institute of Biophysics and Biomedical Engineering
Bulgarian Academy of Science
Sofia, Bulgaria
E-mail: olympia@biomed.bas.bg

Stefka Fidanova
Institute of Information and Communication Technology
Bulgarian Academy of Science
Sofia, Bulgaria
E-mail: stefka@parallel.bas.bg

Peter Vassilev
Institute of Biophysics and Biomedical Engineering
Bulgarian Academy of Science
Sofia, Bulgaria
E-mail: peter.vassilev@gmail.com

Pawel Gepner
Intel Corporation, Pipers Way
Swindon Wiltshire SN3 1RJ
United Kingdom
E-mail: pawel.gepner@intel.com

***Abstract*—In this paper we apply an approach based on the apparatus of the Index Matrices and the Intuitionistic Fuzzy Sets – namely InterCriteria Analysis. The main idea is to use the InterCriteria Analysis to establish the existing relations and dependencies of defined parameters in non-linear model of an *E. coli* fed-batch cultivation process. Moreover, based on results of series of identification procedures we observe the mutual relations between model parameters and considered optimization techniques outcomes, such as execution time and objective function value. Based on InterCriteria Analysis we examine the obtained identification results and discuss the conclusions about existing relations and dependencies between defined, in terms of InterCriteria Analysis, criteria.**

*Index Terms*—InterCriteria Analysis; Index matrices; Intuitionistic Fuzzy Sets; Genetic Algorithm; chromosomes; parameter identification; *E. coli*; fed-batch cultivation process.

## I. Introduction

THE InterCriteria Analysis (ICA) is developed with the aim to gain additional insight into the nature of the criteria involved and discover on this basis existing relations between the criteria themselves [8]. It is based on the apparatus of the Index Matrices (IM) [10], [11], and the Intuitionistic Fuzzy Sets [12] and can be applied for decision making in different areas of science and practice. The approach has been discussed in a several papers [10], [14], [15], [16]. In [8] a possibility of the ICA method for criterion value prediction, proposing two algorithms, is presented. In [16] a discussion on the threshold values in the ICA was further elaborated. But, up to now, considering ICA application, the only applications reported are in one area: namely, EU member states competitiveness analysis [14], [15]. Encouraging results of these first applications of the ICA provoke us to use the method for establishing and identifying the relations between parameters of the mathematical model of an *E. coli* fed-batch cultivation process. The model parameters are further considered as criteria in terms of ICA.

In the case of modelling of cultivation processes ICA approach could be very useful. Cultivation processes are characterized with complex, non-linear dynamic and their modelling is a hard combinatorial optimization problem. On the one hand, the parameter identification is of key importance for modelling process and additional knowledge about the model parameters relations will be extremely useful to improve the model accuracy. On the other hand, the information may be used to improve the performance of the used optimization algorithms if, for instance, some algorithm outcomes are added to the considered criteria. Thus, the relations between model parameters and optimization algorithm performance will be established.

In this paper we applied the ICA to establish the basic relations between the parameters in the model of an *E. coli* fed-batch cultivation process. The existing relations are identified based on results of a series of parameters identification procedures. The use of meta-heuristic techniques such as Genetic Algorithms (GAs) has received more and more attention [3]. These methods offer good solutions, even global optima, within reasonable computing time [17], so we choose to use genetic algorithms for estimation of the model parameters.

The paper is organized as follows. The background of InterCriteria Analysis is given in Section 2. The problem formulation is described in Section 3. The numerical results and a discussion are presented in Section 4. Conclusion remarks are done in Section 5.

## II. InterCriteria Analysis

Here we expand on the idea proposed in [8]. Following [8] and [12] we will obtain an Intuitionistic Fuzzy Pair (IFP) as the degrees of “agreement” and “disagreement” between two criteria applied on different objects. We remind briefly that an IFP is an ordered pair of real non-negative numbers $\langle a, b\rangle$ such that:

$$a + b \leq 1.$$

978-83-60810-66-8/$25.00©2015, IEEE

For clarity, let us be given an IM (see [10]) whose index sets consist of the names of the criteria (for rows) and objects (for columns). The elements of this IM are further supposed to be real numbers (in the general case, this is not required). We will obtain an IM with index sets consisting of the names of the criteria (for rows and for columns) with elements IFPs corresponding to the "agreement" and "disagreement" of the respective criteria.

Two things are further supposed (which are not always guaranteed in practice and, when not fulfilled, present an interesting direction for new research in themselves):

1) All criteria provide an evaluation for all objects (i.e. there are no inapplicable criteria for a given object) and all these evaluations are available (no missing evaluations).
2) All the evaluations of a given criteria can be compared amongst themselves.

Further by $O$ we denote the set of all objects $O_1, O_2, \ldots, O_n$ being evaluated, and by $C(O)$ the set of values assigned by a given criteria $C$ to the objects, i.e.

$$\begin{aligned} O &\stackrel{\text{def}}{=} \{O_1, O_2, \ldots, O_n\}, \\ C(O) &\stackrel{\text{def}}{=} \{C(O_1), C(O_2), \ldots, C(O_n)\}. \end{aligned}$$

Let:

$$C^*(O) \stackrel{\text{def}}{=} \{\langle x, y\rangle |\ x \neq y\ \&\ \langle x, y\rangle \in C(O) \times C(O)\}.$$

In order to compare two criteria we must construct the vector of all internal comparisons of each criteria, which fulfill exactly one of three relations $R$, $\overline{R}$ and $\tilde{R}$. In other words, we require that for a fixed criterion $C$ and any ordered pair $\langle x, y\rangle \in C^*(O)$ it is true:

$$\langle x, y\rangle \in R \Leftrightarrow \langle y, x\rangle \in \overline{R}, \tag{1}$$

$$\langle x, y\rangle \in \tilde{R} \Leftrightarrow \langle x, y\rangle \notin (R \cup \overline{R}), \tag{2}$$

$$R \cup \overline{R} \cup \tilde{R} = C^*(O). \tag{3}$$

From the above it is seen that we need only consider a subset of $C(O) \times C(O)$ for the effective calculation of the vector of internal comparisons (denoted further by $V(C)$) since from (1), (2) and (3) it follows that if we know what is the relation between $x$ and $y$ we also know what is the relation between $y$ and $x$. Thus we will only consider lexicographically ordered pairs $\langle x, y\rangle$. Let, for brevity:

$$C_{i,j} = \langle C(O_i), C(O_j)\rangle.$$

Then for a fixed criterion $C$ we construct the vector:

$$\begin{aligned} V(C) = \{&C_{1,2}, C_{1,3}, \ldots, C_{1,n}, C_{2,3}, C_{2,4}, \ldots, \\ &C_{2,n}, C_{3,4}, \ldots, C_{3,n}, \ldots, C_{n-1,n}\}. \end{aligned}$$

It can be easily seen that it has exactly $\frac{n(n-1)}{2}$ elements. Further, to simplify our considerations, we replace the vector $V(C)$ with $\hat{V}(C)$, where for each $1 \leq k \leq \frac{n(n-1)}{2}$ for the $k$-th component it is true:

$$\hat{V_k}(C) = \begin{cases} 1 \text{ iff } V_k(C) \in R, \\ -1 \text{ iff } V_k(C) \in \overline{R}, \\ 0 \text{ otherwise.} \end{cases}$$

Then when comparing two criteria we determine the "degree of agreement" between the two as the number of matching components (divided by the length of the vector for normalization purposes). This can be done in several ways, e.g. by counting the matches or by taking the complement of the Hamming distance. The "degree of disagreement" is the number of components of opposing signs in the two vectors (again normalized by the length). This also may be done in various ways. A pseudocode of the algorithm used in this study for calculating the degrees of agreement and disagreement between two criteria $C$ and $C'$ is presented below.

**Algorithm 1** Calculating "agreement" and "disagreement" between two criteria

```
Require: Vectors $\hat{V}(C)$ and $\hat{V}(C')$

function DEGREE OF AGREEMENT($\hat{V}(C), \hat{V}(C')$)
    $V \leftarrow \hat{V}(C) - \hat{V}(C')$
    $\mu_{C,C'} \leftarrow 0$
    for $i \leftarrow 1$ to $\frac{n(n-1)}{2}$ do
        if $V_i = 0$ then
            $\mu_{C,C'} \leftarrow \mu_{C,C'} + 1$
        end if
    end for
    $\mu_{C,C'} \leftarrow \frac{2}{n(n-1)}\mu_{C,C'}$
    return $\mu_{C,C'}$
end function

function DEGREE OF DISAGREEMENT($\hat{V}(C), \hat{V}(C')$)
    $V \leftarrow \hat{V}(C) - \hat{V}(C')$
    $\nu_{C,C'} \leftarrow 0$
    for $i \leftarrow 1$ to $\frac{n(n-1)}{2}$ do
        if abs($V_i$) = 2 then                ▷ abs: absolute value
            $\nu_{C,C'} \leftarrow \nu_{C,C'} + 1$
        end if
    end for
    $\nu_{C,C'} \leftarrow \frac{2}{n(n-1)}\nu_{C,C'}$
    return $\nu_{C,C'}$
end function
```

It is obvious (from the way of calculation) that for $\mu_{C,C'}$, $\nu_{C,C'}$, we have:

$$\mu_{C,C'} = \mu_{C',C}, \nu_{C,C'} = \nu_{C',C}.$$

Also, $\langle \mu_{C,C'}, \nu_{C,C'}\rangle$ is an IFP.

## III. Problem Formulation

Let us use the following non-linear differential equation system to describe the *E. coli* fed-batch cultivation process [1], [4]:

$$\frac{dX}{dt} = \mu X - \frac{F_{in}}{V}X, \tag{4}$$

$$\frac{dS}{dt} = -q_S X + \frac{F_{in}}{V}(S_{in} - S), \tag{5}$$

$$\frac{dV}{dt} = F_{in}, \tag{6}$$

where

$$\mu = \mu_{max}\frac{S}{k_S + S}, \quad q_S = \frac{1}{Y_{S/X}}\mu \tag{7}$$

and $X$ is the biomass concentration, [g/l]; $S$ is the substrate concentration, [g/l]; $F_{in}$ is the feeding rate, [l/h]; $V$ is the bioreactor volume, [l]; $S_{in}$ is the substrate concentration in the feeding solution, [g/l]; $\mu$ and $q_S$ are the specific rate functions, [1/h]; $\mu_{max}$ is the maximum value of the $\mu$, [1/h]; $k_S$ is the saturation constant, [g/l]; $Y_{S/X}$ is the yield coefficient, [-].

For the model (Eq. (4)-Eq. (7)) the parameters that will be identified are $\mu_{max}, k_S$ and $Y_{S/X}$.

Let $Z_{\text{mod}} \overset{\text{def}}{=} [X_{\text{mod}}\ S_{\text{mod}}]$ (model predictions for biomass and substrate) and $Z_{\text{exp}} \overset{\text{def}}{=} [X_{\text{exp}}\ S_{\text{exp}}]$ (known experimental data for biomass and substrate). Then putting $Z = Z_{\text{mod}} - Z_{\text{exp}}$, we define the objective function as:

$$J = \|Z\|^2 \to \min, \tag{8}$$

where $\|\|$ denotes the $\ell^2$-vector norm.

For the model parameters identification we use experimental data for biomass and glucose concentration of an *E. coli* MC4110 fed-batch fermentation process. The detailed description of the process condition and experimental data are presented in [2].

To estimate the model parameters we applied consistently 14 differently tuned GA. We use various population sizes – from 5 to 200 chromosomes in the population. The number of generations is fixed to 200. The main GA operators and parameters are summarized in Table I. Because of the stochastic nature of the applied algorithms we perform series of 30 runs for each population size. Thus, we obtain the average, best and worst estimate of the parameters, as well as of the algorithm execution time and value of objective function. The detailed description of identification procedure is given in [7].

To perform ICA three IMs are constructed – the IM $A_1$ (Eq. 9) with the obtained average results, the IM $A_2$ (Eq. 10) with the best obtained results and IM $A_3$ (Eq. 11) with the worst obtained results. In addition to the presented in [7] results here the average, worst and best estimates for the tree model parameters in all 14 cases are given too. Thus, five criteria are considered – $C_1$ is parameter $\mu_{max}$, $C_2$ is parameter $k_S$, $C_3$ is parameter $Y_{S/X}$, $C_4$ is objective function value $J$ and $C_5$ is resulting execution time $T$.

TABLE I
Main GA Operators and Parameters

| **Operator** | **Type** |
|---|---|
| fitness function | linear ranking |
| selection function | roulette wheel selection |
| crossover function | simple crossover |
| mutation function | binary mutation |
| reinsertion | fitness-based |
| **Parameter** | **Value** |
| generation gap | 0.97 |
| crossover probability | 0.75 |
| mutation probability | 0.01 |
| number of generations | 200 |

## IV. Numerical Results and Discussion

Computer specification to run all identification procedures are Intel Core i5-2329 3.0 GHz, 8 GB Memory, Windows 7 (64bit) operating system.

Based on the presented **Algorithm 1** the ICA is implemented in the Matlab 7.5 environment. We obtain IMs that determine the degrees of "agreement" ($\mu_{C,C'}$) and "disagreement" ($\nu_{C,C'}$) between criteria for the three cases.

*1) Case of average results:*

Resulting degrees of "agreement" ($\mu_{C,C'}$) are as follows:

$$\text{IM}_1 = \begin{array}{c|ccccc} & C_1 & C_2 & C_3 & C_4 & C_5 \\ \hline C_1 & \mathbf{1} & 0.91 & 0.41 & 0.74 & 0.26 \\ C_2 & 0.91 & \mathbf{1} & 0.36 & 0.78 & 0.27 \\ C_3 & 0.41 & 0.36 & \mathbf{1} & 0.55 & 0.38 \\ C_4 & 0.74 & 0.78 & 0.55 & \mathbf{1} & 0.11 \\ C_5 & 0.26 & 0.27 & 0.38 & 0.11 & \mathbf{1} \end{array}$$

Resulting degrees of "disagreement" ($\nu_{C,C'}$) are as follows:

$$\text{IM}_2 = \begin{array}{c|ccccc} & C_1 & C_2 & C_3 & C_4 & C_5 \\ \hline C_1 & \mathbf{0} & 0.08 & 0.58 & 0.26 & 0.74 \\ C_2 & 0.08 & \mathbf{0} & 0.62 & 0.21 & 0.71 \\ C_3 & 0.58 & 0.62 & \mathbf{0} & 0.44 & 0.60 \\ C_4 & 0.26 & 0.21 & 0.44 & \mathbf{0} & 0.89 \\ C_5 & 0.74 & 0.71 & 0.60 & 0.89 & \mathbf{0} \end{array}$$

*2) Case of worst results:*

Resulting degrees of "agreement" ($\mu_{C,C'}$) are as follows:

$$\text{IM}_5 = \begin{array}{c|ccccc} & C_1 & C_2 & C_3 & C_4 & C_5 \\ \hline C_1 & \mathbf{1} & 0.79 & 0.34 & 0.88 & 0.14 \\ C_2 & 0.79 & \mathbf{1} & 0.18 & 0.84 & 0.22 \\ C_3 & 0.34 & 0.18 & \mathbf{1} & 0.33 & 0.64 \\ C_4 & 0.88 & 0.84 & 0.33 & \mathbf{1} & 0.07 \\ C_5 & 0.14 & 0.22 & 0.64 & 0.07 & \mathbf{1} \end{array}$$

Resulting degrees of "disagreement" ($\nu_{C,C'}$) are as follows:

$$A_1(average) =$$

$$\begin{array}{c|cccccccccccccc}
 & \mathrm{GA}_5 & \mathrm{GA}_{10} & \mathrm{GA}_{20} & \mathrm{GA}_{30} & \mathrm{GA}_{40} & \mathrm{GA}_{50} & \mathrm{GA}_{60} & \mathrm{GA}_{70} & \mathrm{GA}_{80} & \mathrm{GA}_{90} & \mathrm{GA}_{100} & \mathrm{GA}_{110} & \mathrm{GA}_{150} & \mathrm{GA}_{200} \\
\hline
C_1 & 0.552 & 0.525 & 0.515 & 0.491 & 0.486 & 0.508 & 0.497 & 0.498 & 0.498 & 0.496 & 0.494 & 0.500 & 0.492 & 0.489 \\
C_2 & 0.022 & 0.021 & 0.018 & 0.013 & 0.010 & 0.016 & 0.014 & 0.013 & 0.014 & 0.014 & 0.012 & 0.015 & 0.013 & 0.011 \\
C_3 & 2.032 & 2.025 & 2.019 & 2.023 & 2.023 & 2.020 & 2.022 & 2.022 & 2.022 & 2.021 & 2.023 & 2.021 & 2.021 & 2.023 \\
C_4 & 6.271 & 5.838 & 4.760 & 4.561 & 4.646 & 4.607 & 4.580 & 4.568 & 4.578 & 4.570 & 4.553 & 4.547 & 4.560 & 4.545 \\
C_5 & 4.649 & 6.053 & 7.472 & 11.248 & 12.917 & 14.649 & 16.973 & 19.719 & 21.793 & 24.196 & 26.848 & 29.515 & 39.406 & 51.917
\end{array} \tag{9}$$

$$A_2(best) =$$

$$\begin{array}{c|cccccccccccccc}
 & \mathrm{GA}_5 & \mathrm{GA}_{10} & \mathrm{GA}_{20} & \mathrm{GA}_{30} & \mathrm{GA}_{40} & \mathrm{GA}_{50} & \mathrm{GA}_{60} & \mathrm{GA}_{70} & \mathrm{GA}_{80} & \mathrm{GA}_{90} & \mathrm{GA}_{100} & \mathrm{GA}_{110} & \mathrm{GA}_{150} & \mathrm{GA}_{200} \\
\hline
C_1 & 0.491 & 0.480 & 0.494 & 0.491 & 0.488 & 0.492 & 0.488 & 0.490 & 0.484 & 0.491 & 0.488 & 0.486 & 0.488 & 0.488 \\
C_2 & 0.013 & 0.011 & 0.013 & 0.012 & 0.012 & 0.012 & 0.012 & 0.013 & 0.011 & 0.013 & 0.012 & 0.012 & 0.012 & 0.012 \\
C_3 & 2.023 & 2.024 & 2.018 & 2.023 & 2.020 & 2.023 & 2.020 & 2.019 & 2.019 & 2.020 & 2.019 & 2.021 & 2.019 & 2.018 \\
C_4 & 4.833 & 4.855 & 4.475 & 4.482 & 4.444 & 4.449 & 4.463 & 4.438 & 4.447 & 4.450 & 4.425 & 4.433 & 4.458 & 4.436 \\
C_5 & 4.867 & 5.912 & 7.675 & 11.295 & 13.229 & 15.007 & 17.316 & 20.062 & 22.667 & 24.757 & 26.926 & 30.015 & 39.780 & 52.323
\end{array} \tag{10}$$

$$A_3(worst) =$$

$$\begin{array}{c|cccccccccccccc}
 & \mathrm{GA}_5 & \mathrm{GA}_{10} & \mathrm{GA}_{20} & \mathrm{GA}_{30} & \mathrm{GA}_{40} & \mathrm{GA}_{50} & \mathrm{GA}_{60} & \mathrm{GA}_{70} & \mathrm{GA}_{80} & \mathrm{GA}_{90} & \mathrm{GA}_{100} & \mathrm{GA}_{110} & \mathrm{GA}_{150} & \mathrm{GA}_{200} \\
\hline
C_1 & 0.577 & 0.538 & 0.544 & 0.518 & 0.521 & 0.517 & 0.515 & 0.510 & 0.510 & 0.505 & 0.489 & 0.510 & 0.504 & 0.511 \\
C_2 & 0.015 & 0.026 & 0.024 & 0.018 & 0.019 & 0.018 & 0.018 & 0.016 & 0.017 & 0.015 & 0.012 & 0.016 & 0.015 & 0.016 \\
C_3 & 2.037 & 1.995 & 2.019 & 2.021 & 2.021 & 2.021 & 2.020 & 2.022 & 2.021 & 2.022 & 2.022 & 2.023 & 2.022 & 2.022 \\
C_4 & 9.296 & 9.618 & 5.363 & 5.009 & 4.967 & 4.864 & 4.808 & 4.736 & 4.746 & 4.721 & 4.702 & 4.732 & 4.672 & 4.721 \\
C_5 & 5.600 & 5.632 & 7.301 & 10.827 & 12.496 & 14.399 & 16.801 & 19.500 & 21.715 & 23.915 & 27.051 & 29.188 & 39.921 & 51.309
\end{array} \tag{11}$$

$$\mathrm{IM}_6 = \begin{array}{c|ccccc}
 & C_1 & C_2 & C_3 & C_4 & C_5 \\
\hline
C_1 & \mathbf{0} & 0.20 & 0.65 & 0.12 & 0.86 \\
C_2 & 0.20 & \mathbf{0} & 0.80 & 0.15 & 0.77 \\
C_3 & 0.65 & 0.80 & \mathbf{0} & 0.66 & 0.35 \\
C_4 & 0.12 & 0.15 & 0.66 & \mathbf{0} & 0.93 \\
C_5 & 0.86 & 0.77 & 0.35 & 0.93 & \mathbf{0}
\end{array}$$

*3) Case of best results:*

Resulting degrees of "agreement" ($\mu_{C,C'}$) are as follows:

$$\mathrm{IM}_3 = \begin{array}{c|ccccc}
 & C_1 & C_2 & C_3 & C_4 & C_5 \\
\hline
C_1 & \mathbf{1} & 0.74 & 0.49 & 0.63 & 0.36 \\
C_2 & 0.74 & \mathbf{1} & 0.35 & 0.53 & 0.51 \\
C_3 & 0.49 & 0.35 & \mathbf{1} & 0.71 & 0.30 \\
C_4 & 0.63 & 0.53 & 0.71 & \mathbf{1} & 0.25 \\
C_5 & 0.36 & 0.51 & 0.30 & 0.25 & \mathbf{1}
\end{array}$$

Resulting degrees of "disagreement" ($\nu_{C,C'}$) are as follows:

$$\mathrm{IM}_4 = \begin{array}{c|ccccc}
 & C_1 & C_2 & C_3 & C_4 & C_5 \\
\hline
C_1 & \mathbf{0} & 0.19 & 0.44 & 0.33 & 0.59 \\
C_2 & 0.19 & \mathbf{0} & 0.59 & 0.44 & 0.46 \\
C_3 & 0.44 & 0.59 & \mathbf{0} & 0.26 & 0.68 \\
C_4 & 0.33 & 0.44 & 0.26 & \mathbf{0} & 0.75 \\
C_5 & 0.59 & 0.46 & 0.68 & 0.75 & \mathbf{0}
\end{array}$$

Let us consider the following scheme for defining the consonance and dissonance between each pair of criteria (see Table III).

TABLE II
CRITERIA RELATIONS SORTED BY $\mu_{C,C'}$ VALUES

| **Criteria relation** | **Obtained $\langle \mu_{C,C'}, \nu_{C,C'} \rangle$ values in case of** | | |
|---|---|---|---|
| | **average results** | **worst results** | **best results** |
| $C_1 \leftrightarrow C_2$ | $\langle 0.91, 0.08 \rangle$ | $\langle 0.79, 0.20 \rangle$ | $\langle 0.74, 0.19 \rangle$ |
| $C_2 \leftrightarrow C_4$ | $\langle 0.78, 0.21 \rangle$ | $\langle 0.84, 0.15 \rangle$ | $\langle 0.53, 0.44 \rangle$ |
| $C_1 \leftrightarrow C_4$ | $\langle 0.74, 0.26 \rangle$ | $\langle 0.88, 0.12 \rangle$ | $\langle 0.63, 0.33 \rangle$ |
| $C_3 \leftrightarrow C_4$ | $\langle 0.55, 0.44 \rangle$ | $\langle 0.33, 0.66 \rangle$ | $\langle 0.71, 0.26 \rangle$ |
| $C_1 \leftrightarrow C_3$ | $\langle 0.41, 0.58 \rangle$ | $\langle 0.34, 0.65 \rangle$ | $\langle 0.49, 0.44 \rangle$ |
| $C_3 \leftrightarrow C_5$ | $\langle 0.38, 0.60 \rangle$ | $\langle 0.64, 0.35 \rangle$ | $\langle 0.30, 0.68 \rangle$ |
| $C_2 \leftrightarrow C_3$ | $\langle 0.36, 0.62 \rangle$ | $\langle 0.18, 0.80 \rangle$ | $\langle 0.35, 0.59 \rangle$ |
| $C_2 \leftrightarrow C_5$ | $\langle 0.27, 0.71 \rangle$ | $\langle 0.22, 0.77 \rangle$ | $\langle 0.51, 0.46 \rangle$ |
| $C_1 \leftrightarrow C_5$ | $\langle 0.26, 0.74 \rangle$ | $\langle 0.14, 0.86 \rangle$ | $\langle 0.36, 0.59 \rangle$ |
| $C_4 \leftrightarrow C_5$ | $\langle 0.11, 0.89 \rangle$ | $\langle 0.07, 0.93 \rangle$ | $\langle 0.25, 0.75 \rangle$ |

In the case of the average values of the examined criteria, in accordance with the scale presented in Table III, we found the following pair dependencies:

- There is no observed strong positive consonance or strong negative consonance between any of the ten criteria pairs. Since the observed values depend on the number of objects if we can expand their number, it is possible to

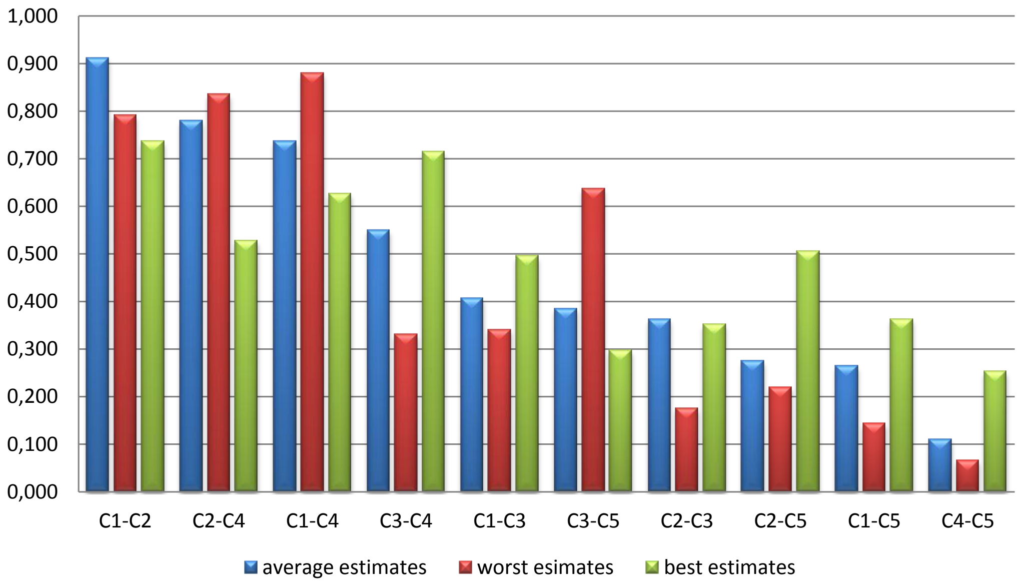


Fig. 1. Degrees of "agreement" ($\mu_{C,C'}$ values) for all cases

TABLE III
CONSONANCE AND DISSONANCE SCALE

| **Interval of** $\mu_{C,C'}$, % | **Meaning** |
|---|---|
| [0-5] | strong negative consonance |
| (5-15] | negative consonance |
| (15-25] | weak negative consonance |
| (25-33] | weak dissonance |
| (33-43] | dissonance |
| (43-57] | strong dissonance |
| (57-67] | dissonance |
| (67-75] | weak dissonance |
| (75-85] | weak positive consonance |
| (85-95] | positive consonance |
| (95-100] | strong positive consonance |

obtain values in these intervals.

- For the pair $C_4 \leftrightarrow C_5$ (i.e., $T \leftrightarrow J$) a negative consonance is identified. Such dependence is logical – for a large number of algorithm iterations (i.e., greater execution time $T$) it is more likely to find a more accurate solution, i.e. smaller value of $J$.
- The pairs $C_2 \leftrightarrow C_5$ (i.e., $k_S \leftrightarrow T$) show identical results – these criteria are in weak dissonance. The third model parameter $Y_{S/X}$ and $T$ are in dissonance. The conclusion is that the total execution time is not dependent solely on one of the model parameters. Logically the triple of these parameters should be in consonance with $T$.
- For the pairs $C_1 \leftrightarrow C_3$ (i.e., $\mu_{max} \leftrightarrow Y_{S/X}$) and $C_2 \leftrightarrow C_3$ (i.e., $k_S \leftrightarrow Y_{S/X}$) a dissonance is observed. Considering the physical meaning of the model parameters [1] it is clear that there is no dependence between these criteria. A strong correlation is expected between criteria $C_1 \leftrightarrow C_2$ (i.e., $\mu_{max} \leftrightarrow k_S$) [1]. The results confirmed these expectation – these criteria are in a positive consonance.
- The observed low value of $\mu_{C_3,C_4}$, i.e., strong dissonance between $Y_{S/X} \leftrightarrow J$ show the low sensitivity of this model parameter. According to [18] the parameter $Y_{S/X}$ has lower sensitivity compared to parameter $\mu_{max}$.
- Due to the established strong correlation between criteria $C_1 \leftrightarrow C_2$ (i.e., $\mu_{max} \leftrightarrow k_S$) we observe that $C_1 \leftrightarrow C_4$ (i.e., $\mu_{max} \leftrightarrow J$) and $C_2 \leftrightarrow C_4$ (i.e., $k_S \leftrightarrow J$) are in, respectively weak dissonance and weak positive consonance. Similarly to the relations with $T$ the conclusion is that the accuracy of the criterion is not dependent solely on one of the model parameters. Logically the triple of these parameters should be in consonance (or strong consonance) with the criterion value. Moreover, taking into account the parameters sensitivity it is clear that the more sensitive parameter will be more linked to the value of $J$.

Due to stochastic nature of considered here GA we observed some different criteria dependences in the rest two cases – case of worst and case of best results:

- In the case of the worst results we found weaker relation between $C_1 \leftrightarrow C_2, C_3 \leftrightarrow C_4, C_2 \leftrightarrow C_3, C_1 \leftrightarrow C_5, C_2 \leftrightarrow C_5$ and $C_4 \leftrightarrow C_5$. For the pairs $C_1 \leftrightarrow C_4, C_2 \leftrightarrow C_4$ and $C_3 \leftrightarrow C_5$ we observed higher value of $\mu_{C,C'}$. Compared to the case of average results there are no large, strongly manifested discrepancies. In case

of discrepancy, the considered criteria pair appears in an adjacent scale according to Table III. For example, pair $C_1 \leftrightarrow C_2$ in case of average results are in positive consonance, while in case of worst results – in weak positive consonance.
- In the case of the best results we identify the same results – in case of discrepancy the considered criteria pair appears in an adjacent scale. However, in this case we observed some larger discrepancies. Taking into account the nature of the GA we consider that the results in case of average criteria values have the highest significance.

## V. Conclusion

In this paper based on the apparatus of the Index Matrices and the Intuitionistic Fuzzy Sets, InterCriteria Analysis of a model parameters identification using Genetic Algorithm is performed. A non-linear model of an *E. coli* fed-batch cultivation process is considered. Series of model identification procedures using Genetic Algorithms are done. The InterCriteria Analysis is applied to explore the existing relations and dependencies of defined model parameters and Genetic Algorithms outcomes – execution time and objective function value. Three case studies are examined – considering average, worst and best results for the obtained model parameters, execution time and objective function value. Applying the InterCriteria Analysis we establish relations and dependencies between the defined criteria. Based on the used scale for defining the consonance and dissonance between each pair of criteria, we discuss which criteria are in consonance and dissonance, as well as the degree of their dependence.

## Acknowledgment

Work presented here is partially supported by the Bulgarian National Scientific Fund under Grant DFNI-I02/5 "InterCriteria Analysis – A New Approach to Decision Making" and by the Polish-Bulgarian collaborative Grant "Parallel and Distributed Computing Practices".

## References

[1] G. Bastin and D. Dochain, *On-line Estimation and Adaptive Control of Bioreactors*, Els. Sc. Publ., 1991.

[2] O. Roeva, T. Pencheva, B. Hitzmann and St. Tzonkov, "A Genetic Algorithms Based Approach for Identification of *Escherichia coli* Fed-batch Fermentation", *Int. J. Bioautomation*, Vol. 1, 2004, pp. 30–41.

[3] O. Roeva (Ed.), *Real-World Application of Genetic Algorithms*, InTech, 2012.

[4] O. Roeva, "Improvement of Genetic Algorithm Performance for Identification of Cultivation Process Models", *Advanced Topics on Evolutionary Computing, Book Series: Artificial Intelligence Series – WSEAS*, 2008, pp. 34–39.

[5] M. Arndt and B. Hitzmann, "Feed Forward/feedback Control of Glucose Concentration during Cultivation of *Escherichia coli*", *8th IFAC Int. Conf. on Comp. Appl. in Biotechn*, Canada, 2001, pp. 425–429.

[6] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley Longman, London, 2006.

[7] O. Roeva, S. Fidanova and M. Paprzycki, "Influence of the Population Size on the Genetic Algorithm Performance in Case of Cultivation Process Modelling", *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS), WCO 2013*, Poland, pp. 371–376.

[8] K. Atanassov, D. Mavrov and V. Atanassova, "Intercriteria Decision Making: A New Approach for Multicriteria Decision Making, Based on Index Matrices and Intuitionistic Fuzzy Sets", *Issues in IFSs and GNs*, Vol. 11, 2014, pp. 1–8

[9] K. Atanassov, "Generalized index matrices", *Comptes Rendus de l'Academie Bulgare des Sciences*, Vol. 40, 1987, No. 11, pp. 15–18.

[10] K. Atanassov," On index matrices, Part 1: Standard cases", *Advanced Studies in Contemporary Mathematics*, Vol. 20, 2010, No. 2, pp. 291–302.

[11] K. Atanassov, "On index matrices, Part 2: Intuitionistic fuzzy case", *Proceedings of the Jangjeon Mathematical Society*, Vol. 13, 2010, No. 2, pp. 121–126.

[12] K. Atanassov, *On Intuitionistic Fuzzy Sets Theory*, Springer, Berlin, 2012.

[13] K. Atanassov, E. Szmidt and J. Kacprzyk, "On intuitionistic fuzzy pairs", *Notes on Intuitionistic Fuzzy Sets*, Vol. 19, 2013, No. 3, pp. 1–13.

[14] V. Atanassova, L. Doukovska, K. Atanassov and D. Mavrov, "InterCriteria Decision Making Approach to EU Member States Competitiveness Analysis", *Proc. of the International Symposium on Business Modeling and Software Design – BMSD'14*, 24-26 June 2014, Luxembourg, Grand Duchy of Luxembourg, 2014, pp. 289–294.

[15] V. Atanassova, L. Doukovska, D. Mavrov and K. Atanassov, "InterCriteria Decision Making Approach to EU Member States Competitiveness Analysis: Temporal and Threshold Analysis", P. Angelov et al. (eds.), Intelligent Systems'2014, *Advances in Intelligent Systems and Computing*, 322, pp. 95–106.

[16] V. Atanassova, D. Mavrov, L. Doukovska and K. Atanassov, "Discussion on the threshold values in the InterCriteria Decision Making approach", *Int. J. Notes on Intuitionistic Fuzzy Sets*, Volume 20, 2014, Number 2, pp. 94–99.

[17] I. Boussaid, J. Lepagnot and P. Siarry, "A Survey on Optimization Metaheuristics", *Information Sciences*, Vol. 237, 2013, pp. 82–117.

[18] O. Roeva, "Sensitivity Analysis of *E. coli* Fed-batch Cultivation Local Models", *Mathematica Balkanica, New Series*, Vol. 25, 2011, Fasc. 4, pp. 395–411.

DOI: 10.15439/2015F329
ACSIS, Vol. 5

# Maximum Exploratory Equivalence in Trees

Luka Fürst, Uroš Čibej, and Jurij Mihelič
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, SI-1000 Ljubljana, Slovenia
Email: {luka.fuerst,uros.cibej,jurij.mihelic}@fri.uni-lj.si

***Abstract*—Many practical problems are modeled with networks and graphs. Their exploration is of significant importance, and several graph-exploration algorithms already exist. In this paper, we focus on a type of vertex equivalence, called *exploratory equivalence*, which has a great potential to speed up such algorithms. It is an equivalence based on graph automorphisms and can, for example, help us in solving the subgraph isomorphism problem, which is a well-known $NP$-hard problem. In particular, if a given pattern graph has nontrivial automorphisms, then each of its nontrivial exploratory equivalent classes gives rise to a set of constraints to prune the search space of solutions. In the paper, we define the maximum exploratory equivalence problem. We show that the defined problem is at least as hard the graph isomorphism problem. Additionally, we present a polynomial-time algorithm for solving the problem when the input is restricted to tree graphs. Furthermore, we show that for trees, a maximum exploratory equivalent partition leads to a globally optimal set of subgraph isomorphism constraints, whereas this is not necessarily the case for general graphs.**

## I. Introduction

Searching for patterns in structured data is one of the most ubiquitous applications of computer algorithms in various scientific areas. Such data is often modeled with graphs, which efficiently represent diverse types of entities (modeled by graph vertices) and relations between them (modeled by graph edges) and also enable a more general and global view on the data. In the era of ever-growing, even planetary-wide (social, citation, traffic, etc.) networks [1], all of which can be naturally modeled by graphs, graph representation and also algorithms on graphs are becoming increasingly important. Applications of graphs arise in various areas, ranging from chemistry [2], [3], economy [4], politics [5], to popular culture [6].

In this paper, we focus on a general technique for speeding up algorithms that search for patterns in graphs. The main idea of our technique is to exploit symmetries of a graph, i.e., to find equivalent vertices in such a way that if two vertices are equivalent then the search algorithm could process only one (and deduce the information about the other one).

The problem of finding equivalent vertices of a graph has already appeared in the literature; see, for example, papers on regular and structural equivalence [7], [8]. To the best of our knowledge, our definition introduces a new form of equivalence. We call it *exploratory equivalence* ($EE$), since its primary intent is to be utilized in graph search algorithms; see [9] for the introductory paper. Nevertheless, the exploitation of symmetries of a problem to reduce the amount of time for exploring the solution search space is not new. See, for example, a method for solving 0/1 integer linear programs having a large symmetry [10] or [11] for a similar method. Another equivalence similar to ours, defined in [12], can also be used for pruning the search space. However, our equivalence is more general and, hence, has a greater pruning power.

To find the symmetries in a graph, the usual approach is to find all graph isomorphisms ($GI$), i.e., structure preserving mappings. In particular, given two graphs, the graph isomorphism problem asks whether they are the same. Similarly, the graph automorphism problem asks whether a graph can be (non-trivially) mapped to itself. The $GI$ problem has a special place in the complexity theory, as it is a canonical example of a possible candidate for an $NP$-intermediate problem. Ladner's theorem [13] tells us that if $P$ is not equal to $NP$, then the class of $NP$-intermediate problems is not empty. As a result, a polynomial-time algorithm is unlikely for the $GI$ problem. Nevertheless, in practice, there are several efficient algorithms and software packages for finding automorphisms of graphs, e.g., NAUTY [14], [15], BLISS [16], [17], SAUCY [18], [19], [20], and NISHE [21]. For some special cases of graphs, e.g., tree graphs [22], polynomial-time algorithms exist.

Based on automorphisms, one could define *automorphic equivalence*, where two vertices are equivalent if and only if there exists an automorphism that maps one to another. Such equivalence classes are also called *orbits*. Notice that exploratory equivalence is similar but not the same as *automorphic equivalence*. Indeed, exploratory equivalence is more restrictive.

In our preceding paper [9], we already presented the definition of exploratory equivalences and the corresponding problem of finding maximum exploratory equivalent partition of graph vertices, i.e., the MAXEXPLOREQ problem. In this paper, we show that the MAXEXPLOREQ is $GI$-hard, which means that a polynomial-time algorithm (in terms of the number of graph vertices) is unlikely to exist. Hence, it is reasonable to restrict the input to MAXEXPLOREQ to selected subclasses of graphs. As the second contribution of this paper, we present a polynomial-time algorithm for solving MAXEXPLOREQ on an arbitrary tree. Thereby we show that the restriction of MAXEXPLOREQ to trees is in $P$. Additionally, we also show that for trees, a maximum exploratory equivalent partition leads to a globally optimal set of subgraph isomorphism search constraints. In particular, when searching for a given tree in a given host graph using the constraints derived from a maximum exploratory equivalent partition of the tree, each of the occurrences of the tree in the host graph will be discovered exactly once. For general graphs, this is not necessarily the case.

The rest of this paper is structured as follows. In the next section, we present mathematical notions needed for the rest of the paper. In Section III, we present a motivational example (based on the subgraph isomorphism problem) for exploratory

978-83-60810-66-8/$25.00©2015, IEEE

equivalence. The definition of the MaxExploreEq problem is presented in Section IV. In Section V, we show that the MaxExploreEq problem is $GI$-hard. In Section VI, we present a polynomial-time algorithm for solving the MaxExploreEq problem on trees and determine its computational complexity. Section VII presents an empirical demonstration of the presented algorithm on the set of all small trees. In Section VIII, we elaborate on the connection between exploratory equivalence and subgraph isomorphism, with a particular emphasis on trees. Finally, Section IX concludes the paper.

## II. Preliminaries

Let $G = (V, E)$ denote a simple undirected graph, where $V = \{1, 2, \ldots, n\}$ is a set of vertices and $E \subseteq V \times V$ is a set of edges. The graph can be labeled; let $\Sigma$ denote a set of labels, and let $\ell_V \colon V \to \Sigma$ and $\ell_E \colon E \to \Sigma$ denote the functions that assign labels to individual vertices and edges, respectively. An unlabeled graph can be viewed as a labeled graph where all vertices and edges have the same label. A *tree* is a connected acyclic undirected graph.

A (graph) *homomorphism* from a graph $G = (V, E)$ to a graph $H = (U, F)$ is a mapping $h \colon V \to U$ such that for each $(i, j) \in E$ it also holds that $(h(i), h(j)) \in F$. To simplify notation, the homomorphism $h \colon V \to U$ will be denoted $h \colon G \to H$. An *endomorphism* is a homomorphism whose domain is equal to its codomain, i.e., $h \colon G \to G$.

An *isomorphism* is a bijective homomorphism, i.e., a mapping $h \colon G \to H$ such that $(i, j) \in E$ if and only if $(h(i), h(j)) \in F$. We write $G \simeq H$ if there exists an isomorphism from $G$ to $H$; such graphs $G$ and $H$ are called isomorphic. A *subgraph isomorphism* $G \to H$ is an isomorphism between the graph $G$ and a subgraph of the graph $H$. A subgraph in $H$ that is isomorphic to $G$ is called an *occurrence* of $G$ in $H$.

An *automorphism* is both an endomorphism and an isomorphism, i.e., a mapping $h \colon G \to G$ such that $(i, j) \in E$ if and only if $(h(i), h(j)) \in E$. Note that every automorphism is a permutation. The set of all automorphisms of a graph $G$ can be defined as

$$\mathrm{Aut}(G) = \{a \in \Pi[n] \mid G \simeq a(G)\} \tag{1}$$

where $\Pi[P]$ denotes the set of all permutations of a set $P$ and $\Pi[n] \equiv \Pi[\{1, 2, \ldots, n\}]$. For example, the set of automorphisms of the graph $G$ in Fig. 1 can be denoted as $\{123456, 123465, 124356, 124365, 215634, 215643, 216534, 216543\}$, where, e.g., 215643 denotes an automorphism $h$ such that $h(1) = 2$, $h(2) = 1$, $h(3) = 5$, $h(4) = 6$, $h(5) = 4$, and $h(6) = 3$.

Given a (finite) set $S$, a family $\{P_1, P_2, \ldots, P_s\}$ of nonempty subsets of $S$ is a *partition* of $S$ if every element in $S$ is exactly in one of the subsets, i.e., $P_i \subseteq S$ and $P_i \neq \emptyset$, where $1 \leq i \leq s$, $\bigcup_{1 \leq i < s} P_i = S$, and $P_i \cap P_j = \emptyset$ for all $1 \leq i, j \leq s$ with $i \neq j$. When the partition $\{P_1, P_2, \ldots, P_s\}$ is given explicitly, we usually use $\{i \in P_1 \mid i \in P_2 \mid \ldots \mid i \in P_s\}$ as a short form, e.g., $\{\{1, 2\}, \{3\}, \{4\}\}$ is shortened to $\{1, 2 \mid 3 \mid 4\}$. In what follows, the order of the sets in a partition is often important. To denote such an *ordered partition*, we use the form $\langle i \in P_1 \mid i \in P_2 \mid \ldots \mid i \in P_s \rangle$, e.g., $\langle 1, 2 \mid 3 \mid 4 \rangle$.

## III. Motivation

Given a pattern graph and a host graph, the goal of the *subgraph isomorphism problem* is to find all (or at least one, depending on the definition) occurrences of the pattern graph in the host graph, i.e., the subgraphs of the host graph that are isomorphic to the pattern graph.

Unfortunately, the decision version of the subgraph isomorphism problem is $NP$-complete [23], while its counting version is $\#P$-complete, since the counting version of the clique problem is $\#P$-complete [24]. Furthermore, not only that it is unlikely that a polynomial-time algorithm exists, but so far no exponential-time algorithm with a lower bound better than what can be achieved by the naive enumeration of the occurrences has been devised [25]. Most algorithms are therefore based on a backtracking approach (e.g., [26], [27]). In particular, the vertices of the pattern graph are matched with those of the host graph until a match is found, using the vertex neighborhood information to prune the search space.

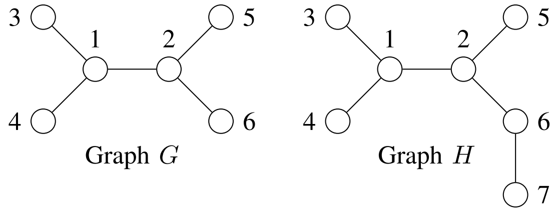


Fig. 1. A sample pattern graph $G$ and host graph $H$.

Let us assume that a given pattern graph $G$ has $m$ nontrivial automorphisms. When searching for the occurrences of $G$ in a given host graph $H$, a search algorithm that establishes all valid matches between $G$ and subgraphs of $H$ discovers *each* of $G$'s occurrences $m$ times, because the vertices of $G$ can be isomorphically mapped to the vertices of each of $G$'s occurrences in $m$ different ways. As an example, consider the pattern graph $G$ and the host graph $H$ in Fig. 1. An algorithm that is unaware of the eight automorphisms of $G$ will find the single occurrence of $G$ in $H$ eight times. In other words, it will establish eight subgraph isomorphisms $h \colon G \to H$:

| $i$ | $h_i(1)$ | $h_i(2)$ | $h_i(3)$ | $h_i(4)$ | $h_i(5)$ | $h_i(6)$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2 | 1 | 2 | 3 | 4 | 6 | 5 |
| 3 | 1 | 2 | 4 | 3 | 5 | 6 |
| 4 | 1 | 2 | 4 | 3 | 6 | 5 |
| 5 | 2 | 1 | 5 | 6 | 3 | 4 |
| 6 | 2 | 1 | 5 | 6 | 4 | 3 |
| 7 | 2 | 1 | 6 | 5 | 3 | 4 |
| 8 | 2 | 1 | 6 | 5 | 4 | 3 |

However, by imposing the constraints $h(1) < h(2)$, $h(3) < h(4)$, and $h(5) < h(6)$ while performing the exhaustive subgraph isomorphism search, the sole occurrence of the graph $G$ in the graph $H$ will be discovered exactly once, and this *regardless of the numbering of H's vertices*.

Motivated by this observation, we recently introduced the so-called *exploratory equivalence* [9], on the basis of which such constraints can be defined and safely imposed during the subgraph isomorphism search. Exploratory equivalence is an automorphisms-based equivalence relation on the vertices of a given graph; an *exploratory equivalent partition* (EE partition) is a partition of the graph vertex set into a set of exploratory equivalence classes. Every EE partition can be directly translated into a safe set of search constraints, where 'safe' means that the constraints never lead the algorithm to miss any occurrence, regardless of the numbering of the host graph vertices. In particular, an EE partition $\{u_{11}, \ldots, u_{1k_1} \mid u_{21}, \ldots, u_{2k_2} \mid \ldots \mid u_{s1}, \ldots, u_{sk_s}\}$ (the vertices $u_{11}$, …, $u_{1k_1}$ constitute the first equivalence class, etc.) gives rise to the constraint set $\{h(u_{11}) < \ldots < h(u_{1k_1})$, $h(u_{21}) < \ldots < h(u_{2k_2})$, …, $h(u_{s1}) < \ldots < h(u_{sk_s})\}$. In the example of Fig. 1, one of the EE partitions of the graph $G$ is $\{1, 2 \mid 3, 4 \mid 5, 6\}$, and it leads to the above-mentioned set of constraints.

If a graph $G$ has nontrivial automorphisms, it has several nontrivial EE partitions. All of them lead to a safe set of search constraints. However, of particular interest is one that gives rise to the set of constraints that results in the largest speedup when searching for the occurrences of $G$. Such an EE partition is called a *maximum EE partition* ('a' instead of 'the' because there can be several of them), and the problem of finding such a partition for a given graph is denoted MAXEXPLOREQ. In our previous paper [9], we defined the problem and showed two algorithms, both of which are polynomial only in the number of automorphisms, rather than in the number of graph vertices. Besides that, the algorithms fail to find a maximum EE partition for all graphs, although counterexamples appear to be very rare; for example, the second algorithm finds a maximum EE partition for all but 2 graphs out of 261080 connected unlabeled undirected 9-vertex graphs.

## IV. PROBLEM DESCRIPTION

Since the MAXEXPLOREQ problem is defined and thoroughly explained in our FedCSIS 2014 paper [9], we provide a relatively brief review of the main definitions.

*Definition 1 (cover):* A set of permutations $A \subseteq \Pi[n]$ *covers* a set $P \subseteq \{1, \ldots, n\}$ if for every permutation $\sigma$ of the set $P$ there exists a permutation $a \in A$ such that $a(i) = \sigma(i)$ for all $i \in P$:

$$\text{cover}(A, P) \equiv \forall \sigma \in \Pi[P]\ \exists a \in A\ \forall i \in P\colon \sigma(i) = a(i). \quad (2)$$

For example, the set Aut($G$) for the graph $G$ of Fig. 1 covers the set $\{3, 5\}$, since it contains both an automorphism for which $a(3) = 3$ and $a(5) = 5$ (12**3**4**5**6) and an automorphism for which $a(3) = 5$ and $a(5) = 3$ (21**5**6**3**4). For the graph $G'$ of Fig. 2, the set Aut($G'$) covers the set $\{1, 3, 5\}$, since it contains an automorphism for each of the 3! permutations of the set $\{1, 3, 5\}$ (**1**2**3**4**5**6 for the permutation 135, **1**6**5**4**3**2 for the permutation 153, **3**2**1**6**5**4 for the permutation 315, etc.)

*Definition 2 (stabilizer):* The *stabilizer* of a set $A \subseteq \Pi[n]$ with respect to a set $P \subseteq \{1, \ldots, n\}$ is the set of all permutations in $A$ that fix all elements of $P$:

$$\text{Stab}(A, P) = \{a \in A \mid \forall i \in P\colon a(i) = i\}. \quad (3)$$

For the graph $G$ of Fig. 1, we have Stab(Aut($G$), $\{1, 2\}$) $=$ {**12**3456, **12**3465, **12**4356, **12**4365}, Stab(Aut($G$), $\{3, 4\}$) $=$ {12**34**56, 12**34**65}, and Stab(Aut($G$), $\{3, 5\}$) $=$ {12**3**4**5**6}.

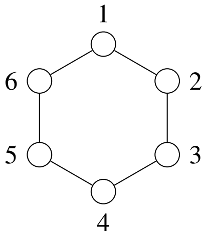


Fig. 2. A sample graph $G'$.

*Definition 3 (EE ordered partition):* For a given graph $G = (V, E)$, an ordered partition $\langle P_1, P_2, \ldots, P_s\rangle$ of $V$ is *exploratory equivalent* if for all $i \in \{1, \ldots, s\}$ we have

$$\text{cover}(A_{i-1}, P_i) \text{ and } A_i = \text{Stab}(A_{i-1}, P_i),$$

where $A_0 = \text{Aut}(G)$.

For the graph $G$ of Fig. 1, one of the EE ordered partitions is $\langle 1, 2 \mid 3, 4 \mid 5, 6\rangle$; the corresponding stabilizers are $A_1 = \{123456, 123465, 124356, 124365\}$, $A_2 = \{123456, 123465\}$, and $A_3 = \{123456\}$.

*Definition 4 (EE partition):* For a given graph $G = (V, E)$, a partition $\langle P_1, P_2, \ldots, P_s\rangle$ of $V$ is *exploratory equivalent* if there exists an exploratory equivalent ordered partition $\langle P_{i_1}, P_{i_2}, \ldots, P_{i_s}\rangle$ for a set of distinct indices $i_j \in \{1, \ldots, s\}$.

Figure 3 shows all EE partitions of the graph $G$ in Fig. 1.

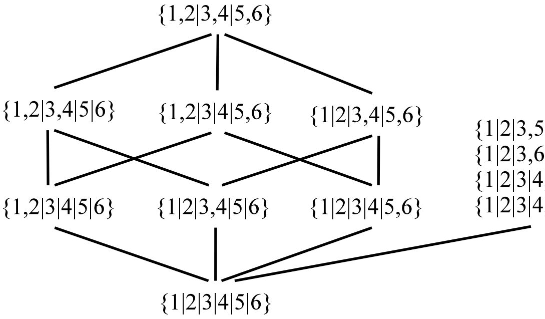


Fig. 3. The Hasse diagram of all EE partitions of the graph $G$ of Fig. 1. (The four partitions on the right-hand side are actually four separate vertices in the diagram.)

As we mentioned in the introduction, an EE partition determines a set of subgraph isomorphism search constraints. For example, the EE partition $\{1 \mid 2 \mid 3, 4 \mid 5, 6\}$ determines the constraints $h(3) < h(4)$ and $h(5) < h(6)$, which can be safely used when searching for subgraph isomorphisms $h\colon G \to H$ in an arbitrary host graph $H$. Since an EE partition set with $k$ vertices represents $k!$ permutations of those vertices, the corresponding constraint reduces the number of discoveries of each occurrence of $G$ in $H$ by a factor of $k!$. The score of an EE partition can thus be defined as follows:

*Definition 5 (score of an EE partition):* The *score* of an EE partition $\{P_1, \ldots, P_s\}$ is $\prod_{i=1}^{s} |P_i|!$.

The goal of the MAXEXPLOREQ problem is to find a maximum EE partition, i.e., one with the highest score:

*Definition 6 (*MaxExploreq*):* Given a graph $G$, find an EE partition with the maximum score.

The sole maximum EE partition of the graph $G$ of Fig. 1 is $\{1, 2 \mid 3, 4 \mid 5, 6\}$, with the score of $2!\,2!\,2! = 8$. One of the two maximum EE partitions of the graph $G'$ of Fig. 2 $\{1, 3, 5 \mid 2 \mid 4 \mid 6\}$, with the score of $3!\,1!\,1!\,1! = 6$. The other is $\{2, 4, 6 \mid 1 \mid 3 \mid 5\}$. Note that $\{1, 3, 5 \mid 2, 4, 6\}$ is *not* an EE partition, since $\mathrm{Stab}(\mathrm{Aut}(G'), \{1, 3, 5\}) = \{123456\}$ and the resulting set of automorphisms covers only singletons.

For convenience, let us also define an exploratory equivalent set and exploratory equivalent vertices:

*Definition 7 (EE set):* For a graph $G = (V, E)$, a set $P \subseteq V$ is *exploratory equivalent* if there exists an EE partition that contains $P$.

For example, in the graph $G$ of Fig. 1, the sets $\{3, 5\}$, $\{5, 6\}$, and $\{3, 6\}$ are all exploratory equivalent. However, the set $\{3, 5, 6\}$ is not.

*Definition 8 (EE vertices):* Vertices $v_1$, …, $v_k$ are *exploratory equivalent* if the set $\{v_1, \ldots, v_k\}$ is exploratory equivalent.

We will now present an alternative interpretation of exploratory equivalence that will be used in some proofs. Let $V' = \{v_1, v_2, \ldots, v_k\} \subseteq V$ be a subset of vertices of an unlabeled graph $G = (V, E)$, and let $Z_1$, $Z_2$, …, $Z_k$ be mutually distinct labels. Let $G^j$ denote a copy of the graph $G$ in which the vertex $v_i$ (for $i \in \{1, \ldots, k\}$) is labeled $\sigma_j(Z_i)$, where $\sigma_j$ (for $j \in \{1, \ldots, k!\}$) represents the $j$-th permutation of the set $\{Z_1, \ldots, Z_k\}$. Now, the set $V'$ is exploratory equivalent if the graphs $G^1$, …, $G^{k!}$ are all mutually isomorphic. For instance, in the case of the graph $G'$ of Fig. 2, the set $\{1, 3, 5\}$ is exploratory equivalent because all 3! graphs in Fig. 4 are mutually isomorphic.

Let $\mathcal{P} = \langle P_1, P_2, \ldots, P_s\rangle$ with $P_i = \{v_{i1}, v_{i2}, \ldots, v_{ik_i}\}$ be an ordered partition of the vertex set of an unlabeled graph $G$, and let $Z_{11}$, $Z_{12}$, …, $Z_{1k_1}$, $Z_{21}$, $Z_{22}$, …, $Z_{2k_2}$, …, $Z_{s1}$, $Z_{s2}$, …, $Z_{sk_s}$ be mutually distinct labels that do not occur at any vertex in the graph $G$. Let $G^{r,j}$ $(1 \leq r \leq s)$ denote a copy of the graph $G$ in which the vertices $v_{i1}$, $v_{i2}$, …, $v_{ik_i}$ (for $i \in \{1, \ldots, r-1\}$) are labeled $Z_{i1}$, $Z_{i2}$, …, $Z_{ik_i}$, respectively, while the vertices $v_{r1}$, $v_{r2}$, …, $v_{rk_r}$ are labeled $\sigma_j(Z_{r1})$, $\sigma_j(Z_{r2})$, …, $\sigma_j(Z_{rk_r})$, respectively, where $\sigma_j$ (for $j \in \{1, \ldots, k_r!\}$) represents the $j$-th permutation of the set $\{Z_{r1}, \ldots, Z_{rk_r}\}$. Now, the partition $\mathcal{P}$ is exploratory equivalent if we have $G^{i,1} \simeq G^{i,2} \simeq \ldots \simeq G^{i,k_i!}$ for each $i \in \{1, \ldots, s\}$. For instance, in the case of the graph $G$ of Fig. 1, the ordered partition $\langle 1, 2 \mid 3, 4 \mid 5, 6\rangle$ is exploratory equivalent because we have $G^{1,1} \simeq G^{1,2}$, $G^{2,1} \simeq G^{2,2}$, and $G^{3,1} \simeq G^{3,2}$ for the graphs of Fig. 5.

Alternatively, the partition $\mathcal{P} = \langle P_1, P_2, \ldots, P_s\rangle$ is exploratory equivalent if the set $P_1$ is exploratory equivalent and if for each $i \in \{2, \ldots, s\}$, the set $P_i$ remains exploratory equivalent after the labels of the vertices of the sets $P_j$ (for all $1 \leq j < i$) have been fixed to $Z_{j1}$, …, $Z_{jk_j}$.

## V. The Complexity of MaxExploreq

In this section, we show that MaxExploreq is at least as hard as the graph isomorphism problem. We have the following theorem:

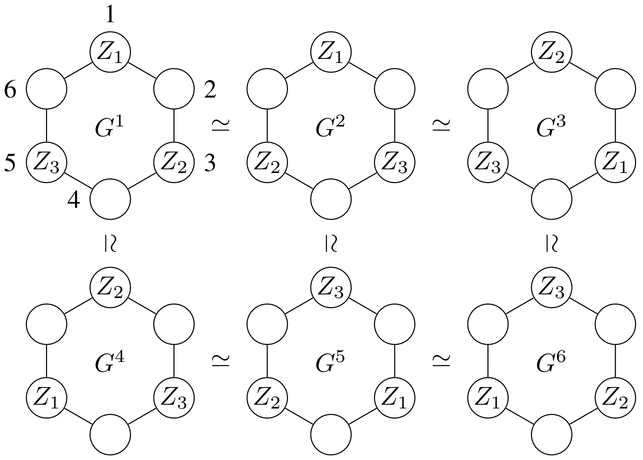


Fig. 4. These 6 isomorphic graphs prove that the set $\{1, 3, 5\}$ is exploratory equivalent for the graph $G'$ of Fig. 2.

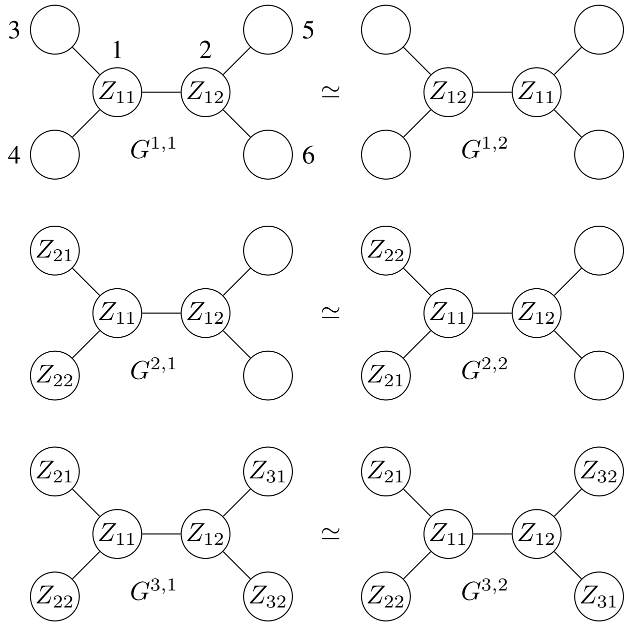


Fig. 5. The three pairs of isomorphic graphs proving that the ordered partition $\langle\{1, 2\}, \{3, 4\}, \{5, 6\}\rangle$ is exploratory equivalent for the graph $G$ of Fig. 1.

*Theorem 1:* The MaxExploreq problem is $GI$-hard.

*Proof:* The theorem can be proved by a polynomial-time reduction of the graph isomorphism problem to the MaxExploreq problem. Let $G$ and $H$ be graphs for which one would like to determine whether they are isomorphic. Let us form a graph $G'$ by adding a vertex $u_0$ to the graph $G$ and connecting it with all the vertices of $G$. In an analogous way, let us form a graph $H'$ from the graph $H$ (we call the added vertex $v_0$). Now we solve the MaxExploreq problem on the graph $G' \cup H'$, i.e., on the disjoint union of the graphs $G'$ and $H'$. We claim that the graphs $G$ and $H$ are isomorphic if and only if any maximum EE partition contains a set with at least one vertex from $G'$ and at least one vertex from $H'$. Let us prove this.

(If) If a maximum EE partition for the graph $G' \cup H'$ contains a set with vertices $u \in V(G')$ and $v \in V(H')$, then there exists an automorphism that maps $u$ to $v$ and $v$ to $u$. Since the graphs $G'$ and $H'$ are both connected, such an automorphism can exist only if the graphs are isomorphic. This implies that the graphs $G$ and $H$ are isomorphic, too.

(Only if) If the graphs $G$ are $H$ isomorphic, then an EE partition for the graph $G' \cup H'$ cannot be maximum unless it contains at least one set with at least one vertex from both $G'$ and $H'$. Indeed, in an EE partition that contains no such set, one can always join the singletons $\{u_0\}$ and $\{v_0\}$ into an EE set $\{u_0, v_0\}$ and thus obtain an EE partition with a higher score, since the vertices $u_0$ and $v_0$, owing to their degree, can only be exploratory equivalent with each other (and they are, if the graphs $G'$ and $H'$, and of course also $G$ and $H$, are isomorphic). ■

Because of its $GI$-hardness, the MAXEXPLOREQ problem for general graphs is unlikely to be solvable in polynomial time. In the rest of this paper, we therefore restrict the problem to trees.

## VI. SOLVING THE MAXEXPLOREQ PROBLEM ON TREES

### A. Prerequisites

Let a graph $T = (V, E)$ be an arbitrary tree. For the sake of simplicity, let us assume that the tree is unlabeled; the algorithm could be fairly straightforwardly generalized to labeled trees. Since $T$ is an arbitrary unrooted tree, we will only speak of *leaves* (vertices with degree 1) but not of the root, parents, and children. Before showing an algorithm for finding a maximum EE partition on $T$, let us present some auxiliary definitions and claims.

*Definition 9 (distance):* The *distance* between vertices $u$ and $v$ in a tree (denoted $d(u, v)$) is the number of edges on the (unique) path from $u$ to $v$.

*Definition 10 (neighborhood):* In a tree, the *neighborhood* of a vertex $u$ at a distance $d$ is the subtree composed of all vertices $v$ such that $d(u, v) \leq d$.

*Definition 11 (eccentricity, center):* The *eccentricity* of a vertex $u$ in a tree is the maximum distance between $u$ and any other vertex, i.e., $e(u) = \max_{v \in V} d(u, v)$. A *center* of the tree is a vertex with minimum eccentricity.

*Theorem 2:* Any tree has either one or two centers. If it has two, they are adjacent.

*Proof:* Let us focus on a longest path (several such paths are possible) in the tree, and let $u$ and $v$ be the two extreme vertices on that path. The distance between $u$ and $v$ is therefore the greatest possible in the tree. The eccentricity of any vertex $w$ on the path is $e(w) = \max\{d(u, w), d(v, w)\}$; it obviously cannot be lower, but if it were greater, we could form a strictly longer path in the tree (passing through $w$, one of $u$ and $v$, and the most remote vertex from $w$), contradicting our assumption. Any center $c$ of the tree has to be located somewhere on the path from $u$ to $v$, for if we had, say, a putative center $c'$ outside of that path, then $e(c') = \max\{d(u, c'), d(v, c')\}$ would be greater than $e(c) = \max\{d(u, c), d(v, c)\}$. Since a center is a vertex with the lowest eccentricity, we have only two possibilities:

- If $d(u, v)$ is odd, the tree has exactly one center $c$, and it is located halfway between $u$ and $w$, such that $d(c, u) = d(c, v)$.
- If $d(u, v)$ is even, the tree has two adjacent centers $c_1$ and $c_2$ such that $d(c_1, u) = d(c_2, v)$.

■

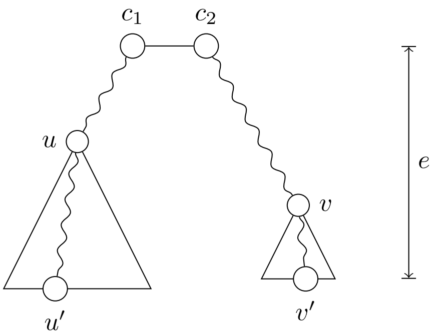


Fig. 6. An illustration of the proof of Lemma 3.

*Lemma 3:* Let $c_1$ and $c_2$ be the center(s) of the tree (by Theorem 2, we may have $c_1 = c_2$). If vertices $u$ and $v$ are exploratory equivalent, we have $d(u, c_1) = d(v, c_2)$.

*Proof:* Let us assume that $d(u, c_1) \neq d(v, c_2)$. Without loss of generality, we may further assume that $d(u, c_1) < d(v, c_2)$. Consider Fig. 6. Let $u'$ be a leaf such that $d(c_1, u') = e(c_1) = e(c_2) = e$ (since $c_1$ is a center, such a leaf must exist). Likewise, let $v'$ be a leaf such that $d(c_2, v') = e$. Since $d(u, c_1) < d(v, c_2)$, we have $d(u, u') > d(v, v')$. This means that there is a leaf at the distance of $d(u, u')$ from $u$, but there cannot be any leaf at the same distance from $v$. The neighborhoods of $u$ and $v$ at the distance $d(u, u')$ are therefore non-isomorphic, which implies that the vertices $u$ and $v$ cannot be automorphically mapped to each other. Consequently, the vertices $u$ and $v$ are not exploratory equivalent. ■

*Definition 12 (centrifugal subtree):* Let $u$ be a vertex connected with vertices $v_1, \ldots, v_k$, and let $v_i$, for some $i \in \{1, \ldots, k\}$, be the sole vertex on the path from $u$ to the center(s) of the tree. The *centrifugal subtree* of the vertex $u$ is the tree composed of the vertex $u$ and of all vertices on the paths starting at $u$, passing through $v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k$, respectively, and finishing at leaves.

Informally, the centrifugal tree of a given vertex $u$ contains the vertex $u$ and all vertices 'below' it in the direction away from the center(s). The triangles in Fig. 6 represent the centrifugal subtrees of the vertices $u$ and $v$.

*Lemma 4:* If vertices $u$ and $v$ of the tree $T$ are exploratory equivalent, they have isomorphic centrifugal subtrees.

*Proof:* If the centrifugal subtrees are not isomorphic, the vertices $u$ and $v$ cannot be automorphically mapped to each other, since there exists some distance $d$ at which their neighborhoods are not isomorphic. Therefore, the vertices cannot be exploratory equivalent. ■

*Lemma 5:* If vertices $v_1, \ldots, v_k$ of the tree $T$ are connected to the same vertex and all their centrifugal subtrees are mutually disjoint and isomorphic, then the vertices $v_1$, …, $v_k$ are exploratory equivalent.

*Proof:* It is easy to see that the vertices $v_1$, …, $v_k$ (and with them the entire corresponding centrifugal subtrees) can be automorphically mapped to each other in all $k!$ possible ways, which means that they are exploratory equivalent. ■

*Lemma 6:* Let the tree have two distinct centers, $c_1$ and $c_2$. If the centrifugal subtrees of $c_1$ and $c_2$ are isomorphic, then the centers $c_1$ and $c_2$ are exploratory equivalent.

*Proof:* Here, the same argument applies as in the proof of Lemma 5. ■

*Lemma 7:* Let $\mathcal{P} = \langle P_1, \ldots, P_s \rangle$ be an ordered partition of the vertex set $V$ such that the following holds:

- The set $P_1$ is exploratory equivalent in the sense of Lemma 5 or Lemma 6.
- Every set $P_i$ with $i > 1$ is exploratory equivalent in the sense of Lemma 5.
- If the vertices of $P_j$, together with their common neighbor, all belong to the centrifugal subtree of some vertex $v \in P_i$, then $j > i$.

If all the above conditions are met, the partition $\mathcal{P}$ is exploratory equivalent.

*Proof:* By Lemmas 5 and 6, the set $P_1 = \{u_{11}, \ldots, u_{1k_1}\}$ is exploratory equivalent. Let us fix the labels of $u_{11}$, …, $u_{1k_1}$ to $Z_{11}$, …, $Z_{1k_1}$. Is the set $P_2 = \{u_{21}, \ldots, u_{2k_1}\}$ still exploratory equivalent? Yes, owing to the third condition in the lemma, it holds that for each vertex $u_{1i}$ ($i \in \{1, \ldots, k_1\}$) the centrifugal subtrees of the vertices $u_{21}, \ldots, u_{2k_2}$ are either disjoint from the centrifugal tree of $u_{1i}$ or completely contained within it. In both cases, the vertices $u_{21}$, …, $u_{2k_2}$ (and with them the entire centrifugal subtrees) can be automorphically mapped to each other in all $k_2!$ possible ways, even if the vertices $u_{11}$, …, $u_{1k_1}$ have distinct unique labels. The second case is illustrated in Fig. 7. Since the same reasoning applies all the way to the set $P_s$, we can conclude that the partition is indeed exploratory equivalent. ■

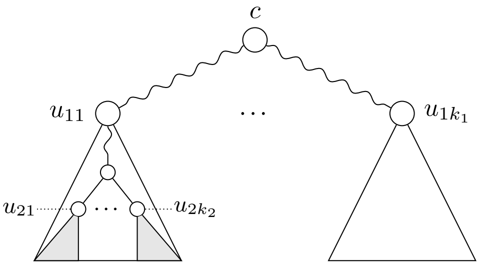


Fig. 7. An illustration of the second case in the proof of Lemma 7. The subtrees represented by the two large triangles are isomorphic, and so are those represented by the two small ones.

For the tree of Fig. 8, an ordered EE partition in the sense of Lemma 7 is $\langle 17 \mid 15, 16 \mid 11 \mid 12 \mid 13 \mid 14 \mid 1, 2, 3 \mid 4, 5 \mid 6, 7 \mid 8, 9, 10 \rangle$. Non-singleton sets are represented by different colors.

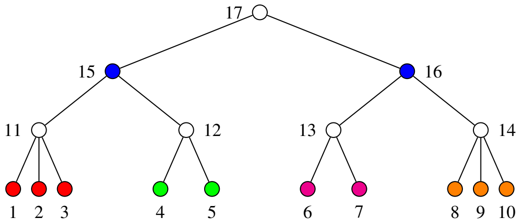


Fig. 8. A sample tree.

*Lemma 8:* If $u$ and $v$ are EE vertices at a distance greater than 2, then there also exist EE vertices $u'$ and $v'$ at a distance of at most 2.

*Proof:* Let us first assume that $d = d(u, v)$ is even. Let $w$ be the sole vertex such that $d(u, w) = d(v, w) = d/2$. Now let $u'$ and $v'$ be the neighbors of $w$ on the paths from $w$ to $u$ and $w$ to $v$, respectively. The distance between $u'$ and $v'$ is therefore 2. We claim that the vertices $u'$ and $v'$ are exploratory equivalent if so are $u$ and $v$. Indeed! The automorphism that maps $u$ to $v$ and vice versa maps the entire centrifugal subtree of $u'$ to the centrifugal subtree of $v'$ and vice versa. In particular, $u'$ is mapped to $v'$ and vice versa, which means that $u'$ and $v'$ are exploratory equivalent, too. (However, note that the sets $\{u, v\}$ and $\{u', v'\}$ cannot *both* be part of the same EE partition!)

If $d(u, v)$ is odd, then it follows from Lemma 3 that the tree $T$ has two distinct centers, $c_1$ and $c_2$, and that $d(u, c_1) = d(v, c_2)$. An automorphism that maps $u$ to $v$ (and $v$ to $u$) also maps $c_1$ to $c_2$ (and $c_2$ to $c_1$), implying that the vertices $c_1$ and $c_2$ are exploratory equivalent, too. ■

*Lemma 9:* If $u_1$, …, $u_k$ are EE vertices, then there also exist EE vertices $u'_1$, …, $u'_k$ such that $d(u_i, u_j) \leq 2$ for all distinct pairs $i, j \in \{1, \ldots, k\}$. Furthermore, if $k > 2$, then $d(u_i, u_j) = 2$ for all distinct pairs $i, j \in \{1, \ldots, k\}$.

*Proof:* The first part of the lemma is a straightforward generalization of Lemma 8. As for the second part, observe that if distinct vertices $u$, $v$, and $w$ are exploratory equivalent, we cannot have $d(u, v) = d(v, w) = 1$ and $d(u, w) = 2$; such vertices would then form a 3-vertex line subgraph $u - v - w$, which can never have more than 2 automorphisms, but a set of three vertices can be exploratory equivalent only if at least 3! automorphisms exist. The case $d(u, v) = d(u, w) = d(v, w) = 1$ is clearly impossible in a tree, and so $d(u, v) = d(u, w) = d(v, w) = 2$ remains as the only possibility. ■

*Lemma 10:* There exists a maximum EE partition $\mathcal{P} = \{P_1, \ldots, P_s\}$ of the tree $T$ such that for each $i \in \{1, \ldots, s\}$ the distance between each pair of vertices in $P_i$ is at most 2.

*Proof:* Let $\mathcal{R} = \{R_1, \ldots, R_s\}$ be a maximum EE partition such that a set $R \in \mathcal{R}$ does not conform to the conditions in the lemma. By Lemma 9, the set $R$ can be replaced by the corresponding EE set $R'$ of vertices at a distance of at most 2. We now claim that the partition $\mathcal{R}' = \mathcal{R} \setminus \{R\} \cup \{R'\}$ is also exploratory equivalent. To see this, consider the operation performed in the proof of Lemma 8. Let $v$ be the vertex such that $d(v, u_1) = \ldots = d(v, u_k)$. (If $k = 2$, such a vertex might not exist, but then we have $d(c_1, u_1) = d(c_2, u_2)$, and the same logic applies.) By replacing the EE set $R = \{u_1, \ldots, u_k\}$ with the EE set $R' = \{u'_1, \ldots, u'_k\}$ (where $u'_1, \ldots, u'_k$ are the neighbors of $v$ on the paths from $v$ to $u_1, \ldots, u_k$, respectively), the resulting partition remains exploratory equivalent, since an automorphism that maps $u_i$ to $u_j$ also maps the entire path from $v$ to $u_i$ to the path from $v$ to $u_j$ and since the selection of $R$ into an EE partition precludes the selection of any other set containing vertices on different paths from $v$ to $u_1, \ldots, u_k$. However, the opposite is not necessarily the case. While the choice of $R'$ does, of course, preclude the selection of $R$, it might not rule out everything 'between' $R'$ and $R$. For instance, in the example shown in Fig. 8, it would be unwise to select the EE set $R_1 = \{1, 8\}$ into an EE partition, since the most we could then possibly attain would be the partition $\{1, 8 \mid 2, 3 \mid 4, 5 \mid 6, 7 \mid 9, 10 \mid \text{singletons}\}$. However, if we instead choose the set $R'_1 = \{15, 16\}$, we can obtain the clearly better (and indeed maximum) partition $\{15, 16 \mid 1, 2, 3 \mid 4, 5 \mid 6, 7 \mid 8, 9, 10 \mid \text{singletons}\}$. ■

The above lemma tells us that when searching for a maximum EE partition, we may safely ignore any pairs of vertices at the distance greater than 2. This important fact is the basis for the algorithm we show below.

### *B. The algorithm*

We are now ready to present a polynomial-time algorithm that constructs a maximum EE partition for a given tree. The algorithm is shown as Alg. 1. At the beginning, the algorithm assigns the so-called *ornament* $*$ to each leaf of the given tree; all other vertices are assigned the ornament $\epsilon$. The algorithm then proceeds in a reverse breadth-first fashion: in each iteration, the vertices connected to the (current) leaves that have at most one $\epsilon$-ornamented neighbor receive their ornaments, constructed from the ornaments of their leaf neighbors. Simultaneously, the leaves are removed from the tree. The output of the algorithm is a partition of the vertex set of the original tree.

Figures 9 and 10 provide two examples for Alg. 1. The numbers beside the vertices indicate the order in which the ornaments are assigned to the vertices (of course, the order within each iteration is arbitrary), while the boxes show the ornaments. The vertices with the same non-white color belong to the same set in the returned partition. For the tree of Fig. 9, the algorithm thus produces the partition $\langle 9, 10 \mid 7, 8 \mid 5, 6 \mid 4 \mid 3 \mid 2 \mid 1\rangle$. The partition for the tree of Fig. 10 is $\langle 8 \mid 6, 7 \mid 1 \mid 5 \mid 4 \mid 3 \mid 2\rangle$.

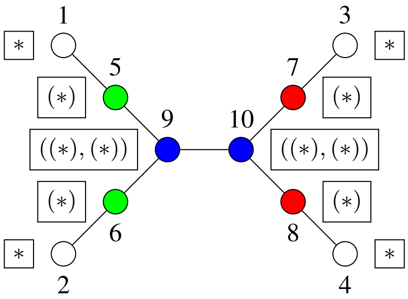


Fig. 9. The execution of Alg. 1 on a sample tree.

**Algorithm 1** An algorithm for solving the MAXEXPLOREQ problem on a given tree $T$.

```
function FINDMAXPARTITION($T = (V, E)$)
    for all $v \in V$ do $\mathrm{orn}(v) \leftarrow \epsilon$
    $L \leftarrow$ the set of leaves of $T$
    for all $v \in L$ do $\mathrm{orn}(v) \leftarrow *$
    $t \leftarrow 0$
    while $|V| > 2$ do
        $W = \{w \in V \mid (\exists v \in L\colon (w, v) \in E)$ and
            only one neighbor $w'$ of $w$ has $\mathrm{orn}(w') = \epsilon\}$
        for all $w \in W$ do
            $\langle r_1, \ldots, r_m\rangle \leftarrow$ the leaves connected to $w$,
                lexicographically sorted by their ornaments
            $\mathrm{orn}(w) \leftarrow (\mathrm{orn}(r_1), \ldots, \mathrm{orn}(r_m))$
            $R \leftarrow \{r_1, \ldots, r_m\}$
            while $R \neq \emptyset$ do
                $r \leftarrow$ any vertex from $R$
                $t \leftarrow t + 1$
                $P_t \leftarrow \{r' \in R \mid \mathrm{orn}(r') = \mathrm{orn}(r)\}$
                $R \leftarrow R \setminus P_t$
                remove each vertex in $P_t$ from $T$
        $L \leftarrow$ the set of leaves of $T$
    if $|V| = 1$ then return $\langle V, P_t, P_{t-1}, \ldots, P_1\rangle$
    else
        $\{u, v\} \leftarrow V$
        if $\mathrm{orn}(u) = \mathrm{orn}(v)$ then return $\langle V, P_t, \ldots, P_1\rangle$
        else return $\langle \{u\}, \{v\}, P_t, \ldots, P_1\rangle$
```

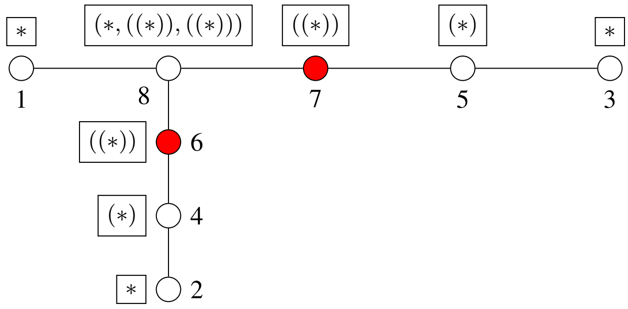


Fig. 10. The execution of Alg. 1 on a sample tree.

*Lemma 11:* In each iteration, the algorithm removes from the current tree all vertices that are farthest from the center(s) of the current tree.

*Proof:* In each iteration, the algorithm removes all leaves except those whose neighbor is connected to at least two non-leaves. However, such leaves cannot be at the greatest distance from the center(s). ■

*Corollary 12:* After each iteration, the center(s) of the resulting tree are coincident with those of the original tree.

*Corollary 13:* The (at most two) vertices that remain in the set $V$ after the main loop of the algorithm are exactly the center(s) of the tree.

*Lemma 14:* At the end of the algorithm, two vertices have equal ornaments if and only if their centrifugal subtrees are isomorphic.

*Proof:* Since the algorithm proceeds from the leaves towards the centers, it builds the ornaments of individual vertices from the ornaments of their centrifugal subtrees. By construction, the ornament of a vertex reflects the structure of its centrifugal subtree. The lexicographical ordering ensures that two vertices with isomorphic centrifugal subtrees also have equal ornaments. As for the 'only if' part, consider that vertices with non-isomorphic centrifugal subtrees cannot possibly have equal ornaments; any valid ornament takes the form $(s_1, s_2, \ldots, s_k)$, where $s_1, \ldots, s_k$ are individual subornaments, and there is exactly one way to split a valid ornament into valid constituents. Therefore, it cannot happen that two non-isomorphic subtrees 'accidentally' receive equal ornaments. ■

*Lemma 15:* For a given tree, the ordered partition produced by the algorithm is exploratory equivalent.

*Proof:* First, the properties from Lemmas 5, 6, and 14 ensure that each set from the EE partition is *individually* exploratory equivalent. Indeed, the algorithm adds a non-singleton set to the partition only if all vertices from that set have the same neighbor and isomorphic centrifugal subtrees. Second, does the returned ordered partition $\langle P_s, P_{s-1}, \ldots, P_1 \rangle$ (where $s = t + 1$ or $t + 2$, depending on the situation after the main loop) conform to the conditions in Lemma 7, which guarantee 'EE-ness'? It does! The first two conditions are clearly met; the two centers, if they exist and if they are exploratory equivalent, constitute the set $P_s$. As for the third condition, consider that the algorithm 'peels' the tree from the leaves towards the centers and that the EE sets are stacked into the partition in the reverse order. These facts ensure that the centrifugal subtree of a vertex in $P_j$ can be a subtree of the centrifugal subtree of a vertex in $P_i$ only if $j > i$. ■

*Lemma 16:* The EE partition produced by the algorithm contains all possible EE sets $P = \{u_1, \ldots, u_k\}$ such that for each distinct pair $i, j \in \{1, \ldots, k\}$ we have $d(u_i, u_j) \leq 2$.

*Proof:* By Lemma 3, the vertices $\{u_1, \ldots, u_k\}$ of an EE set all have the same distance from the center. If the distance between each pair of them is at most 2, then they must be connected with the same vertex or, if $k = 2$, the vertices $u_1$ and $u_2$ can also be the two centers of the tree. By Lemma 3, EE vertices are always located at the same distance from the tree center(s); by Lemma 4, they must also have isomorphic centrifugal subtrees. The algorithm produces *all* such sets that fulfill these conditions: (1) it considers all sets of vertices that have the pairwise distance of exactly 2 and are located at the same distance from the tree center(s); (2) if the tree has two centers, the algorithm will, at the very end, certainly check whether they have the same ornaments; (3) the algorithm adds each such set of vertices to the output partition provided that the vertices have isomorphic centrifugal subtrees. ■

*Theorem 17:* For a given tree, Alg. 1 produces a maximum exploratory equivalent (ordered) partition.

*Proof:* By Lemma 15, the partition is exploratory equivalent. By Lemma 10, every tree has a *maximum* EE partition such that all pairwise distances in each EE set are at most 2. Since, by Lemma 16, the algorithm constructs an EE partition out of *all* such EE sets in the input tree, and since each of these sets contains as many vertices as possible (owing to line 17 in Alg. 1), the output EE partition certainly has the maximum score. ■

We have just proved that the algorithm indeed solves the MAXEXPLOREQ problem for trees. To get an estimate on the algorithm's complexity, we proceed as follows. For each vertex of the tree, one has to sort the signatures of its children (the neighbors of that vertex in its centrifugal subtree). Each vertex has $O(n)$ children, and the length of each ornament is $O(n)$, giving $O(n^2 \log n)$ to sort the signatures of the children of each vertex. Since there are $O(n)$ vertices, the total complexity of the algorithm is $O(n^3 \log n)$. We have the following theorem:

*Theorem 18:* Algorithm 1 is a polynomial-time algorithm for tree graphs.

## VII. MAXEXPLOREQ ON SMALL TREES

In order to give us some insight into the problem, we performed a small empirical study of MAXEXPLOREQ on the set of small trees. This analysis will show us how the symmetries (that we can detect and exploit with MAXEXPLOREQ) are present in the studied set of trees.

Since trees are a ubiquitous structure, there are a lot of applications that can benefit from the symmetries found with our algorithm. An application that directly relates to the set of small trees is graphlet counting. In [28], the authors present a method for counting graphlets by exploiting many symmetries of small graphs (up to 5 nodes). Their method is currently considered one of the best, and with the use of MAXEXPLOREQ some of those symmetries could also be used to count larger graphlets much faster than with the straightforward approach.

For the analysis of this set of trees, we generated all nonisomorphic unlabeled trees of sizes 2 to 20 (let us call the set $T_2^{20}$) and computed MAXEXPLOREQ on every generated tree. Table I gives the number of trees for each size; in parentheses, we give the number of trees that have only the trivial automorphism, i.e., all sets in the MAXEXPLOREQ partition are singletons. Figure 11 shows the distribution of MAXEXPLOREQ score in $T_2^{20}$. This histogram shows that maximum EE partitions are non-trivial in a vast majority of trees.

To view the potential of MAXEXPLOREQ in more detail, let us examine trees of different sizes separately. For each separate set, we computed the median MAXEXPLOREQ score. The resulting chart is shown in Fig. 12. From this chart, we can see the potential speedup of at least half of the trees in the set of all trees with the same size. For example, for the trees of size 15, half of the trees have the potential speedup of 12, and for larger trees the median value is even larger, implying an almost exponential growth of the median value.

The two charts shown above demonstrate features of the MAXEXPLOREQ score, but they do not show the structure of individual partitions in any way. To show a feature of the MAXEXPLOREQ partitions, we measured the frequencies of the largest set in the partition (for each tree in $T_2^{20}$). Figure 13 shows the frequencies of these sets. In this histogram, we can see that most of the partitions are composed of pairs and triplets; the frequency of other partitions drops exponentially.

TABLE I. NUMBER OF NONISOMORPHIC TREES OF A SPECIFIED SIZE. IN PARENTHESES, THE NUMBER OF TREES WITHOUT AUTOMORPHISMS IS GIVEN.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| size | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| #trees | 1 (0) | 1 (0) | 2 (0) | 3 (0) | 6 (0) | 11 (1) | 23 (1) | 47 (3) | 106 (6) | 235 (15) |
| size | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | |
| #trees | 551 (29) | 1301 (67) | 3159 (139) | 7741 (313) | 19320 (671) | 48629 (1487) | 123867 (1487) | 317955 (7264) | 823065 (16137) | |

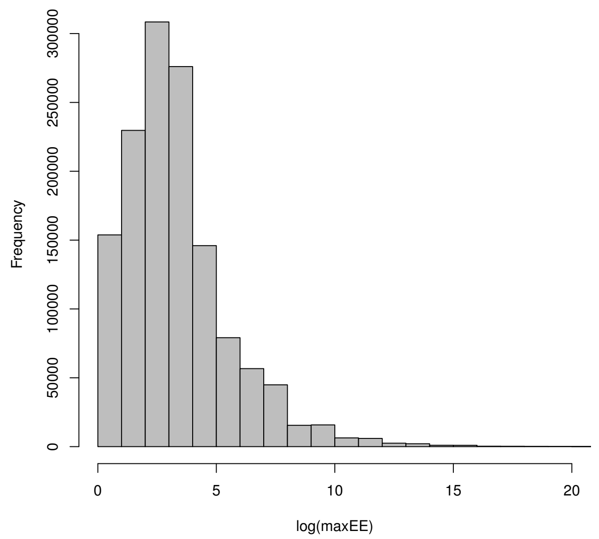


Fig. 11. The histogram of the frequencies of MAXEXPLOREQ values (logarithmic x axis) in $T_2^{20}$

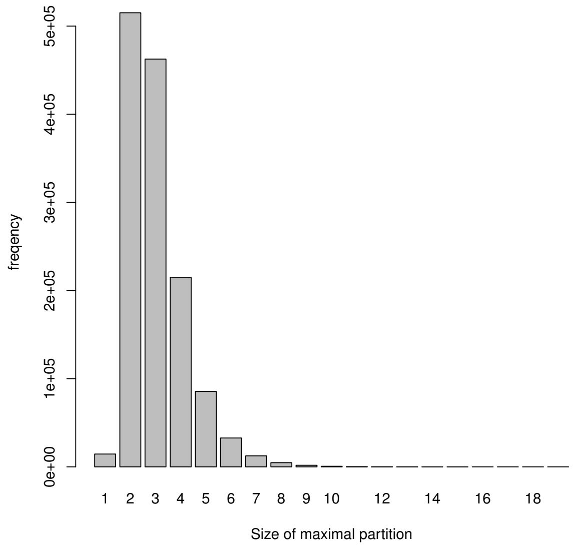


Fig. 13. Frequencies of the largest set in the MAXEXPLOREQ partition for each tree.

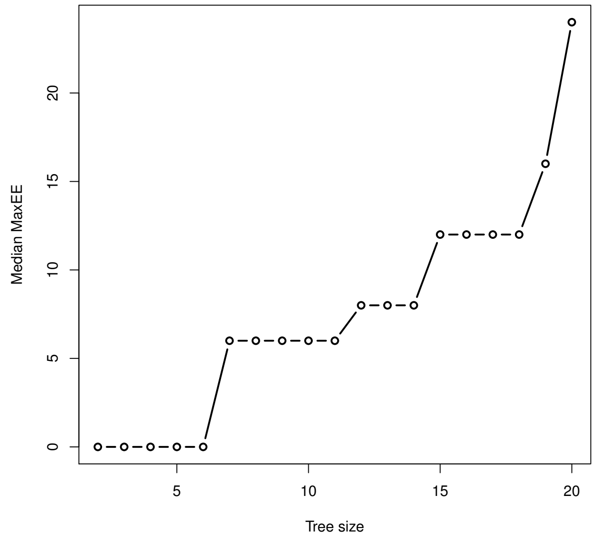


Fig. 12. The median values of MAXEXPLOREQ for all trees in $T_2^{20}$ computed separately for all trees of the same size.

## VIII. EXPLORATORY EQUIVALENCE AND THE SUBGRAPH ISOMORPHISM PROBLEM

We motivated exploratory equivalence by its application to the subgraph isomorphism problem. In this section, we will establish the relationship between these two concepts in more depth. First, note that there is a bijection between the set of automorphisms of a pattern graph $G$ and the set of isomorphisms between $G$ and each occurrence of $G$ in an arbitrary host graph $H$:

*Lemma 19:* If $G'$ is an occurrence of a graph $G$ in a graph $H$, then for each automorphism of $G$ there exists an isomorphism $G \to G'$, and vice versa.

*Proof:* A subgraph isomorphism between the graphs $G$ and $H$ is an isomorphism between two copies of the graph $G$ ($G$ and $G'$ in our case). An automorphism of $G$ can be interpreted in exactly the same way: as an isomorphism between two copies of $G$. ■

*Lemma 20:* Let $G = (V, E)$ be a pattern graph, and let $G'$ be its occurrence in a host graph $H$. If a set $\{v_1, v_2, \ldots, v_k\} \subseteq V$ is exploratory equivalent, then there exists an isomorphism $h' \colon G \to G'$ such that $h'(v_1) < h'(v_2) < \ldots < h'(v_k)$.

*Proof:* Without loss of generality, we may assume that the graph $G'$ consists of the vertices $u_1, u_2, \ldots, u_k$ such that $u_1 < u_2 < \ldots < u_k$. Let $h_0 \colon G \to G'$ be an isomorphism such that $h_0(v_i) = u_{\sigma(i)}$ (for each $i \in \{1, \ldots, k\}$), where $\sigma$ is some permutation of the set $\{1, \ldots, k\}$. Since the set $\{v_1, v_2, \ldots,$

$v_k\}$ is exploratory equivalent, there exists an automorphism of $G$ for each of the $k!$ permutations of the set. One of those automorphisms (e.g., $h$) has the property that $h(v_i) = v_{\sigma^{-1}(i)}$ for each $i \in \{1, \ldots, k\}$. Now, let us define the isomorphism $h' = h_0 \circ h$. For each $i \in \{1, \ldots, k\}$, we have $h'(v_i) = h_0(h(v_i)) = h_0(v_{\sigma^{-1}(i)}) = u_i$ and hence $h'(v_1) < h'(v_2) < \ldots < h'(v_k)$. ∎

*Theorem 21:* Let $G$ be a pattern graph, and let $G'$ be its occurrence in a host graph $H$. If $\mathcal{P} = \langle P_1, \ldots, P_s\rangle$ (where $P_i = \{v_{i1}, \ldots, v_{ik_i}\}$ for $i \in \{1, \ldots, s\}$) is an EE ordered partition, then there exists an isomorphism $h' \colon G \to G'$ such that $h'(v_{i1}) < h'(v_{i2}) < \ldots < h'(v_{ik_i})$ for all $i \in \{1, \ldots, s\}$.

*Proof:* Since $P_1 = \{v_{11}, \ldots, v_{1k_1}\}$ is an EE set, Lemma 20 ensures the existence of an isomorphism $h_1 \colon G \to G'$ with the property $h_1(v_{11}) < h_1(v_{12}) < \ldots < h_1(v_{1k_1})$. Now, the fact that $\mathcal{P}$ is an EE ordered partition implies that the set $P_2$ remains exploratory equivalent even if we assign unique labels to the vertices in $P_1$. In other words, even if the set of all possible isomorphisms $h \colon G \to G'$ is restricted to those that satisfy $h(v_{11}) < h(v_{12}) < \ldots < h(v_{1k_1})$, there will still exist an isomorphism $h_2 \colon G \to H$ such that $h_2(v_{21}) < h_2(v_{22}) < \ldots < h_2(v_{2k_2})$ (in addition to $h_2(v_{11}) < \ldots < h_2(v_{1k_1})$). The same reasoning applies for $P_3$, $P_4$, etc., up to $P_s$. Therefore, there exists an automorphism $h'$ such that $h'(u_{i1}) < \ldots < h'(u_{ik_i})$ for all $i \in \{1, \ldots, s\}$. ∎

From Theorem 21, it follows that if $\mathcal{P} = \{P_1, \ldots, P_s\}$ is an EE partition of a pattern graph $G$, we can safely impose the constraints $h(v_{i1}) < \ldots < h(v_{ik_i})$ (for each $i \in \{1, \ldots, s\}$) while searching for subgraph isomorphisms $h \colon G \to H$ in an arbitrary host graph $H$. However, these constraints are not necessarily optimal in the sense of redundancy elimination: they might not reduce the number of residual isomorphisms between $G$ and each of its occurrences in $H$ to 1. For example, consider the graphs $G$ and $H$ in Fig. 14. The set of automorphisms of $G$ (and simultaneously the set of $G$-to-$H$ isomorphisms) is $\{1234,\ 2341,\ 3412,\ 4123,\ 4321,\ 3214,\ 2143,\ 1432\}$, and the maximum EE partition is $\{1, 3 \mid 2, 4\}$, giving the set of constraints $\{h(1) < h(3),\ h(2) < h(4)\}$. However, these constraints still retain two isomorphisms, 1234 and 2143. (In fact, two isomorphisms would remain regardless of how the vertices of $H$ were numbered.) This means that in this case, the set of constraints resulting from the maximum EE partition does *not* eliminate the entire redundancy in subgraph isomorphism search and hence cannot be regarded as optimal. Incidentally, the optimal set of constraints is $\{h(1) < h(3) < h(4)\}$, but the partition $\{1, 3, 4 \mid 2\}$ is not exploratory equivalent.

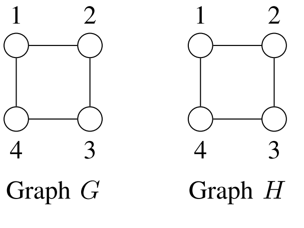


Fig. 14. A sample pattern graph $G$ and host graph $H$.

In the case of general graphs, a maximum EE partition might lead to a suboptimal set of constraints because the number of automorphisms might be greater than the score of a maximum EE partition. In a tree, however, these two values are exactly the same.

We will state the following lemma without a formal proof. We can employ the same techniques as in the proofs of lemmas and theorems of Section VI. In addition, note that the set of automorphisms forms a group: if $h$ and $h'$ are automorphisms, then $h \circ h'$ and $h' \circ h$ are automorphisms, too.

*Lemma 22:* For any tree $T$, the following properties hold:

- If $T$ contains an automorphism that maps a vertex $u$ to a vertex $v$, then the vertices $u$ and $v$ are exploratory equivalent.
- Let $c_1$ and $c_2$ be the centers of the tree (we may also have $c_1 = c_2$). If the tree has an automorphism $h$ such that $h(u) = v$, then (1) $d(u, c_1) = d(v, c_2)$ and (2) the centrifugal subtrees of $u$ and $v$ are isomorphic.
- For each nontrivial tree automorphism $h$, there exists a pair of vertices $u$ and $v$ such that $1 \le d(u, v) \le 2$ and $h(u) = v$.
- Let the set $P = \{v_1, \ldots, v_k\}$ be exploratory equivalent. Let $h$ and $h'$ be automorphisms such that $h(v_1) = h'(v_1) = \sigma(v_1)$, …, $h(v_k) = h'(v_k) = \sigma(v_k)$ for some permutation $\sigma$ of the set $P$. If, on top of that, $h(u) = w_1$ and $h'(u) = w_2$ with $\{u, w_1, w_2\} \cap \{v_1, \ldots, v_k\} = \emptyset$ and $w_1 \neq w_2$, then the set $W = \{w_1, w_2\}$ is exploratory equivalent independently of the set $P$, which means that the sets $P$ and $W$ can both be part of the same EE partition.
- If the sets of tree vertices $P = \{u_1, \ldots, u_p\}$ and $Q = \{v_1, \ldots, v_q\}$ are both exploratory equivalent and if they cannot be extended by any other vertices without becoming non-EE, then the number of distinct automorphisms permuting the sets $P$ and $Q$ is equal to $p!\,q!$ only if (1) the centrifugal subtrees of the vertices in $P$ are all pairwise disjoint from the centrifugal subtrees of the vertices in $Q$ or (2) the vertices $v_1, \ldots,$ $v_q$ are all part of the centrifugal subtree of a vertex $u_i$ for some $i \in \{u_1, \ldots, u_p\}$ (or vice versa). Otherwise, the number of distinct automorphisms permuting the sets $P$ and $Q$ is $p! = q!$. In this case, we must have $p = q$, and each of the vertices of $Q$ is located in the centrifugal subtree of a different vertex of $P$ (or the other way around).

*Theorem 23:* Let $\mathcal{P} = \langle P_1, P_2, \ldots, P_s\rangle$ be the maximum EE ordered partition obtained by Alg. 1 for a given tree $T$. Then the number of automorphisms of $T$ is $|P_1|! \ldots |P_s|!$.

*Proof:* If $s = 1$, the vertices of $P_1$ can be mapped to each other in all $|P_1|!$ ways, which means that there are at least $|P_1|!$ automorphisms. However, the number of automorphisms is exactly $|P_1|!$. Suppose there were two automorphisms, $h$ and $h'$, for the same permutation of the set $P_1$. In particular, if $P_1 = \{v_1, \ldots, v_k\}$, suppose that $h(v_i) = h'(v_i) = \sigma(v_i)$ for all $i \in \{1, \ldots, k\}$ and for some permutation $\sigma$ of $P_1$. For $h$ and $h'$ to be distinct, we must have, say, $h(u) = w_1$ and $h'(u) = w_2$ with $w_1 \neq w_2$ and $u, w_1, w_2 \in V \setminus P_1$. However, in this case, the set $\{w_1, w_2\}$ is exploratory equivalent independently of the set $P_1$ (Lemma 22) and is therefore part of the same maximum EE partition as the set $P_1$.

Now, let us assume that the theorem holds for some $s > 1$, and let us verify that it also holds for $s + 1$. Indeed: for any fixed permutation of the vertices in the set $P_1 \cup \ldots \cup P_s$, there are exactly $|P_{s+1}|!$ automorphisms, one for each permutation of the set $P_{s+1}$, which, together with the inductive assumption, gives the property stated in the theorem. The number of automorphisms cannot possibly be more than that; if it were, that would imply the exploratory equivalence of some set not present in $\mathcal{P}$ (independently of the sets in $\mathcal{P}$) or the fact that the set $P_s$ is not at the bottom of the centrifugal subtree containment hierarchy (which would, in turn, imply that $\mathcal{P}$ is not an EE ordered partition). ■

Theorem 23 implies the optimality of the subgraph isomorphism constraints derived from a maximum EE partition for an arbitrary tree. In the search for occurrences of a pattern tree $T$ in an arbitrary host graph $H$, the use of these constraints reduces the number of generated isomorphisms between $T$ and each of its occurrences in $H$ to 1, thus eliminating the automorphism-induced redundancy completely.

## IX. Conclusion

Recently, we defined the so-called MaxExplorEq problem, the goal of which is to find a maximum exploratory equivalent (EE) partition of the vertex set of a given graph $G$. This problem is closely related to the problem of finding occurrences of a graph $G$ in a graph $H$ (the subgraph isomorphism problem), since every EE partition of $G$ determines a set of redundancy reduction constraints that can be safely imposed during the subgraph isomorphism search. In the MaxExplorEq problem, we try to find an EE partition that gives rise to the optimal set of constraints in terms of redundancy elimination in subgraph isomorphism search.

In this paper, we proved that MaxExplorEq is $GI$-hard, which means that it is unlikely to be solvable in polynomial time. For this reason, we restricted the MaxExplorEq problem to an important subclass of graphs — the class of trees. By devising a polynomial-time algorithm, we showed that the restricted MaxExplorEq problem belongs to the complexity class $P$. Our algorithm finds a maximum EE partition in time $O(n^3 \log n)$, where $n$ is the number of vertices of the input tree. Note that in contrast to the algorithms presented in our previous paper [9], Alg. 1 does not require or enumerate the set of automorphisms of the given tree. If it did, it could not possibly run within polynomial-time bounds, since the number of automorphisms for a tree with $n$ vertices can be up to $(n-1)!$.

Besides that, we showed that the score of the maximum EE partition is equal to the number of automorphisms in the case of trees, but not necessarily in the case of general graphs. For any tree, a maximum EE partition thus gives rise to an optimal set of subgraph isomorphism search constraints.

To demonstrate the large potential of MaxExplorEq, we performed a small empirical study on the set of all trees of sizes 2 to 20. The study demonstrates that large speedups could be obtained in various search algorithms, especially for finding tree-shaped patterns in larger structures. The automorphisms on trees have been, of course, well known for a long time; however, our algorithm finds the partition of nodes that can be completely interchanged in search algorithms, and thus we give an explicit recipe on how to exploit these symmetries.

Could we apply the approach presented in this paper to general graphs? The $GI$-hardness of the MaxExplorEq problem does not offer much hope to find a polynomial-time algorithm for arbitrary graphs. Nevertheless, the lemmas and theorems of Section VI — if, of course, they could really be extended to arbitrary graphs in some way — might at least give rise to a relatively efficient branch-and-bound algorithm for finding a maximum EE partition. However, a number of problems will have to be solved before arriving at a viable algorithm. A general graph might have an arbitrary number of centers, and it is not yet clear whether the concepts such as ‘centrifugal subtree’ could be generalized at all.

## References

[1] J. Leskovec and E. Horvitz, “Geospatial structure of a planetary-scale social network,” *IEEE Transactions on Computational Social Systems*, vol. 1, no. 3, pp. 156–163, 2014. doi: 10.1109/TCSS.2014.2377789

[2] D. K. Agrafiotis, V. S. Lobanov, M. Shemanarev, D. N. Rassokhin, S. Izrailev, E. P. Jaeger, S. Alex, and M. Farnum, “Efficient Substructure Searching of Large Chemical Libraries: The ABCD Chemical Cartridge,” *Journal of Chemical Information and Modeling*, pp. 3113–3130, 2011. doi: 10.1021/ci200413e

[3] J. M. Barnard, “Substructure searching methods: Old and new,” *Journal of Chemical Information and Computer Sciences*, vol. 33, no. 4, pp. 532–538, 1993. doi: 10.1021/ci00014a001

[4] M. O. Jackson, *Social and Economic Networks*. Princeton, NJ, USA: Princeton University Press, 2008. ISBN 0691134405, 9780691134406

[5] D. Knoke, *Political Networks: The Structural Perspective*, ser. Structural Analysis in the Social Sciences. Cambridge University Press, 1994. ISBN 9780521477628. [Online]. Available: http://books.google.si/books?id=9djTlBLaOccC

[6] B. Hopkins, “Kevin Bacon and graph theory,” *Problems, Resources, and Issues in Mathematics Undergraduate Studies (PRIMUS)*, vol. 14, no. 1, pp. 5–11, 2004. doi: 10.1080/10511970408984072

[7] L. D. Sailer, “Structural equivalence: Meaning and definition, computation and application,” *Social Networks*, vol. 1, pp. 73–90, 1978. doi: 10.1016/0378-8733(78)90014-X

[8] M. G. Everett and S. P. Borgatti, “Regular equivalence: General theory,” *Journal of mathematical sociology*, vol. 19, no. 1, pp. 29–52, 1994. doi: 10.1080/0022250X.1994.9990134

[9] J. Mihelič, L. Fürst, and U. Čibej, “Exploratory equivalence in graphs: Definition and algorithms,” in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems (FedCSIS), Warsaw, Poland, September 7-10, 2014*, 2014. doi: 10.15439/2014F352 pp. 447–456.

[10] F. Margot, “Pruning by isomorphism in branch-and-cut,” *Mathematical Programming*, vol. 94, pp. 71–90, 2002. doi: 10.1007/s10107-002-0358-2

[11] J. Ostrowski, J. Linderoth, F. Rossi, and S. Smriglio, “Orbital branching,” in *Integer Programming and Combinatorial Optimization, 12th International IPCO Conference, Ithaca, NY, USA, June 25–27, 2007, Proceedings*, 2007. doi: 10.1007/978-3-540-72792-7_9 pp. 104–118.

[12] D. Erwin and F. Harary, “Destroying automorphisms by fixing nodes,” *Discrete mathematics*, vol. 306, pp. 3244–3252, 2006. doi: 10.1016/j.disc.2006.06.004

[13] R. Ladner, “On the structure of polynomial time reducibility,” *Journal of the ACM*, vol. 22, pp. 155–171, 1975. doi: 10.1145/321864.321877

[14] B. D. McKay, “Practical graph isomorphism,” in *Congressus Numerantium 30: 10th Manitoba Conference on Numerical Mathematics and Computing, Winnipeg, Canada, 1980*, 1981, p. 45–87.

[15] B. D. McKay and A. Piperno, "Practical graph isomorphism, II," *Journal of Symbolic Computation*, vol. 60, pp. 94–112, 2013. doi: 10.1016/j.jsc.2013.09.003

[16] T. Junttila and P. Kaski, "Engineering an efficient canonical labeling tool for large and sparse graphs," in *Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments, ALENEX 2007, New Orleans, LA, USA, January 6, 2007*, 2007. doi: 10.1137/1.9781611972870.13

[17] ——, "Conflict propagation and component recursion for canonical labeling," in *Theory and Practice of Algorithms in (Computer) Systems – First International ICST Conference, TAPAS 2011, Rome, Italy, April 18–20, 2011, Proceedings*, ser. LNCS 6595. Springer, 2011. doi: 10.1007/978-3-642-19754-3_16 pp. 151–162.

[18] P. T. Darga, M. H. Liffiton, K. A. Sakallah, and I. L. Markov, "Exploiting structure in symmetry detection for CNF," in *Proceedings of the 41st Design Automation Conference, DAC 2004, San Diego, CA, USA, June 7-11, 2004*, 2004. doi: 10.1145/996566.996712 pp. 530–534.

[19] P. T. Darga, K. A. Sakallah, and I. L. Markov, "Faster symmetry discovery using sparsity of symmetries," in *Proceedings of the 45th Design Automation Conference, DAC 2008, Anaheim, CA, USA, June 8-13, 2008*, 2008. doi: 10.1145/1391469.1391509 pp. 149–154.

[20] H. Katebi, K. A. Sakallah, and I. L. Markov, "Symmetry and satisfiability: An update," in *Theory and Applications of Satisfiability Testing – SAT 2010, 13th International Conference, Edinburgh, UK, July 11-14, 2010, Proceedings*, ser. LNCS 6175. Springer, 2010. doi: 10.1007/978-3-642-14186-7_11 pp. 113–127.

[21] M. N. Velev and R. E. Bryant, "Effective use of boolean satisfiability procedures in the formal verification of superscalar and VLIW microprocessors," in *Proceedings of the 38th Design Automation Conference, DAC 2001, Las Vegas, NV, USA, June 18-22, 2001*, 2001. doi: 10.1145/378239.378469 pp. 226–231.

[22] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*. Addison-Wesley, 1974. ISBN 978-0201000290

[23] S. A. Cook, "The complexity of theorem-proving procedures," in *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing (STOC), Shaker Heights, Ohio, USA, May 3-5, 1971*, 1971. doi: 10.1145/800157.805047 pp. 151–158.

[24] S. Arora and B. Barak, *Computational complexity: a modern approach*. Cambridge University Press, 2009. ISBN 9780521424264

[25] F. V. Fomin and D. Kratsch, *Exact Exponential Algorithms*. Springer, 2011. ISBN 978-3-642-16532-0

[26] J. R. Ullmann, "An Algorithm for Subgraph Isomorphism," *Journal of the ACM*, vol. 23, pp. 31–42, 1976. doi: 10.1145/321921.321925

[27] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub)graph isomorphism algorithm for matching large graphs." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1367–1372, 2004. doi: 10.1109/TPAMI.2004.75

[28] T. Hočevar and J. Demšar, "A combinatorial approach to graphlet counting," *Bioinformatics*, vol. 30, no. 4, pp. 559–565, 2014. doi: 10.1093/bioinformatics/btt717

DOI: 10.15439/2015F293
ACSIS, Vol. 5

# Evaluation of Metaheuristics for Robust Graph Coloring Problem

Zbigniew Kokosiński and Łukasz Ochał
Cracow University of Technology,
Faculty of Electrical and Computer Eng.,
Dept. of Automatic Control and Information Technology,
Warszawska 24, 31-155 Cracow, Poland
Email: {zk@pk.edu.pl; lukasz.ochal@gmail.com}

***Abstract*—In this paper a new formulation of the robust graph coloring problem (RGCP) is proposed. In opposition to classical GCP defined for the given graph *G(V,E)* not only elements of E but also Ē can be subject of color conflicts in edge vertices. Conflicts in Ē are assigned penalties 0<P(e)<1. In addition to satisfying constraints related to the number of colors and/or a threshold of the acceptable sum of penalties for color conflicts in graph complementary edges (rigidity level), a new bound called the relative robustness threshold (RRT) is proposed. Then two metaheuristics – SA, TS and their parallel analogues PSA and PTS – for that version of RGCP are presented and experimentally compared. For comparison we use DIMACS graph coloring instances in which a selected percentage of graph edges E is randomly moved to Ē. Since graph densities and chromatic numbers of DIMACS GCP instances are known in advance, the RGCP instances generated on their basis are more suitable for testing algorithms than totally random instances used so far. The results of the conducted experiments are presented and discussed.**

## I. Introduction

The classical graph $k$-colorability problem belongs to the class of NP-hard combinatorial problems [9]. This decision problem is defined for an undirected graph $G=(V,E)$ and positive integer $k \leq |V|$ : is there an assignment of available $k$ colors to graph vertices, providing that adjacent vertices receive different colors? In optimization version of the basic problem called GCP, a conflict-free coloring with minimum number of colors $k$ is searched.

Many particular colorings of graph vertices and/or edges represent solutions of variety of practical problems that can be modeled by graphs with specific constraints put on the elements of the sets $V$ and $E$. With additional assumptions many variants of the coloring problem can be defined such as equitable coloring, sum coloring, contrast coloring, harmonious coloring, circular coloring, consecutive coloring, list coloring, total coloring etc. [14], [18].

One of the most interesting variants of GCP is the robust graph coloring problem (RGCP) [24]. It models a class of vertex coloring problems in which adjacency relation between graph vertices is not “stable”. In certain circumstances nonadjacent vertices $u$ and $v$ can become adjacent and the edge $(u,v)$ is assigned a penalty $0<P(u,v)<1$ when there is a color conflict : $c(u)=c(v)$. If all penalties $P(u,v)$ are known it is possible to define requirements for solution feasibility. A conflict-free coloring of all vertices in $E$ is required, while a number of penalized color conflicts in the set of complementary edges $\bar{E}$ , not exceeding certain threshold (f.i. rigidity level) can be tolerated. Given a number of colors $k$, the coloring with a lower rigidity level is more robust. In general, feasibility conditions can be expressed in terms of the maximum number of colors used and an upper bound on a cost function.

Similarly as GCP also RGCP is known to be NP-hard [26]. Therefore, application of approximate algorithms and metaheuristics for solving this problem is reasonable [1], [10], 11]. In literature no r-approximation algorithms for RGCP were reported so far. Research conducted in this area contains a number of algorithms and metaheuristics for RGCP [2], [19], [20], [24], [25] a new formulation of specific robust coloring problems [4] and a combination of system robustness and fuzziness [8], [12]. Research results were gathered and summarized in [24]. Other recent papers on system robustness are [7] and [23].

In research papers [19], [20] there appears a problem in experimental verification of the investigated metaheuristic methods. How to measure quality of the solution, when nothing is known about chromatic properties of the given graph? How to validate the assumed penalty threshold for a feasible solution? In our approach RGCP is redefined in order to allow the system designer to use a new cost function - the relative robustness of the solution, which can be expressed by a percentage – the relative robustness $RR=100\%$ means a conflict-free vertex coloring in both edges $E$ and complementary edges $\bar{E}$ with $P(e)>0$. This view is very natural and meets common expectations of system designers. For experimental verification two metaheuristics - Tabu Search (TS), Simulated Annealing (SA) and their parallel versions – PTS and PTA – are used. In standard and parallel versions they were applied earlier in similar research [3], [6], [16], [17], [19], [20], [21], [22]. As input graphs we used DIMACS graph coloring instances [27], [28], [29] in

978-83-60810-66-8/$25.00©2015, IEEE

which a constant percentage of graph edges *E*, denoted by *E'* is assigned penalties *0<P(e)<1*. Since graph densities and chromatic numbers of DIMACS GCP instances are known in advance [15], the RGCP instances generated on their basis are more suitable for testing algorithms than totally random instances used so far. The presented results justify both theoretical assumptions and application of parallel metaheuristics for solving RGCP problem.

In the next section RGCP problem is defined together with its new formulation. TS/PTS and SA/PSA algorithms and their parameters are presented in sections III. Then, in section IV, computer experiments are described and their results discussed. In conclusion some general suggestions related to the obtained results and future research in this area are derived.

## II. Robust Graph Coloring Problem

GCP is defined for an undirected graph *G(V,E)* as an assignment of available colors {*1, . . . , k*} to graph vertices providing that adjacent vertices receive different colors and the number of colors *k* is minimal. The resulting coloring *c* is called conflict–free and *k* is called graph chromatic number *χ(G)*.

### A. RGCP – a simple formulation

RGCP is defined for undirected weighted graph *G(V,E)* with function *w(e)*= $p_{uv} \in [0,1]$, as an assignment of available colors {*1, . . ., k*} to graph vertices, providing that

$$\forall (u,v) \in E \ : (p_{uv} = 1) \Rightarrow \ c(u) \neq c(v) \qquad (1)$$

and rigidity level for $\bar{E}$ is minimum

$$RL(c) = \sum_{((u,v)\in\overline{E})\wedge(c(u)=c(v))\wedge(p(u,v)>0)} p_{uv} \qquad (2)$$

In some cases weight (penalty) $p_{uv}$ may be considered as a probability of edge existence; in the classical vertex coloring $p_{uv} \in \{0,1\}$.

For most practical problems it suffices, that

$$RL(c) \leq T \qquad (3)$$

where :

*T* – is an assumed threshold.

The question is what value of *T* is reasonable for the modeled problem? How it reflects system robustness? What level of *T* should be guaranteed for the given system? In order to answer such questions an alternative formulation of RGCP problem is proposed. The alternative formulation of RGCP does not change the nature of the problem, but allows the system designer to apply the relative robustness level instead of the absolute robustness level which is not known in advance.

### B. RGCP – an alternative formulation

Let us characterize system robustness more precisely. The robustness threshold is given by the following formula

$$RT(c) = \frac{\mathrm{RRT(c)}}{100\%} \sum_{(e\in\overline{E})\wedge(p((u,v))>0)} p_{uv} \ , \qquad (4)$$

where:

*RRT(c)* – is a relative robustness threshold set up by the designer and expressed in [%].

Thus, our optimization goal is to find a coloring *c* satisfying equation (1) and inequalities (5) and (6):

$$\sum_{((u,v)\in\overline{E})\wedge(c(u)\neq c(v))\wedge(p(u,v)>0)} p_{uv} \geq RT(c) \qquad (5)$$

$$\frac{\sum_{((u,v)\in\overline{E})\wedge(c(u)\neq c(v))\wedge(p(u,v)>0)} p_{uv}}{\sum_{((u,v)\in\overline{E})\wedge(p(u,v)>0)} p_{uv}} \cdot 100\% \geq RRT(c) \qquad (6)$$

### C. Generation of RGCP instances

Parametrized RGCP instances can be generated by a random modification of GCP instances: a percentage of *E'ϵ E* is moved to $\bar{E}$ with weights *0<p(u,v)<1*. In Fig. 1 three out of seven edges are selected at random and assigned new values *p(u,v)*.

### D. Cost function for RGCP

In problem formulation the priority is given to conflict-free coloring of edges with *P(u,v)=1* . Otherwise, the solution is not feasible. Feasible solutions have the cost function:

$$cf(c) = \sum_{(u,v)\in E\cup\overline{E}} q_{uv} \qquad (7)$$

where:

$$q_{uv} = \left| \begin{array}{ll} 1, & if\,(c(u)=c(v)\, and\ p_{uv}=1 \\ p_{uv} & if\,(c(u)=c(v)\, and\ p_{uv}<1 \\ 0 & if\,(c(u)\neq c(v)) \end{array} \right. \qquad (8)$$

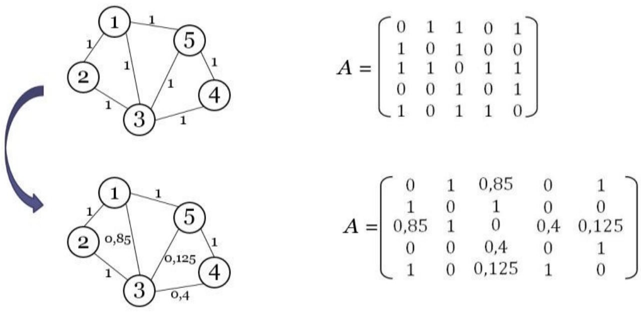


Fig. 1 Generation of RGCP graph instances from GCP instances (42,86% of *E* moved to $\bar{E}$; $\bar{E}$ percentage increased from 30 to 60%).

## III. PTS AND PSA METAHEURISTICS

The applications of basic metaheuristics for RGCP was reported in [19]. The first parallel metaheuristic for RGCP – Parallel Evolutionary Algorithm – was presented in [5]. In the present paper we deal with two other popular parallel metaheuristics PTS and PSA. The details of implementation are skipped here for the sake of brevity. In order to determine their parameters at first we investigate algorithms TS and SA.

### *A. Parameters for Tabu Search Algorithm*

Tabu Search metaheuristic presented in [19] is adapted for parallelization. PTS algorithm includes three TS processes that periodically exchange information when 1/3 and 2/3 of the required RR is obtained. There are at least two key parameters of TS/PTS algorithms that have to be set [6]: tMAX and MaxTenure. This parameters were found experimentally. The results of conducted experiments are shown in Table I and Table II.

The values of parameters recommended for TS and PTS algorithms are as follows: tMAX=10 and MaxTenure=15. As a selection criterion majority of optimum solutions with respect to relative robustness RR was used.

### *B. Parameters for Simulated Annealing Algorithm*

A Simulated Annealing metaheuristic for RGCP presented in [19] is adapted for parallelization. There are three important parameters of SA and PSA algorithms that have to be set [21]: MinIteration, ControlFactor (speed of convergence) and Tmax. These parameters were also found experimentally. The results of conducted experiments are shown in Tables III, IV, and V. We can assume Tmin=0,25.

PSA algorithm includes also three SA processes that periodically exchange information when 1/3 and 2/3 of the required RR is obtained. All processes resume computations with new best solution. The values of parameters recommended for PSA are the following: MinIteration=5, ControlFactor=0,9 and Tmax=10.

Table I Efficiency of TSA with tMax (MaxTenure=10)

| Graphs | tMax : 5 | | | **tMax : 10** | | | tMax : 15 | | |
|---|---|---|---|---|---|---|---|---|---|
| | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] |
| queen5.5_40<br>χ(g)=5<br>dens.=54% | 6,3 | 0,4 | 91,0 | **4,9** | **0,3** | **93,0** | 6,2 | 0,5 | 91,2 |
| games120_4<br>χ(g)=9<br>dens.=9% | **0** | 260 | **100** | **0** | **247** | **100** | **0** | 253 | **100** |
| myciel7_40<br>χ(g)=8<br>dens.=13% | 1,8 | 795 | 99,8 | **0** | 766 | **100** | **0** | **745** | **100** |

Table II Efficiency of TSA with MaxTenure (tMAX=10)

| Graphs | MaxTenure : 5 | | | MaxTenure : 10 | | | **MaxTenure : 15** | | |
|---|---|---|---|---|---|---|---|---|---|
| | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] |
| queen5.5_40<br>χ(g)=5<br>dens.=54% | 5,2 | 0,5 | 92,6 | 6,5 | 0,4 | 90,7 | **3,8** | **0,3** | **94,6** |
| games120_4<br>χ(g)=9<br>dens.=9% | **0** | 252 | **100** | **0** | 260 | **100** | **0** | **250** | **100** |
| myciel7_40<br>χ(g)=8<br>dens.=13% | 0,2 | 879 | **100** | 0,7 | **776** | **100** | 0,4 | 833 | **100** |

Table III Efficiency of SA algorithm with MinIteration (Tmin=0,25; Tmax=10; ControlFactor=0,9)

| Graphs | **MinIteration : 5** | | | MinIteration : 10 | | | MinIteration : 15 | | |
|---|---|---|---|---|---|---|---|---|---|
| | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] |
| queen5.5_40<br>χ(g)=5<br>dens.=54% | 7,7 | **0,3** | 89,1 | **6,5** | 0,6 | **90,8** | 7,9 | 0,9 | 88,8 |
| games120_4<br>χ(g)=9<br>dens.=9% | **3,7** | 29,9 | **98,6** | 15,9 | **14,2** | 94,0 | 9,3 | 18,3 | 96,5 |
| myciel7_40<br>χ(g)=8<br>dens.=13% | **30** | 121 | **97,2** | 11k | 58,8 | - | 31k | **45,7** | - |

Table IV Efficiency of SA algorithm with ControlFactor (Tmin=0,25; Tmax=10; MinIteration=5)

| Graphs | ControlFactor : 0,85 | | | **ControlFactor : 0,9** | | | ControlFactor : 0,95 | | |
|---|---|---|---|---|---|---|---|---|---|
| | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] |
| queen5.5_40<br>χ(g)=5<br>dens.=54% | 14 | **0,2** | 79,3 | **6,4** | 0,3 | **90,9** | 7,9 | 0,5 | 88,9 |
| games120_4<br>χ(g)=9<br>dens.=9% | 104 | **18,3** | 96,1 | **4,1** | 28,7 | **98,5** | 6,4 | 57,3 | 97,6 |
| myciel7_40<br>χ(g)=8<br>dens.=13% | 2k | **75,2** | - | 13,8 | 119 | 98,8 | **5,4** | 240 | **99,5** |

Table V Efficiency of SA algorithm with Tmax (Tmin=0,25; MinIteration=5; ControlFactor=0,9)

| Graphs | Tmax : 5 | | | **Tmax : 10** | | | Tmax : 15 | | |
|---|---|---|---|---|---|---|---|---|---|
| | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] | c.f. | time [s] | RR [%] |
| queen5.5_40<br>χ(g)=5<br>dens.=54% | 10 | **0,2** | 85,8 | **9,5** | 0,3 | **86,6** | 12,0 | 0,3 | 83,1 |
| games120_4<br>χ(g)=9<br>dens.=9% | **4,0** | **24,2** | **98,5** | 5,4 | 29,8 | 98,0 | 6,9 | 30,7 | 97,4 |
| myciel7_40<br>χ(g)=8<br>dens.=13% | 19 | **94,1** | 98,3 | **14,5** | 122 | **98,7** | 18,1 | 137 | 98,3 |

Similarly, as a selection criterion for a given parameter the majority of optimum solutions with respect to relative robustness RR was used.

SA/PSA algorithm has SA2/PSA2 version with automatic computation of Tmax (the initial temperature), cf. [21].

The computed parameters of TS and SA metaheuristics were used for parallel versions of both iterative methods.

## IV. Experimental Results

The application has been written in C++, while GUI in C#, accordingly. Microsoft Visual Studio 2008 v.9.0 was used. All computer experiments we performed on a machine with Intel Core 2 Duo, CPU P8400, 2,27 GHz, 4,00 GB of RAM memory.

### *A. Goals of optimization and algorithms*

The main purpose of optimization was obtaining the best available robustness with minimum number of colors used. It is possible to:

a. compute maximal RR for the given number of colors (algorithms: TS, SA, SA2)

b. compute the above in parallel (algorithms: PTS, PSA2)

c. find a robust coloring with minimal number of colors for the given RR (algorithms TS, SA$^C$, SA2$^C$)

d. compute the above in parallel (algorithms PTS$^C$, PSA2$^C$)

The program solves RGCP problem providing value of cost function, relative robustness RR and the number of colors. There is a pool of algorithm's variants to choose from, including sequential and parallel versions.

In next subsections a number of experiments performed with the help of that program is reported.

### *B. TS versus SA*

The first experiment was devoted to efficiency comparison of sequential versions of the two basic metaheuristics. For comparison 9 DIMACS graphs were selected the number of colors was set up to *k*= *χ(G)*. The results are shown in Fig. 2. For most combinations of test graphs and the size of the set *E'* the TS outperforms SA in terms of relative robustness RR of the modeled system. Typically, TS was able to achieve 100% RR and never less than 95%. SA issued a bit worse results: for only three graphs maximum RR=100% was obtained. In majority of cases RR was within the range 91–99%. In a single case when SA algorithm failed to achieve a conflict-free coloring for a graph with density 46%, the value *k* was incremented. Basically, more dense graph are more difficult to color. SA is simpler than TS, much faster for bigger graphs and its power relies on randomization in a higher degree than TS which is more precise in searching for a good solution, checking all color combinations for all vertices in each iteration. Regardless of *E'* size both algorithms delivered solutions with similar values of cost function and RR. However, when *E'* size is bigger the number of iterations required to obtain a conflict-free coloring decreases in both methods and the speed of TS decreases. The graph density is more essential than the graph size.

### *C. TS$^C$ versus SA$^C$ and SA2$^C$*

Three subsequent experiments were based on eight graphs instances with the percentage of *E'* equal 60%. The number of colors was computed that allows to achieve the given level of system relative reliability RR on the levels 70%, 85%, and 95% respectively.

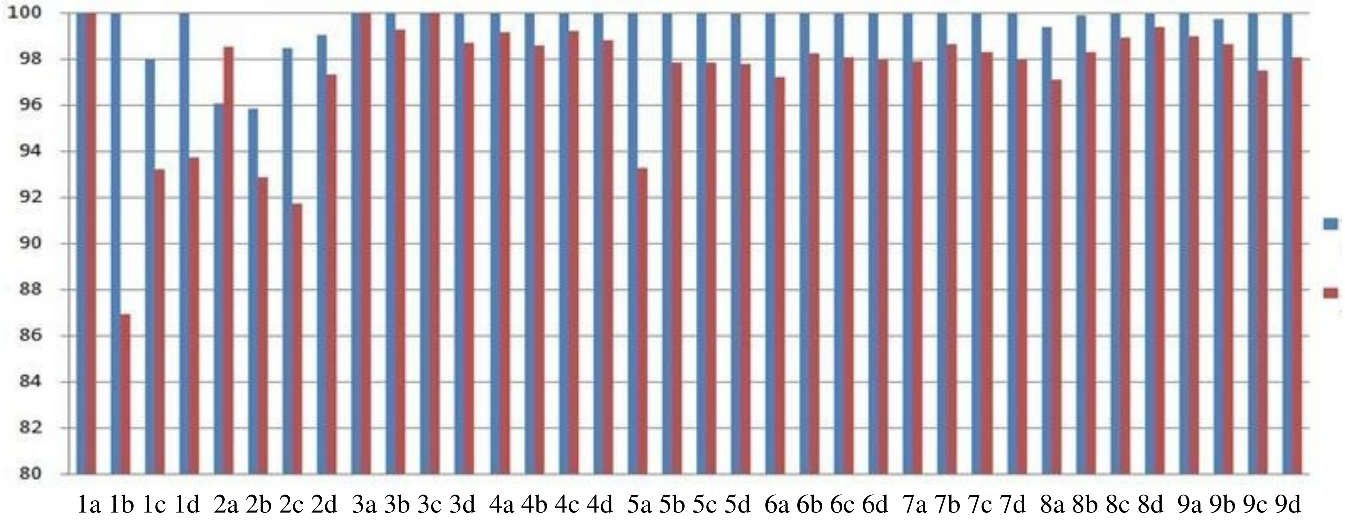


Fig. 2 Relative robustness *RR* [TS-blue, SA-red]. Graphs: 1-queen5.5, 2-queen6.6, 3-myciel, 4-huck, 5-david, 6-games120, 7-anna, 8-mulsol.i.4, 9-myciel7. *E'* : a=10%, b=20%, c=40%, d=60%. Number of colors k= *χ(G)*.

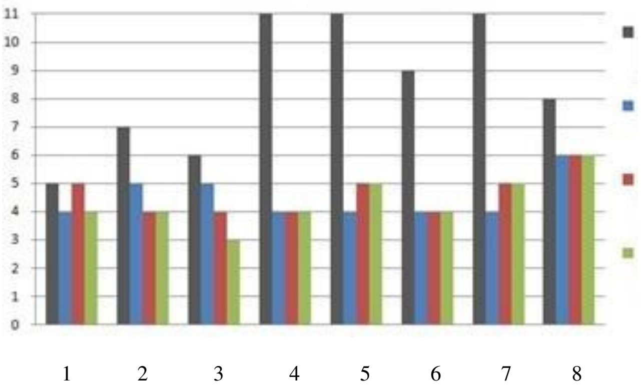


Fig. 3 Number of colors required for *RR*=70%, [$\chi(G)$,TS$^C$,SA$^C$,SA2$^C$]. Basic graphs: 1-queen5.5, 2-queen6.6, 3-myciel, 4-huck, 5-david, 6-games120, 7-anna, 8-myciel7; *E'* – 60%.

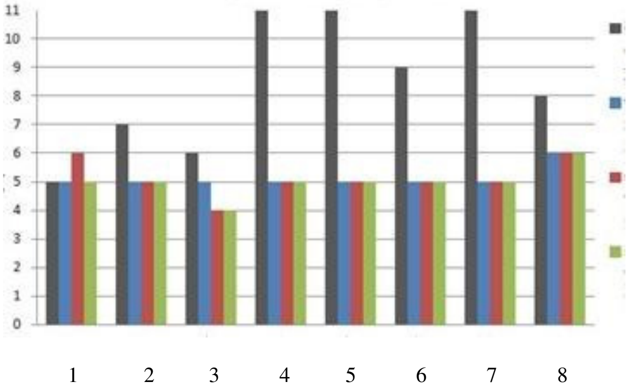


Fig. 4 Number of colors required for *RR*=85%, [$\chi(G)$,TS$^C$,SA$^C$,SA2$^C$]. Graphs: 1-queen5.5, 2-queen6.6, 3-myciel, 4-huck, 5-david, 6-games120, 7-anna, 8-myciel7; *E'* – 60%.

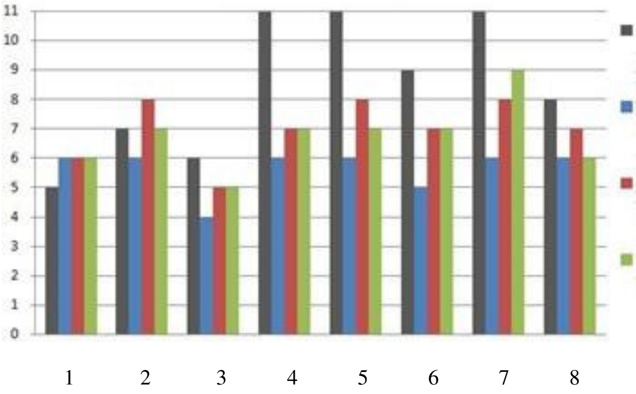


Fig. 5 Number of colors required for *RR*=95%, [$\chi(G)$,TS$^C$,SA$^C$,SA2$^C$]. Graphs: 1-queen5.5, 2-queen6.6, 3-myciel, 4-huck, 5-david, 6-games120, 7-anna, 8-myciel7; *E'* – 60%.

In Fig. 3-5 the order of bars characterizing experiments for the given input graph is as follows: $\chi(G)$, TS$^C$, SA$^C$ and SA2$^C$. The results depicted in Fig. 3 present the number of colors used by the corresponding methods for the set of all 8 graphs with RR=70%. The average number of colors used is as follows : TS$^C$=4,5 , SA$^C$=4,625 and SA2$^C$=4,375 , with the average sum of $\chi(G)$ equal 8,5.

Similarly, the results depicted in Fig. 4 can be characterized in short by average number of colors used by the corresponding methods for the set of all 8 graphs RR=85% : TS$^C$=5,125 , SA$^C$=5,125 and SA2$^C$=5,0 with the same sum of $\chi(G)$.

Finally, the general results depicted in Fig. 5 can be summarized by average number of colors used by the corresponding methods for the set of all 8 graphs with RR=95%: TS$^C$=5,625 , SA$^C$=7,0 and SA2$^C$=6,75 with respect to the sum of $\chi(G)$ as above.

### *D. PTS$^C$ versus PSA2$^C$*

The experiment reported in subsection C was then repeated for parallel metaheuristics PTS$^C$ and PSA2$^C$ (with an automatic computing of initial temperature Tmax).

Three subsequent experiments were based on eight graphs instances with the percentage of *E'* equal 60%. The number of colors was computed that allows to achieve the given level of system relative reliability RR on the levels 70%, 85%, and 95% respectively

Results of the research concerning minimization of colors in a conflict free robust graph coloring with fixed RR level can be summarized by the average number of colors used by the corresponding sequential and parallel methods for the set of all eight graphs from subsection C: PTS$^C$=5,5, TS$^C$=5,625, PSA2$^C$=6,625 and SA2$^C$=6,75 when the average $\chi(G)$ is 8,5. As expected , the results obtained by parallel metaheuristics are slightly improved in comparison to classical metaheuristics.

In addition total computation time of sequential and parallel versions of both metaheuristics was compared for the set of all eight graphs from subsection C. Average processing time of PTS$^C$ is 666,8 [s] while TS$^C$ 693,2 [s]. The average processing time of PSA2$^C$ is 674,8 [s] while SA2$^C$ 435,9 [s]. Solutions generated by PTS$^C$ are often repeatable while PSA2$^C$ results are less stable and with similar quality as those from SA2$^C$ .

## V. Conclusions

In this paper new formulation of RGCP problem is given that seems to be more appropriate for designers of robust systems. Relative robustness is a versatile measure for characterization of any robust system modeled by a graph. For experimental verification two popular parallel metaheuristics TS/PTS and SA/PSA were used.

We proposed a new method of test instance generation by random modification of a given percentage *E'* of graph edges *E*. DIMACS hard-to-color graph instances were used for modification. The results confirm that the proposed approach and the used tools can be efficiently used for practical applications.

An interesting goal of the future research is to apply to RGCP – and verify experimentally – more metaheuristics like Parallel Evolutionary Algorithm (PEA), Parallel

Immune Algorithm (PIA), Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO) and others [10], [11]. For particular applications the robustness measures can be modified to reflect specific properties of the given system.

## References

[1] E. Alba (Ed.), *Parallel metaheuristic - a new class of algorithms,* John Wiley & Sons, 2005. DOI: 10.1002/0471739383

[2] C. Archetti, N. Bianchessi and A. Hertz, "A branch-and-price algorithm for the robust graph coloring problem", *Les Cahiers du Gerad,* G-2011–75, Montreal, 2011.

[3] H. Bouziri, M. Jouini, "A tabu search approach for the sum coloring problem", *Electronic Notes in Discrete Mathematics,* Vol. 36, pp. 915–922, 2010 DOI:10.1016/j.endm.2010.05.116

[4] R. L. Bracho, J. R. Rodriguez, F. J. Martinez : "Algorithms for Robust Graph Coloring on Paths", in *Proc. 2nd International Conference on Electrical and Electronics Engineering,* Mexico, pp. 9–12, IEEE 2005. DOI: 10.1109/ICEEE.2005.1529561

[5] G. Chrząszcz, *Parallel evolutionary algorithm for robust scheduling in power systems,* M.Sc. thesis, Cracow University of Technology, (in Polish) 2009.

[6] J. Dąbrowski, "Parallelization techniques for tabu search", in *Proc. 8th Int. Conference on Applied Parallel Computing: State of the Art in Scientific Computing.* – 2007.

[7] S. Deleplanque, J.-P. Derutin, A. Quilliot, "Anticipation in the Dial-a-Ride Problem: an introduction to the robustness", *Proc. of the 2013 Federated Conference on Computer Science and Information Systems,* FedCSIS'2015, Kraków, Poland, pp. 299–305, 2013.

[8] A. Dey, R. Pradhan, A. Pal, T Pal, "The Fuzzy Robust Graph Coloring Problem", in S. C., Biswal B. N., Udgata S. K., Mandal J.K. (Eds.) in *Proc. of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014,* Vol. 1, Advances in Inteligent Systems and Computing Proceedings, Vol. 327, pp. 805–813, Springer 2015. DOI: 10.1007/978-3-319-11933-5_91

[9] R. Garey, D. S. Johnson, *Computers and intractability. A guide to the theory of NP-completeness,* Freeman, 1979.

[10] M. Gendreau, J. Y. Potvin (Eds.), *Handbook of metaheuristics,* International Series in Operations Research & Management Science, Springer US, 2010. DOI: 10.1007/978-1-4419-1665-5

[11] F. Glover, G. A. Kochenberger (Eds.), *Handbook of metaheuristics,* Kluwer 2003.

[12] B. Gładysz, "Fuzzy robust courses scheduling problem", *Fuzzy Optimization and Decision Making,* Vol. 6, pp. 155–161, 2007. DOI: 10.1007/s10700-007-9303-0

[13] G. Hutchinson, "Partitioning algorithms for finite sets", *Comm. ACM,* No. 6, pp. 613–614, 1963.

[14] T. R. Jensen, B. Toft, *Graph coloring problems,* Wiley Interscience, 1995.

[15] D. S. Johnson, M. A. Trick: Cliques, coloring and satisfiability: Second DIMACS Implementation Challenge, *DIMACS Series in Discr. Math. and Theor. Comp. Sc.* Vol. 26, 1996.

[16] Z. Kokosiński, M. Kołodziej, K. Kwarciany, "Parallel genetic algorithm for graph coloring problem", in *Proc. of the International Conference on Computational Science, ICCS'2004,* LNCS, Vol. 3036, pp. 215–222, 2008. DOI: 10.1007/978-3-540-24685-5_27

[17] Z. Kokosiński, "Parallel metaheuristics in graph coloring", *Bulletin of the National University "Lviv Politechnic",* Series: Computer sciences and information technologies, No. 744 , pp. 209–214, 2012.

[18] M. Kubale (Ed.): *Graph colorings,* American Mathematical Society, 2004. DOI: 10.1090/conm/352

[19] A. Lim, F. Wang, "Metaheuristic for robust graph coloring problem", in *Proc. 16th IEEE Int. Conference on Tools with Artificial Intelligence,* ICTAI. – 2004.

[20] A. Lim, F. Wang, "Robust graph coloring for uncertain supply chain management", in *Proc. 38th Annual Hawaii Int. Conf. on System Science, HICSS 2005,* p. 81b, IEEE 2005. DOI : 10.1109/HICSS.2005.526

[21] S. Łukasik, Z. Kokosiński, G. Świętoń, "Parallel simulated annealing algorithm for graph coloring problem", in *Proceedings of the Int. Conference Parallel Processing and Applied Mathematics, PPAM'2007,* LNCS, Vol. 4967, pp. 229–238, 2008. DOI: 10.1007/978-3-540-68111-3_25

[22] A. Pahlavani, K. Eshghi, "A hybrid algorithm of simulated annealing and tabu search for graph colouring problem", *International Journal of Operational Research,* Vol.11, No.2, pp. 136–159, 2011. DOI: 10.1504/IJOR.2011.040694

[23] D. Ruta, "Robust Method of Sparse Feature Selection for Multi-Label Classification with Naive Bayes", *Proc. of the 2014 Federated Conference on Computer Science and Information Systems,* FedCSIS'2014, Warsaw, Poland, pp. 375–380, 2014. DOI: 10.15439/2014F502

[24] F. Wang, Z. Xu, "Metaheuristics for robust graph coloring", *Journal of Heuristics,* Vol. 19, No.4, pp.529–548, 2013. DOI: 10.1007/s10732-011-9180-4

[25] M. Xu, Y. Wang, and A. Wei: "Robust graph coloring based on the matrix semi-tensor product with application to examination timetabling", *Control Theory and Technology,* Vol. 12, No. 2, pp. 187–197, 2014. DOI: 10.1007/s11768-014-0153-7

[26] J. Yáñez J., J. Ramirez, "The robust coloring problem", *European Journal of Operational Research,* Vol.148, No.3, pp. 546–558, 2003. DOI: 10.1016/S0377-2217(02)00362-4

[27] COLOR web site. Available at : http://mat.gsia.cmu.edu/COLOR/instances.html .

[28] DIMACS ftp site. Available at : ftp://dimacs.rutgers.edu/pub/challenge/graph/benchmarks/ .

DOI: 10.15439/2015F261
ACSIS, Vol. 5

# Small Populations, High-Dimensional Spaces: Sparse Covariance Matrix Adaptation

Silja Meyer-Nieberg
Department of Computer Science,
Universität der Bundeswehr München,
Werner-Heisenberg Weg 37,
85577 Neubiberg, Germany
Email: silja.meyer-nieberg@unibw.de

Erik Kropat
Department of Computer Science,
Universität der Bundeswehr München,
Werner-Heisenberg Weg 37,
85577 Neubiberg, Germany
Email: erik.kropat@unibw.de

***Abstract*—Evolution strategies are powerful evolutionary algorithms for continuous optimization. The main search operator is mutation. Its extend is controlled by the covariance matrix and must be adapted during a run. Modern Evolution Strategies accomplish this with covariance matrix adaptation techniques. However, the quality of the common estimate of the covariance is known to be questionable for high search space dimensions. This paper introduces a new approach by changing the coordinate system and introducing sparse covariance matrix techniques. The results are evaluated in experiments.**

## I. Introduction

EVOLUTIONARY COMPUTATION has a long research tradition. The field comprises today the main classes genetic algorithms, genetic programming, evolution strategies, evolutionary programming, and differential evolution. Evolution strategies (ESs), on which the research presented in this paper focuses, are primarily used for optimizing continuous functions. The function is not required to be analytical.

Evolution strategies rely on mutation, i.e., on the random perturbation of candidate solutions to navigate the search space. The process must be controlled in order to achieve good performance. For this, modern ESs apply covariance matrix adaptation in several variants [1]. Nearly all approaches take the sample covariance matrix into account. This estimator is known to be problematic in the case of small sample sizes compared to the search space dimensionality. Since the population size in evolution strategies is typically considerably smaller, this paper argues that the adaptation process may profit from the introduction of different estimators.

So far, evolutionary algorithms or related approaches have only seldom considered statistical estimation methods targeted at high-dimensional spaces. The reason may be twofold: The improved quality of the estimators induces increased computational costs which may lower the convergence velocity of the algorithm. In addition, the estimators are developed and analyzed for samples of independently, identically distributed random variables. Since evolutionary algorithms deploy selection based on rank or fitness, the assumption of the same distribution is not valid. This may be the reason as to why the literature research has resulted in only one previous approach [2]. There, the authors considered Gaussian based estimation of distribution algorithms. The problem they were faced with concerned a non-positive definiteness of the estimated covariance matrix. Therefore, Dong and Yao augmented the algorithm with a shrinkage procedure to guarantee positive definiteness. Shrinkage is one of the common methods to improve the quality of the sample covariance, see e.g. [3]. While the approach in [2] resembles the Ledoit-Wolf estimator [3], it adapted the shrinkage intensity during the run.

This paper extends the work presented in [4], [5], where Ledoit-Wolf shrinkage estimators were analyzed, combined with a maximum entropy approach, and integrated into evolution strategies. While the results were promising, the question remained how to adapt the parameter of the estimator. Therefore, in this paper, another computational simple estimation method is investigated: thresholding.

The paper is structured as follows. First, modern evolution strategies with covariance adaptation are introduced. Afterwards, a short motivation as to why we think that the covariance computation in ESs may profit from estimation theory for high-dimensional spaces is provided. The next section describes the new approach developed and is followed by the experimental section which compares the new approach against the original ES. Conclusions and a discussion of potential future research constitute the last part of the paper.

### A. Modern Evolution Strategies

This section provides a short introduction into evolutionary algorithms focussing on evolution strategies and covariance matrix adaptation. Evolutionary algorithms (EAs) [6] in general are population-based stochastic search and optimization algorithms used when only direct function measurements are possible.

Their iterative search process requires the definition of termination criteria and stops if these are fulfilled. In each generation, a series of operations is performed: selection for reproduction, followed by offspring creation, i.e. recombination and mutation processes, and finally survivor selection. The initial population of candidate solutions is either drawn randomly from the permissable search space or is initialized based on information already obtained. First of all, the offspring population has to be created. For this, a subset of the parents is determined during *parent selection*. The creation

978-83-60810-66-8/$25.00©2015, IEEE

of the offspring is based on recombination and mutation. Recombination combines traits from two or more parents resulting in one or more intermediate offspring. In contrast, mutation is an unary operator changing the components of an individual randomly. After the offspring have been created, survivor selection is performed to determine the next parent population. The different variants of evolutionary algorithms adhere to the same principles in general, but they may differ in the representation of the solutions and how the selection, recombination, and mutation processes are realized.

*a) Evolution Strategies:* Evolution strategies (ESs) [7], [8] are used for continuous optimization $f : \mathbb{R}^N \rightarrow \mathbb{R}$. Several variants have been introduced see e.g. [9], [1]. In many cases, a population of $\mu$ parents is used to create a set of $\lambda$ offspring, with $\mu \leq \lambda$. For recombination, $\rho$ parents are chosen uniformly at random without replacement and are then recombined. Recombination usually consists of determining the (weighted) mean or centroid of the parents [9]. The result is then mutated by adding a normally distributed random variable with zero mean and covariance matrix $\sigma^2\mathbf{C}$. While there are ESs that operate without recombination, the mutation process is seen as the essential process. It is often interpreted as the main search operator. After the offspring have been created, the individuals are evaluated using the function to be optimized or a derived function which allows an easy ranking of the population. Only the rank of an individual is important for the selection. In the case of continuous optimization, the old parent population is typically discarded with the selection considering only the $\lambda$ offspring of which the $\mu$ best are chosen.

The covariance matrix which is central to the mutation must be adapted during the run: Evolution strategies with ill-adapted parameters converge only slowly or may even fail in the optimization. Therefore, research on methods for adapting the scale factor $\sigma$ or the full covariance matrix has a long research tradition in ESs dating back to their origins [7]. The next section describes one of the current approaches.

*b) Updating the Covariance Matrix:* To our knowledge, covariance matrix adaptation comprises two main classes: one applied in the *covariance matrix adaptation evolution strategy* (CMA-ES) [10] and an alternative used in the *covariance matrix self-adaptation evolution strategy* (CMSA-ES) [11]. Both consider information from the present population combining it with information from the search process so far. The CMA-ES is one of the most powerful evolution strategies and often referred to as the standard in ESs. However, as pointed out in [11], its scaling behavior with the population size may not be good. Beyer and Sendhoff [11] showed that the CMSA-ES performs comparably to the CMA-ES for smaller populations but that is less computational expensive for larger population sizes.

Therefore, the present paper focuses on the CMSA-ES leaving the CMA-ES for future research. The CMSA-ES uses weighted intermediate recombination, in other words, the weighted centroid $\mathbf{m}^{(g)}$ of the $\mu$ best individuals of the population is computed. To create the offspring, random vectors are drawn from the multivariate normal distribution $\vec{\mathcal{N}}(\mathbf{m}^{(g)}, (\sigma^{(g)})^2\mathbf{C}^{(g)})$. The notation of covariance matrix as $(\sigma^{(g)})^2\mathbf{C}^{(g)}$ illustrates that the actual covariance matrix is interpreted as the combination of a general scaling factor (or step-size or mutation strength) with a rotation matrix. Following the usual practice in literature on evolution strategies the latter matrix $\mathbf{C}^{(g)}$ is referred to as *covariance matrix* in the remainder of the paper.

The covariance matrix update is based upon the common estimate of the covariance using the newly created population. Instead of considering all offspring for deriving the estimates, though, it introduces a bias towards good search regions by taking only the $\mu$ best individuals into account. Furthermore, it does not estimate the mean anew but uses the weighted mean $\mathbf{m}^{(g)}$. Following [10],

$$\mathbf{z}_{m:\lambda}^{(g+1)} := \frac{1}{\sigma^{(g)}}\left(\mathbf{x}_{m:\lambda}^{(g+1)} - \mathbf{m}^{(g)}\right) \tag{1}$$

are determined with $\mathbf{x}_{m:\lambda}$ denoting the $m$th best of the $\lambda$ particle according to the fitness ranking. The rank-$\mu$ update then obtains the covariance matrix as

$$\mathbf{C}_{\mu}^{(g+1)} \;:=\; \sum_{m=1}^{\mu} w_m \mathbf{z}_{m:\lambda}^{(g+1)} (\mathbf{z}_{m:\lambda}^{(g+1)})^{\mathrm{T}} \tag{2}$$

which is usually a positive semi-definite matrix since $\mu \ll N$. The weights $w_m$ should fulfill $w_1 \geq w_2 \geq \ldots \geq w_\mu$ with $\sum_{m=1}^{\mu} w_i = 1$. To derive reliable estimates larger population sizes are required which would lower the algorithm's speed. Therefore, past covariance matrices are taken into account via the convex combination of (2) with the sample covariance being shrunk towards the old covariance

$$\mathbf{C}^{(g+1)} \;:=\; (1 - \frac{1}{c_\tau})\mathbf{C}^{(g)} + \frac{1}{c_\tau}\mathbf{C}_{\mu}^{(g+1)} \tag{3}$$

with the weights usually set to $w_m = 1/\mu$ and

$$c_\tau = 1 + \frac{N(N+1)}{2\mu}, \tag{4}$$

see [11]. As long as $\mathbf{C}^{(g)}$ is positive semi-definite, (3) will result in a positive definite matrix.

*c) Step-Size Adaptation:* The CMSA implements the step-size using *self-adaptation* first introduced in [7] and developed further in [8]. Here, evolution is used for fitting the strategy parameters of the mutation process. In other words, the scaling parameter or in its full form, the complete covariance matrix, undergoes recombination, mutation, and indirect selection processes. The working principle is based on an indirect stochastic linkage between good individuals and appropriate parameters: Well adapted parameters should result more often in better offspring than too large or too small values or misleading directions. Although self-adaptation has been developed to adapt the whole covariance matrix, it is applied today mainly to adapt the step-size or a diagonal covariance matrix. In the case of the mutation strength, usually a log-normal distribution

$$\sigma_l^{(g)} \;=\; \sigma_{\text{base}}\exp(\tau\mathcal{N}(0,1)) \tag{5}$$

is used for mutation. The parameter $\tau$ is called the *learning rate* and is usually chosen to scale with $1/\sqrt{2N}$. The baseline $\sigma_{\text{base}}$ is either the mutation strength of the parent or if recombination is used the recombination result. For the step-size, it is possible to apply the same type of recombination as for the positions although different forms – for instance a multiplicative combination – could be used instead. The self-adaptation of the step-size is referred to as *$\sigma$-self-adaptation* ($\sigma$SA) in the remainder of this paper.

The newly created mutation strength is then directly used in the mutation of the offspring. If the resulting offspring is sufficiently good, the scale factor is passed to the next generation.

Self-adaptation with recombination has been shown to be "robust" against noise [12] and is used in the CMSA-ES as the update rule for the scaling factor.

### B. Concerning the Covariance Matrix Adaptation ...

In the case of $\lambda > 1$, the sample covariance (2) appears in nearly any adaptation process. Disregarding the distortion due to selection, the sample covariance as the maximum likelihood estimator of the true covariance matrix is known as a good and reliable estimate if $\mu \gg N$. Evolution strategies typically operate with $\mu < N$, however. For example, following [13] the sizes of the parent and offspring populations in the standard CMA-ES should be chosen as $\lambda = \lfloor \log(3N) \rfloor + 4$ and $\mu = \lfloor \lambda/2 \rfloor$.

Unfortunately, $\mu < N$ leads to problems with respect to the covariance estimation. This is a well-known problem in statistics [14], [15], giving raise to a broad range on literature on alternative estimators e.g. [15], [16], [17], [18], [19], [20], [21], [22], [23]. The quality of a maximum likelihood estimate may be insufficient – especially for high-dimensional spaces, see e.g. [16]. For example, Marčenko and Pastur showed that if $N/\mu \not\to 0$ but $N/\mu \in (0,1)$, instead, the eigenvalues of the covariance matrix are distributed in the interval $((1-\sqrt{N/\mu})^2, (1+\sqrt{N/\mu})^2)$ in the case of the standard normal distribution [17].

Equation (3) actually attempts to counteract the singularity of the population covariance matrix by using the well-known concept of shrinking. However, some distinctive differences are present. First of all, the target is a full covariance matrix whereas shrinkage typically considers simpler regulation forms as e.g. a diagonal matrix. Secondly, the parameter is usually determined via optimizing a performance measure.

Seeing that evolution strategies already apply some kind of shrinkage, some questions arise: Can we improve the estimator further by not only "shrinking" the population or sample covariance matrix but by applying further concepts stemming from the estimation of high-dimensional covariance matrices? And considering that (3) is one regulation technique among several, is it possible to find another well-performing substitute? Or did research in evolution strategies already happen upon the best technique possible?

## II. A Sparse Covariance Matrix Adaptation

This section introduces the new covariance adaptation technique which uses thresholding to transform the population covariance matrix. The decision for thresholding is based upon the comparatively computational efficiency of the approach.

### A. Space Transformation

The ideal covariance matrix for the search depends on the function landscape which is unknown in practical applications. Considering the smooth test functions of typical black-box optimization suites, shows that the Hessians of several functions, as e.g. the separable functions, can be classified as sparse or approximately sparse matrices following the definitions introduced later.

Therefore, sparse structures of the covariance matrix suffice which is exemplified by the separable CMA-ES [24] which restricts the covariance to a diagonal matrix in case of separability to allow fast progress to the optimal solution. For the general case, a spare structure may not be suitable, however.

For this reason, the paper does not require sparseness of the original covariance matrix, although it would be interesting to see how such a variant would perform on the test suites. Instead, it considers a transformation. As argued in [25], an change of the coordinate system may improve the performance of an evolution strategy. Therefore, an adaptive encoding was introduced. In each iteration, the covariance matrix is adapted following the rules of the CMA-ES. Its spectral decomposition is used to change the basis. The creation of new search points is carried out in the eigenspace of the current covariance matrix and the main search parameters of the CMA-ES are updated there. After selection, the covariance matrix is adapted and utilized for a renewed decoding and encoding.

This paper also addresses a change of the coordination system. However, we address the covariance matrix adaptation and estimation itself which in [25] occurs in the original space. Here, we argue that a switch to the eigenspace of the old covariance matrix $\mathbf{C}^{(g)}$ may be beneficial for the estimation of the covariance matrix itself.

Let the covariance matrix $\mathbf{C}^{(g)}$ be a symmetric, positive definite $N \times N$ matrix. The condition holds for the original adaptation since (3) combines a positive definite with a positive semi-definite matrix. As we will see below, in the case of thresholding the condition may not always be fulfilled. Assuming a positive definite matrix allows carrying out a spectral decomposition: Let $\mathbf{v}_1, \dots, \mathbf{v}_N$ denote the $N$ eigenvectors with the eigenvalues $\lambda_1, \dots, \lambda_N$, $\lambda_j > 0$. Note, the eigenvectors form a orthonormal basis of $\mathbb{R}^N$, i.e., $\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i = 1$ and $\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j = 0$, if $i \neq j$. We define $\mathbf{V} := (\mathbf{v}_1, \dots, \mathbf{v}_N)$ as the modal matrix. It then holds that $\mathbf{V}^{-1} = \mathbf{V}^{\mathrm{T}}$. Switching to the eigenspace of $\mathbf{C}^{(g)}$ results in the representation of the covariance matrix

$$\Lambda^{(g)} = \mathbf{V}\mathbf{C}^{(g)}\mathbf{V}^{\mathrm{T}} \tag{6}$$

as a diagonal matrix with the eigenvalues as the diagonal entries. Diagonal matrices are sparse matrices, thus for the estimation of the covariance matrix the more efficient procedures for sparse structures could be used. However, it is not the goal to re-estimate $\mathbf{C}^{(g)}$ but to estimate the true covariance matrix of the distribution indicated by the sample $\mathbf{z}_{1;\lambda},\ldots,\mathbf{z}_{\mu;\lambda}$.

Before continuing, it should be noted that several definitions of sparseness exist. Usually, it is demanded that the number of non-zero elements in a row may not exceed a predefined limit $s_0(N) > 0$, i.e.,

$$\max_i \sum_{j=1}^{N} \delta(|a_{ij}| > 0) \leq s_0(N), \tag{7}$$

which should grow only slowly with $N$. The indicator function $\delta$ fulfills $\delta(\cdot) = 1$ if the condition is met and is zero otherwise. This definition can, however, be relaxed to a more general definition of sparseness, also referred to as approximate sparseness. Cai and Liu [22] consider the following uniformity class of sparse matrices

**Definition 1.** *Let $s_0(N) > 0$ and let $\cdot \succ 0$ denote positive definiteness. Then a class of sparse covariance matrices is defined as*

$$\begin{aligned}\mathcal{U}_q^* :=& \mathcal{U}_q^*(s_0(N)) \\ =& \left\{\Sigma : \Sigma \succ 0, \max_i \sum_{j=1}^{p} (\sigma_{ii}\sigma_{jj})^{\frac{(1-q)}{2}} |\sigma_{ij}|^q \leq s_0(N)\right\}\end{aligned} \tag{8}$$

*for some $0 \leq q < 1$.*

Definition 1 requires the entries of the covariance matrix to lie within a weighted $l_q$ ball. The weight is given by the variances. Cai and Liu [22] introduce a thresholding estimator that requires the assumption above. Its convergence rate towards the true covariance depends on $s_0(N)(\log(N)/\mu)^{(1-q)/2}$. Therefore, the number $s_0(N) > 0$ should again grow only "slowly" for $N \to \infty$.

Definition 1 leads to the main assumption of the paper. Consider an evolution strategy in the search space. The new sample that is the offspring population has been created with the help of the old covariance matrix. The covariance matrix of the selected sample differs from the previous. The deviations of from its structure stem from finite sampling characteristics and rank-based selection. Assuming that the form of the covariance matrix will not change considerably in one iteration, the new underlying covariance matrix should be sparse in the eigenspace of the old covariance, however.

**Assumption 1.** *Let $\Sigma^{(g+1)}$ denote the true covariance matrix of the selected offspring. Consider the old covariance $\mathbf{C}^{(g)}$ with its modal matrix $\mathbf{V}$. Then $\hat{\Lambda} = \mathbf{V}\Sigma^{(g+1)}\mathbf{V}^{\mathrm{T}}$ is approximately sparse, i. e. $\hat{\Lambda} \in \mathcal{U}_q^*$ for some $0 \leq q < 1$.*

Assuming the validity of the assumption, we change the coordinate system in order to perform the covariance matrix estimate. Reconsider the normalized (apart from the covariance matrix) mutation vectors $\mathbf{z}_{1;\lambda},\ldots,\mathbf{z}_{\mu;\lambda}$ that were associated with the $\mu$ best offspring. Their representation in the eigenspace reads

$$\hat{\mathbf{z}}_{m;\lambda} \quad = \quad \mathbf{V}^{\mathrm{T}}\mathbf{z}_{m;\lambda} \text{ for } m = 1,\ldots,\mu. \tag{9}$$

The transformed population covariance is then estimated as

$$\hat{\mathbf{C}}_\mu = \sum_{m=1}^{\mu} w_m \hat{\mathbf{z}}_{m;\lambda}\hat{\mathbf{z}}_{m;\lambda}^{\mathrm{T}}. \tag{10}$$

The estimate (10) will be used to compute the final estimator. In the next section, we discuss potential estimators for sparse covariance matrices.

### B. Sparse Covariance Matrix Estimation

In recent years, covariance matrix estimation in high-dimensional spaces has received a lot of attention. In the case of sparse covariance matrices, banding, tapering, and thresholding can be applied, see e.g. [26] All three make use of the fact that many entries of the matrix that shall be estimated are actually zero or at least very small. Banding and tapering differ from thresholding in that they assume a specific matrix structure in other words they assume an ordering of the variables which is for instance often the case in time-series analysis. Banding and tapering approaches typically lead to consistent estimators if $\log(N)/\mu \to 0$.

Thresholding does not assume a natural order of the variables. Instead, it discards entries which are smaller than a given threshold $\epsilon > 0$. For a matrix $\mathbf{A}$, the thresholding operator $T_\epsilon(\mathbf{A})$ is defined as

$$T_\epsilon(\mathbf{A}) := (a_{ij}\delta(|a_{ij}| \geq \epsilon))_{N\times N}. \tag{11}$$

The choice of the threshold is critical for the quality of the resulting estimate.

Equation (11) represents a example of universal thresholding with a hard thresholding function. Equation (11) can be extended in several ways. On the one hand, the threshold may depend on the entry itself, and on the other hand, instead of the hard threshold applied, a generalized shrinkage function $s_\lambda(\cdot)$ can be used. Following [22] , the function $s_\lambda(\cdot)$ should have the following properties

i) $\exists c > 0$: $s_\lambda(x) \leq c|y|$ $\forall\, x, y$ which satisfy $|x - y| \leq \lambda$,
ii) $s_\lambda(x) = 0$ $\forall x \leq \lambda$,
iii) $|s_\lambda(x) - x| \leq \lambda$ $\forall x \in \mathbb{R}$.

Several functions have been introduced that fulfill i)-iii), as e.g. the soft-thresholding

$$s_\lambda(x) = \operatorname{sign}(x)(|x| - \lambda)_+ \tag{12}$$

or the Lasso

$$s_\lambda(x) = |x|(1 - |\frac{\lambda}{x}|^\eta)_+ \tag{13}$$

with $(x)_+ := \max(0, x)$. In this paper, the threshold $\lambda_{ij}$ is defined component-wise and not universal. Since its correct choice is difficult to decide a priori, adaptive thresholding is applied as in [22], setting

$$\lambda_{ij} := \lambda_{ij}(\delta) = \delta\sqrt{\frac{\hat{\theta}_{ij}\log N}{\mu}} \tag{14}$$

with $\delta > 0$ can be either chosen as a constant or adapted data driven. The variable $\hat{\theta}_{ij}$ that appears in (14) is obtained as

$$\hat{\theta}_{ij} = \frac{1}{\mu}\sum_{m=1}^{\mu} [(\hat{z}_{mi} - \overline{Z^i})((\hat{z}_{mj} - \overline{Z^j}) - \hat{c}_{ij}^{\mu}]^2 \tag{15}$$

```
Require: $\lambda$, $\mu$, $\mathbf{C}^{(0)}$, $\mathbf{m}^{(0)}$, $\sigma^{(0)}$, $\tau$, $c_\tau$
 1: $g = 0$
 2: while termination criteria not met do
 3:     for $l = 1$ to $\lambda$ do
 4:         $\sigma_l = \sigma^{(g)} \exp(\tau \mathcal{N}(0,1))$
 5:         $\mathbf{x}_l = \mathbf{m}^{(g)} + \sigma_l \mathcal{N}(0, \mathbf{C}^{(g)})$
 6:         $f_l = f(\mathbf{x}_l)$
 7:     end for
 8:     Select $(\mathbf{x}_{1:\lambda}, \sigma_{1:\lambda}), \dots, (\mathbf{x}_{\mu:\lambda}, \sigma_{\mu:\lambda})$
 9:     $\mathbf{m}^{(g+1)} = \sum_{m=1}^{\mu} w_m \mathbf{x}_{m:\lambda}$
10:     $\sigma^{(g+1)} = \sum_{m=1}^{\mu} w_m \sigma_{m:\lambda}$
11:     $\mathbf{z}_{m;\lambda} = \frac{\mathbf{x}_{m;\lambda} - \mathbf{m}^{(g)}}{\sigma^{(g)}}$ for $m = 1, \dots, \mu$
12:     $\mathbf{V}, \mathbf{D} \leftarrow \mathrm{spectral}(\mathbf{C}^{(g)})$
13:     $\hat{\mathbf{z}}_{m;\lambda} = \mathbf{V}^{\mathrm{T}} \mathbf{z}_{m;\lambda}$ for $m = 1, \dots, \mu$
14:     $\hat{\mathbf{C}}_\mu = \sum_{m=1}^{\mu} w_m \hat{\mathbf{z}}_{m;\lambda} \hat{\mathbf{z}}_{m;\lambda}^{\mathrm{T}}$
15:     $\hat{\mathbf{C}}_{\mathrm{thres}} = T_{S_{\lambda_{ij}}}(\hat{\mathbf{C}}_\mu)$
16:     $\mathbf{C}_\mu = \mathbf{V}^{\mathrm{T}} \hat{\mathbf{C}}_{\mathrm{thres}} \mathbf{V}$
17:     $\mathbf{C}^{(g+1)} = (1 - \frac{1}{c_\tau}) \mathbf{C}^{(g)} + \frac{1}{c_\tau} \mathbf{C}_\mu$
18:     $g = g + 1$
19: end while
```

Fig. 1. The CMSA-ES with thresholding. The generation counter $g$ is sometimes left out in order to simplify the notation. The symbol *spectral* stands for the spectral decomposition of the matrix into the modal matrix $\mathbf{V}$ and the diagonal matrix containing the eigenvalues $\mathbf{D}$. Rank-based deterministic selection of the $\mu$ best offspring is performed in line 8 based on the fitness $f$.

with $\hat{c}_{ij}^{\mu}$ denoting the $(i,j)$-entry of $\hat{\mathbf{C}}_\mu^{(g+1)}$, $\hat{z}_{mi}$ the $i$th component of $\hat{\mathbf{z}}_{m:\lambda}$, and $\overline{Z^i} := (1/\mu) \sum_{m=1}^{\mu} \hat{z}_{mi}$. Other thresholds have been introduced, see e.g. [27] and will be considered in future work.

While thresholding respects symmetry and non-negativeness properties, it results only in asymptotically positive definite matrices. Thus, for finite sample sizes, it does neither preserve nor induce positive definiteness in general. This holds for hard thresholding as well as for most cases of potential thresholding functions. As shown in [28], a positive semi-definiteness can only be guaranteed for a small class of functions for general matrices. In the case that the condition number of the matrix is sufficiently small, the group of functions that preserve positive definiteness can be widened to include also polynomials. In [27], procedures are discussed that result in positive definite matrices. As this paper aims for a proof of concept, it does not consider repair mechanisms.

### C. Evolution Strategies with Sparse Covariance Adaptation

Component-wise adaptive thresholding can be integrated readily into evolution strategies. Figure 1 illustrates the main points of the algorithm. There are several ways to design the operator $T_{S_{\lambda_{ij}}}$. The first choice concerns the thresholding function $s_{\lambda_{ij}}(\cdot)$. The second question concerns whether thresholding should be applied to all entries of the covariance matrix (11) or only to the off-diagonal elements. This question is difficult to decide beforehand in the application context considered. Therefore, two variants are investigated

1) CMSA-Thres-ES (abbreviated to Thres): An evolution strategy with CMSA which applies thresholding in the eigenspace of the covariance, using the operator

$$T_{S_{\lambda_{ij}}}(\mathbf{A})_{ij} = s_{\lambda_{ij}}(a_{ij}) \qquad (16)$$

and

2) CMSA-Diag-ES (abbreviated to Diag): An ES with covariance matrix adaptation which uses thresholding in the eigenspace of the covariance and excepts the diagonal elements with

$$T_{S_{\lambda_{ij}}}(\mathbf{A})_{ij} = \begin{cases} a_{ij} & \text{if } i = j \\ s_{\lambda_{ij}}(a_{ij}) & \text{if } i \neq j \end{cases}. \qquad (17)$$

In statistics, thresholding is often applied only to the off-diagonal entries. Keeping the diagonal unchanged may however result in a too strong reliance on the structure of the old covariance matrix in our case. This may make a change of the search directions difficult. Therefore, both variants are taken into account.

## III. Experiments

The experiments are performed for the search space dimensions $N = 10$ and 20. Since we aim for a general approach, the performance of the new techniques should also be analyzed for lower dimensional spaces. The maximal number of fitness evaluations is set to $\mathrm{FE}_{\max} = 2 \times 10^5 N$. The start position of the algorithms is randomly chosen from $[-4, 4]^N$. The population size were chosen as $\lambda = \lfloor \log(3N) + 8 \rfloor$ and $\mu = \lceil \lambda/2 \rceil$. The weights $w_m$ were set to $w_m = 1/\mu$.

A run terminates before reaching the maximal number of evaluations, if the difference between the best value obtained so far and the optimal fitness value $|f_{\mathrm{best}} - f_{\mathrm{opt}}|$ is below a predefined target precision set to $10^{-8}$. For each fitness function and dimension, 15 runs are used in accordance to the practice of the black box optimization workshops, see below. If the search stagnates, indicated by changes of the best values being below $10^{-8}$ for $10 + \lceil 30N/\lambda) \rceil$ generations, the ES is restarted. The Lasso thresholding function (13) with $\eta = 4$ was chosen as the thresholding function and by performing a preliminary series of experiments the scaling factor $\delta$ in (15) was set to $\delta = 2\max(\hat{\mathbf{C}}_\mu)$. Both choices can be probably improved. Since the paper strives for a first proof of concept, a detailed investigation of good parameter settings will be performed in future research.

### A. Test Suite

For the experiments, the algorithms were implemented in MATLAB. The paper uses black box optimization benchmarking (BBOB) software framework and the test suite introduced for the black box optimization workshops, see [29]. The goal of the workshop is to benchmark and to compare metaheuristics and other direct search methods for continuous optimization. The framework[1] allows the plug-in of algorithms adhering to a common interface and provides a comfortable way of generating the results in form of tables and figures.

[1] Latest version under http://coco.gforge.inria.fr

| | |
|---|---|
| Sphere | $f(\mathbf{x}) = \|\mathbf{z}\|^2$ |
| Rosenbrock | $f(\mathbf{x}) = \sum_{i=1}^{N-1} 200(z_i^2 - z_{i+1})^2 + (z_i - 1)^2$ |
| Ellipsoidal | $f(\mathbf{x}) = \sum_{i=1}^{N} 10^{6\frac{i-1}{N-1}} z_i^2$ |
| Discus | $f(\mathbf{x}) = 10^6 z_1^2 + \sum_{i=2}^{N} z_i^2$ |
| Rastrigin | $f(\mathbf{x}) = 10\big(N - \sum_{i=1}^{N} \cos(2\pi z_i)\big) + \|\mathbf{z}\|^2$ |

TABLE I
SOME OF THE TEST FUNCTIONS USED FOR THE COMPARISON OF THE ALGORITHMS.

The test suite contains noisy and noise-less functions with the position of the optimum changing randomly from run to run. This paper focuses on the noise-less test suite which contains 24 functions [30]. They can be divided into four classes: separable functions (function ids 1-5), functions with low/moderate conditioning (ids 6-9), functions with high conditioning (ids 10-14), and two groups of multimodal functions (ids 15-24). Among the unimodal functions with only one optimal point, there are separable functions given by the general formula

$$f(\mathbf{x}) = \sum_{i=1}^{N} f_i(x_i) \qquad (18)$$

which can be solved by optimizing each component separately. The simplest member of this class is the (quadratic) sphere with $f(\mathbf{x}) = \|\mathbf{x}\|^2$. Other functions include ill-conditioned functions, like for instance the elliposoidal function, and multimodal functions (Rastrigin) which represent particular challenges for the optimization (Table I). The variable $\mathbf{z}$ denotes a transformation of $\mathbf{x}$ in order to keep the algorithm from exploiting certain particularities of the function, see [30].

### *B. Performance Measure*

The following performance measure is used in accordance to [29]. The expected running time (ERT) gives the expected value of the function evaluations ($f$-evaluations) the algorithm needs to reach the target value with the required precision for the first time, see [29]. In this paper, we use

$$\text{ERT} = \frac{\#(FEs(f_{\text{best}} \geq f_{\text{target}}))}{\#succ} \qquad (19)$$

as an estimate by summing up the fitness evaluations $FEs(f_{\text{best}} \geq f_{\text{target}})$ of each run until the fitness of the best individual is smaller than the target value, divided by all successfull runs.

### *C. Results and Discussion*

The findings are interesting – indicating advantages for thresholding in many but not in all cases. The result of the comparison depends on the function class. In the case of the separable functions with ids 1-5, the strategies behave on the whole very similar in the case of both dimensionalities 10D and 20D. This can be seen in the empirical cumulative distribution functions plots in Fig. 2 and Fig. 3 for example.

Concerning the particular functions, differences are revealed as Tab. II and Tab. III show for the expected running time (ERT) which is provided for several precision targets. The expected running time is provided relative to the best results achieved during the black-box optimization workshop in 2009. The first line of the outcomes for each function reports the ERT of the best algorithm of 2009. However, not only the ERT values but also the number of successes is important. The ERT can only be measured if the algorithm achieved the respective target in the run. If the number of trials where is the full optimization objective has been reached is low then the remaining targets should be discussed with care. If only a few runs contribute to the result, the findings may be strongly influenced by initialization effects. To summarize, only a few cases end with differences that are statistically significant. To achieve this, the algorithm has to perform significantly better than both competing methods – the other thresholding variant and the original CMSA-ES.

In the case of the sphere (function with id 1), slight advantages for the thresholding variants are revealed. A similar observation can be made for the second function, the separable ellipsoid. Here, both thresholded ESs are faster, with the one that only shrinks the off-diagonal elements significantly (Tab. III). This is probably due to the enforced more regular structure.

No strategy is able to reach the required target precision in the case of the separable Rastrigin (id 3) and the separable Rastrigin-Bueche (id 4). Since all strategies only achieve the lowest target precision of $10^1$, a comparison is not performed. The linear slope is solved fast by all, with the original CMSA-ES the best strategy.

In the case of the function class containing test functions with low to moderate conditioning, different findings can be made for the two search space dimensionalities. This is also shown by the empirical cumulative distribution functions plots in Fig. 2 and Fig. 3, especially for $N = 10$. Also in the case of $N = 10$, Table II shows that the strategies with thresholding achieve a better performance in a majority of cases. In addition, thresholding that is not applied to the diagonal appears to lead to a well-performing strategy with the exception of f9, the rotated Rosenbrock function, where it lead to the largest expected running times.

The results for f6, the so-called attractive sector, in 10D are astonishing. While the original CMSA-ES could only reach the required target precision in six of the 15 runs, the thresholding variants were able to succeed 14 times (CMSA-Thres-ES) and 13 times (CMSA-Diag-ES). The latter achieved lower expected running times, though. This does not transfer to 20D. Here, only a minority of runs were successfull for all strategies. Experiments with a larger number of fitness evaluations must be conducted in order to investigate the findings more closely.

The same holds for the step ellipsoid (id 7) which cannot be solved with the target precision required by any strategy. Concerning the lower precision targets, sometimes the CMSA-ES and sometimes the CMSA-Diag-ES appears superior. However, more research is required, since the number of runs entering the data for some of the target precisions is low and

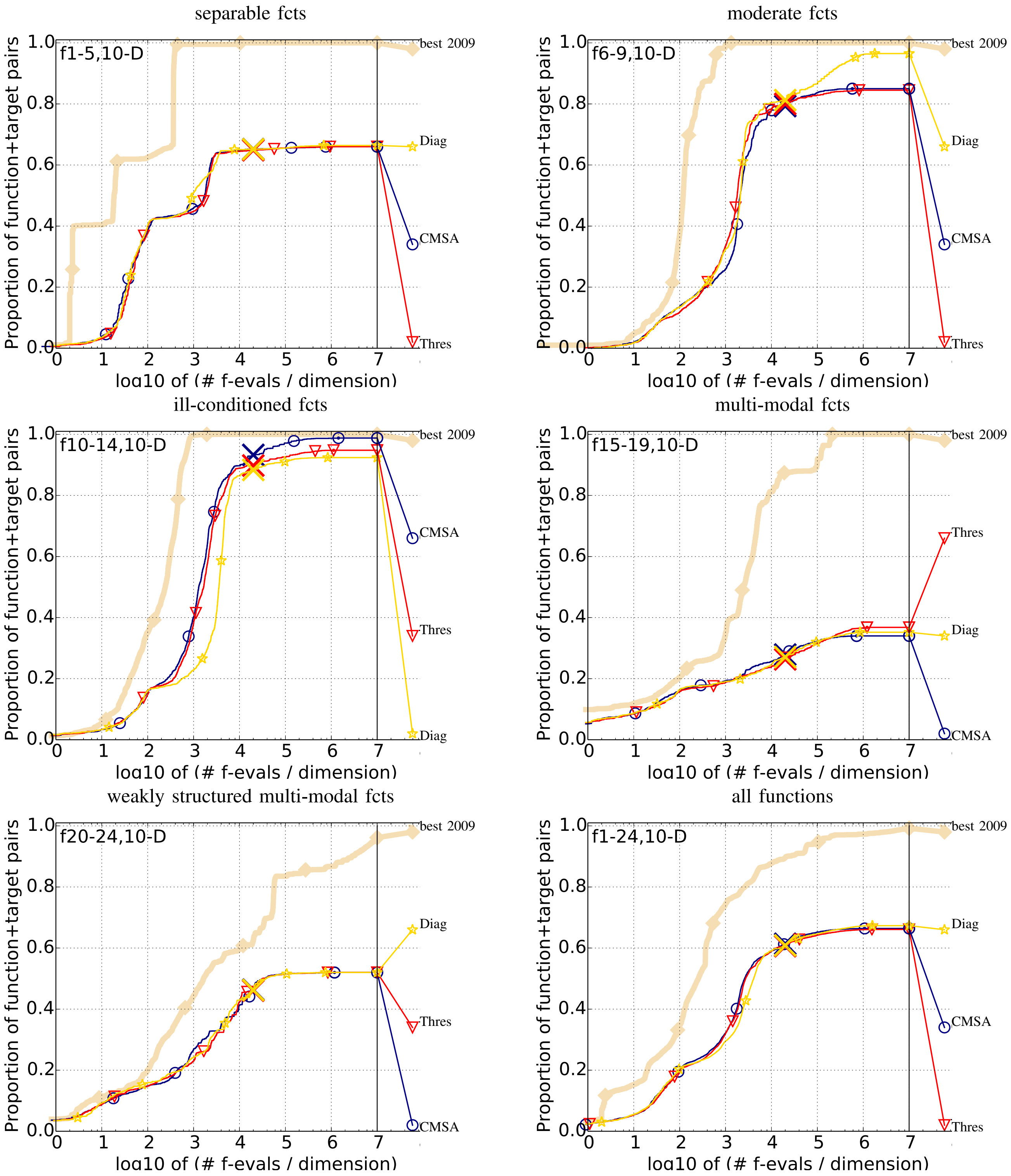


Fig. 2. Bootstrapped empirical cumulative distribution of the number of objective function evaluations divided by dimension (FEvals/DIM) for 50 targets in $10^{[-8..2]}$ for all functions and subgroups in 10-D. The "best 2009" line corresponds to the best ERT observed during BBOB 2009 for each single target.

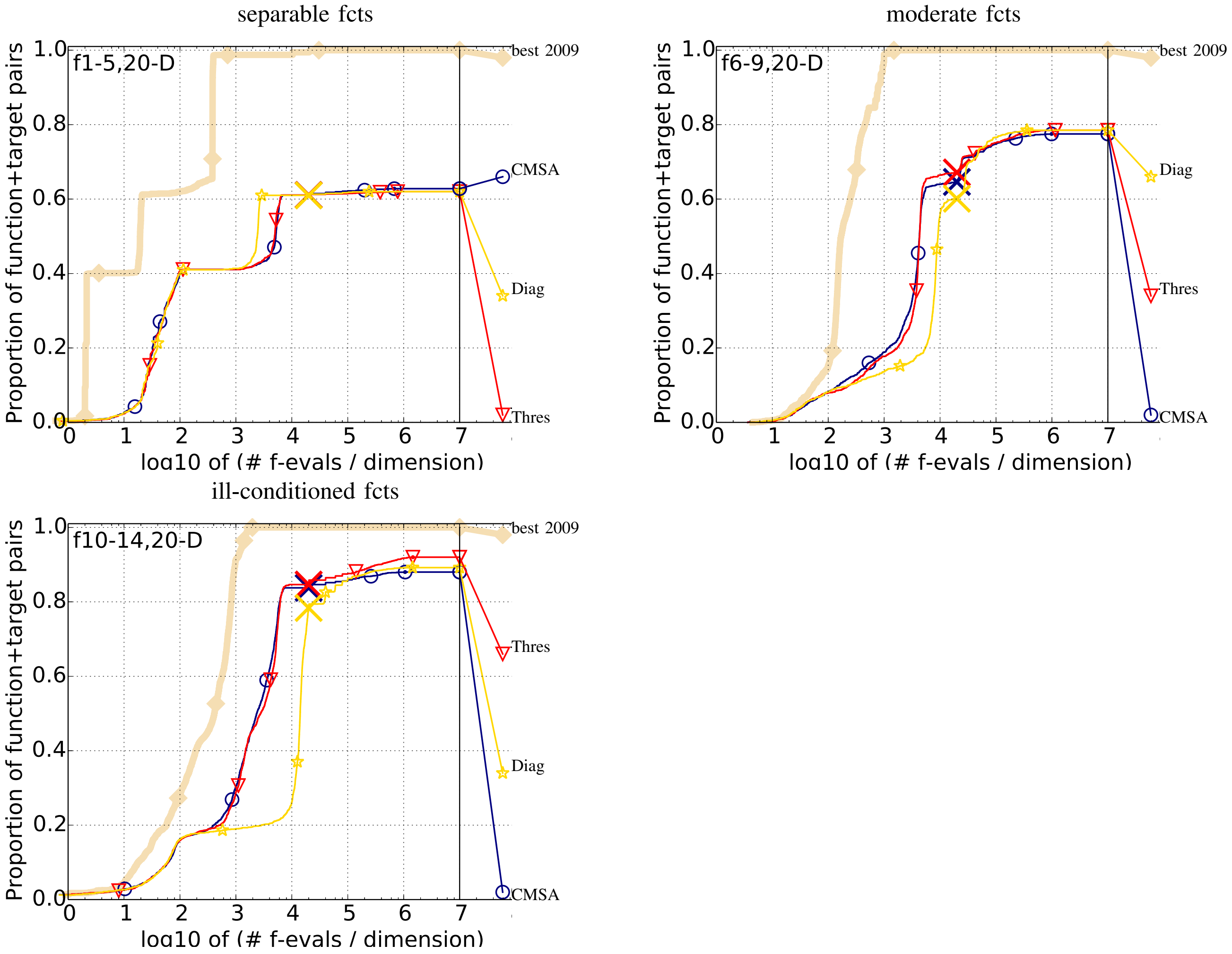


Fig. 3. Bootstrapped empirical cumulative distribution of the number of objective function evaluations divided by dimension (FEvals/DIM) for 50 targets in $10^{[-8..2]}$ for all functions and subgroups in 20-D. The "best 2009" line corresponds to the best ERT observed during BBOB 2009 for each single target.

initial positions may be influential.

On the original Rosenbrock function (id 8), the CMSA-ES and the CMSA with thresholding show a similar behavior with the CMSA-ES performing better. In contrast, the thresholding variant that leaves the diagonal unchanged exhibits larger expected running times. The roles of the original CMSA-ES and the CMSA-Thres-ES reverse for the rotated Rosenbrock (id 9). Here, the best results can be observed for the thresholding variant. Again, the CMSA-Diag-ES performs worst.

In the case of ill-conditioned functions, the findings are mixed. In general, thresholding without including the diagonal does not appear to improve the performance. The strategy performs worst of all – an indicator that keeping the diagonal unchanged may be sometimes inappropriate due to the space transformation. However, since there are interactions with the choice of the thresholding parameters which may have resulted in comparatively too large diagonal elements, we need to address this issue further before coming to a conclusion. First of all for $N = 10$, all strategies are successfull in all cases for the ellipsoid (id 10), the discus (id 11), the bent cigar (id 12), and the sum of different powers (id 14). Only the CMSA-ES reaches the optimization target in the case of the sharp ridge (id 13). This, however, only twice. The reasons for this require further analysis. Either the findings may be due to a violation of the sparseness assumption or considering that this is only a weak assumption the choice of the thresholding parameters and the function should be reconsidered.

All strategies exhibit problems in the case of the group of multi-modal functions, Rastrigin (id 15), Weierstrass (id 16), Schaffer F7 with condition number 10 (id 17), Schaffer F7 with condition 1000 (id 18), and Griewank-Rosenbrock F8F2 (id 19). Partly, this may be due to the maximal number of fitness evaluations permitted. Even the best performing methods of the 2009 BBOB workshop required more evaluations than we allowed in total. Thus, experiments with larger values for the maximal function evaluations should be conducted in future research. Concerning the preliminary targets with lower precision, the CMSA-ES achieves the best results in a majority of cases. However, the same argumentation as for the step ellipsoid applies.

In the case of $N$ = 20, the number of function evaluations that were necessary in the case of the best algorithms of 2009 to reach even the lower precision target of $10^{-1}$ exceeds the budget chosen here. Therefore, the function group is excluded from the analysis for $N$ = 20 and not shown in Figure 3 and Table III.

The remaining group consists of multi-modal functions with weak global structures. Here, especially the functions with numbers 20 (Schwefel $x \sin(x)$), 23 (Kaatsuuras), and 24 (Lunacek bi-Rastrigin) represent challenges for the algorithms. In the case of $N$ = 10, they can only reach the first targets of $10^1$ and $10^0$. Again, the maximal number of function evaluations should be increased to allow a more detailed analysis on these functions. For the case of the remaining functions, function 21, Gallagher 101 peaks, and function 22, Gallagher 21 peaks, the results indicate a better performance for the CMSA-ES versions with thresholding compared with the original algorithm. Again due to similar reasons as for the first group of multi-modal functions, the results are only shown for $N$ = 10.

## IV. Conclusions and Outlook

This paper adressed covariance matrix adaptation techniques for evolution strategies. The original versions are based on the sample covariance – an estimator known to be problematic. Especially in high-dimensional search spaces, where the population size does not exceed the search space dimensionality, the agreement of the estimator and the true covariance may be low. Therefore, thresholding, a comparably computationally simple estimation technique, has been integrated into the covariance adaptation process. Thresholding stems from estimation theory for high-dimensional spaces and assumes an approximately sparse structure of the covariance matrix. The matrix entries are therefore thresholded, meaning a thresholding function is applied. The paper considered adaptive entry-wise thresholding. Since the covariance matrix cannot be assumed to be sparse in general, a basis transformation was carried out and the thresholding process was performed in the transformed space. The performance of the resulting new covariance matrix adapting evolution strategies was compared to the original variant on the black-box optimization benchmarking test suite. Two main variants were considered: A CMSA-ES which subjected the complete covariance to thresholding and a variant which left the diagonal elements unchanged. While the latter is more common in statistics, it is not easy to justify its preferation in optimization. The first findings were interesting with the new variants performing better for several function classes. While this is promising, more experiments and analyses are required and will be performed in future research. This concerns e.g. which variant to use since it depended on the function which of the two performed best. Open questions concern among others the choice of the thresholding function and the scaling parameter for the threshold. In this paper, it was selected by a small series of experiments. Making the parameter completely data driven and thus depending on the current sample is the goal of ongoing research.

If the assumption that the representation of true covariance $\Sigma^{(g+1)}$ of the offspring population in the eigenspace of the previous covariance $C^{(g)}$ is approximately sparse should be violated in some cases, then it may be worthwhile to take a closer look at the convex combination of the new and the old covariance matrix. Further work will thus also consider applying thresholding to the traditionally obtained covariance.

## References

[1] T. Bäck, C. Foussette, and P. Krause, *Contemporary Evolution Strategies*, ser. Natural Computing. Springer, 2013.

[2] W. Dong and X. Yao, "Covariance matrix repairing in gaussian based EDAs," in *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, 2007. doi: 10.1109/CEC.2007.4424501 pp. 415–422.

[3] O. Ledoit and M. Wolf, "A well-conditioned estimator for large dimensional covariance matrices," *Journal of Multivariate Analysis Archive*, vol. 88, no. 2, pp. 265–411, 2004.

[4] S. Meyer-Nieberg and E. Kropat, "Adapting the covariance in evolution strategies," in *Proceedings of ICORES 2014*. SCITEPRESS, 2014, pp. 89–99.

[5] ——, "A new look at the covariance matrix estimation in evolution strategies," in *Operations Research and Enterprise Systems*, ser. Communications in Computer and Information Science, E. Pinson, F. Valente, and B. Vitoriano, Eds. Springer International Publishing, 2015, vol. 509, pp. 157–172. ISBN 978-3-319-17508-9. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-17509-6_11

[6] A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*, ser. Natural Computing Series. Berlin: Springer, 2003.

[7] I. Rechenberg, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Stuttgart: Frommann-Holzboog Verlag, 1973.

[8] H.-P. Schwefel, *Numerical Optimization of Computer Models*. Chichester: Wiley, 1981.

[9] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies: A comprehensive introduction," *Natural Computing*, vol. 1, no. 1, pp. 3–52, 2002.

[10] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

[11] H.-G. Beyer and B. Sendhoff, "Covariance matrix adaptation revisited - the CMSA evolution strategy -," in *PPSN*, ser. Lecture Notes in Computer Science, G. Rudolph *et al.*, Eds., vol. 5199. Springer, 2008. ISBN 978-3-540-87699-1 pp. 123–132.

[12] H.-G. Beyer and S. Meyer-Nieberg, "Self-adaptation of evolution strategies under noisy fitness evaluations," *Genetic Programming and Evolvable Machines*, vol. 7, no. 4, pp. 295–328, 2006.

[13] N. Hansen, "The CMA evolution strategy: A comparing review," in *Towards a new evolutionary computation. Advances in estimation of distribution algorithms*, J. Lozano *et al.*, Eds. Springer, 2006, pp. 75–102.

[14] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate distribution," in *Proc. 3rd Berkeley Symp. Math. Statist. Prob. 1*, Berkeley, CA, 1956, pp. 197–206.

[15] ——, "Estimation of a covariance matrix," in *Rietz Lecture, 39th Annual Meeting*. Atlanta, GA: IMS, 1975.

[16] J. Schäffer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics,," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, p. Article 32, 2005.

[17] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Sbornik: Mathematics*, vol. 1, no. 4, pp. 457–483, 1967.

[18] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008. doi: 10.1093/biostatistics/kxm045. [Online]. Available: http://biostatistics.oxfordjournals.org/content/9/3/432.abstract

[19] E. Levina, A. Rothman, and J. Zhu, "Sparse estimation of large covariance matrices via a nested lasso penalty," *Ann. Appl. Stat.*, vol. 2, no. 1, pp. 245–263, 03 2008. doi: 10.1214/07-AOAS139. [Online]. Available: http://dx.doi.org/10.1214/07-AOAS139

TABLE II

EXPECTED RUNNING TIME (ERT IN NUMBER OF FUNCTION EVALUATIONS) DIVIDED BY THE RESPECTIVE BEST ERT MEASURED DURING BBOB-2009 IN DIMENSION 10. THE ERT AND IN BRACES, AS DISPERSION MEASURE, THE HALF DIFFERENCE BETWEEN 90 AND 10%-TILE OF BOOTSTRAPPED RUN LENGTHS APPEAR FOR EACH ALGORITHM AND TARGET, THE CORRESPONDING BEST ERT IN THE FIRST ROW. THE DIFFERENT TARGET $\Delta f$-VALUES ARE SHOWN IN THE TOP ROW. #SUCC IS THE NUMBER OF TRIALS THAT REACHED THE (FINAL) TARGET $f_\text{opt} + 10^{-8}$. THE MEDIAN NUMBER OF CONDUCTED FUNCTION EVALUATIONS IS ADDITIONALLY GIVEN IN *italics*, IF THE TARGET IN THE LAST COLUMN WAS NEVER REACHED. ENTRIES, SUCCEEDED BY A STAR, ARE STATISTICALLY SIGNIFICANTLY BETTER (ACCORDING TO THE RANK-SUM TEST) WHEN COMPARED TO ALL OTHER ALGORITHMS OF THE TABLE, WITH $p = 0.05$ OR $p = 10^{-k}$ WHEN THE NUMBER $k$ FOLLOWING THE STAR IS LARGER THAN 1, WITH BONFERRONI CORRECTION BY THE NUMBER OF INSTANCES.

| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **f1** | 22 | 23 | 23 | 23 | 23 | 23 | 23 | 15/15 | **f13** | 387 | 596 | 797 | 1014 | 4587 | 6208 | 7779 | 15/15 |
| CMSA | 4.0(3) | 8.6(3) | 14(4) | 19(4) | 26(6) | 38(8) | 50(5) | 15/15 | CMSA | 15(24) | **19**(10) | **31**(31) | **58**(48) | **28**(32) | **89**(17) | **185**(103) | 2/15 |
| Thres | 4.2(2) | 9.2(3) | 14(4) | 18(3) | 24(3) | 35(5) | 46(7) | 15/15 | Thres | **6.3**(5) | 35(86) | 56(25) | 107(156) | 72(89) | 461(427) | ∞ *2e5* | 0/15 |
| Diag | **3.2**(1) | **7.5**(2) | **12**(3) | **18**(2) | **23**(4) | **34**(3) | **44**(4) | 15/15 | Diag | 12(7) | 52(113) | 71(54) | 164(232) | 117(330) | ∞ | ∞ *2e5* | 0/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f2** | 187 | 190 | 191 | 191 | 193 | 194 | 195 | 15/15 | **f14** | 37 | 98 | 133 | 205 | 392 | 687 | 4305 | 15/15 |
| CMSA | 65(34) | 85(21) | 96(29) | 105(16) | 109(25) | 113(21) | 129(57) | 15/15 | CMSA | 1.1(0.5) | 2.2(0.5) | 2.8(0.5) | 3.5(0.9) | 4.3(1) | 8.7(2) | 4.8(3) | 15/15 |
| Thres | 71(35) | 88(23) | 100(27) | 109(22) | 113(13) | 120(18) | 125(14) | 15/15 | Thres | 0.90(0.7) | 2.2(1) | **2.6**(0.7) | **3.2**(1) | **4.3**(1) | **8.7**(2) | **4.4**(3) | 15/15 |
| Diag | **55**(44) | **73**(54) | **88**(57) | **97**(70) | **101**(57) | **107**(67) | **111**(72) | 15/15 | Diag | **0.64**(0.6) | **1.8**(0.8) | 2.7(0.8) | 3.6(0.6) | 9.1(2) | 23(8) | 9.1(3) | 15/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f3** | 1739 | 3600 | 3609 | 3636 | 3642 | 3646 | 3651 | 15/15 | **f15** | 4774 | 39246 | 73643 | 74669 | 75790 | 77814 | 79834 | 12/15 |
| CMSA | 20(49) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 | CMSA | 7.0(10) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Thres | 33(36) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 | Thres | 11(14) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Diag | **11**(6) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 | Diag | **5.8**(6) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f4** | 2234 | 3626 | 3660 | 3695 | 3707 | 3744 | 28767 | 12/15 | **f16** | 425 | 7029 | 15779 | 45669 | 51151 | 65798 | 71570 | 15/15 |
| CMSA | 60(59) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 | CMSA | **1.0**(1) | 1.8(4) | **13**(14) | **31**(57) | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Thres | 119(128) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 | Thres | 1.1(0.7) | **1.8**(2) | 87(89) | 64(58) | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Diag | **38**(40) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 | Diag | 2.6(3) | 2.6(3) | 27(37) | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f5** | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 15/15 | **f17** | 26 | 429 | 2203 | 6329 | 9851 | 20190 | 26503 | 15/15 |
| CMSA | **12**(5) | **17**(4) | **17**(12) | **17**(10) | **17**(10) | **17**(8) | **17**(5) | 15/15 | CMSA | **0.71**(0.5) | **18**(51) | **23**(38) | **30**(16) | **140**(142) | ∞ | ∞ *2e5* | 0/15 |
| Thres | 14(8) | 19(9) | 21(8) | 21(7) | 21(7) | 21(8) | 21(8) | 15/15 | Thres | 39(140) | 34(176) | 37(63) | 67(123) | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Diag | 13(7) | 17(9) | 18(9) | 18(5) | 18(9) | 18(11) | 18(7) | 15/15 | Diag | 1.1(1) | 34(62) | 23(44) | 58(90) | 146(141) | ∞ | ∞ *2e5* | 0/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f6** | 412 | 623 | 826 | 1039 | 1292 | 1841 | 2370 | 15/15 | **f18** | 238 | 836 | 7012 | 15928 | 27536 | 37234 | 42708 | 15/15 |
| CMSA | **1.4**(0.2) | 3.3(3) | 11(29) | 14(32) | 19(25) | 25(17) | 163(268) | 6/15 | CMSA | 68(234) | **129**(263) | **124**(173) | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Thres | 1.8(1) | 5.4(1) | 7.0(15) | 6.9(14) | 10(19) | 20(56) | 30(52) | 14/15 | Thres | 88(33) | 313(539) | 130(104) | **184**(455) | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Diag | 1.6(0.7) | **2.8**(1) | **4.3**(5) | **4.4**(3) | **4.7**(4) | **13**(84) | **21**(45) | 13/15 | Diag | **5.1**(16) | 189(315) | 192(225) | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f7** | 172 | 1611 | 4195 | 5099 | 5141 | 5141 | 5389 | 15/15 | **f19** | 1 | 1 | 10609 | 9.8e5 | 1.4e6 | 1.4e6 | 1.4e6 | 15/15 |
| CMSA | 4.0(3) | **26**(50) | **85**(72) | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 | CMSA | 18(3) | 1.5e5(8e4) | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Thres | 5.7(5) | 103(70) | 230(313) | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 | Thres | 19(11) | 1.1e5(1e5) | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Diag | **2.4**(4) | 32(35) | 212(409) | **552**(402) | **548**(622) | **548**(467) | ∞ *2e5* | 0/15 | Diag | **15**(2) | **8.7e4**(1e5) | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f8** | 326 | 921 | 1114 | 1217 | 1267 | 1315 | 1343 | 15/15 | **f20** | 32 | 15426 | 5.5e5 | 5.7e5 | 5.7e5 | 5.8e5 | 5.9e5 | 15/15 |
| CMSA | **3.3**(0.7) | 17(8) | 18(7) | 18(3) | 18(7) | 19(11) | 19(5) | 15/15 | CMSA | 1.9(0.9) | 25(11) | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Thres | 8.5(5) | 18(12) | 18(9) | 18(7) | 18(7) | 18(7) | 18(9) | 15/15 | Thres | **1.7**(0.8) | 21(13) | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Diag | 6.6(22) | **17**(2) | **18**(5) | **17**(5) | **17**(5) | **18**(4) | **18**(4) | 15/15 | Diag | 2.1(1) | **20**(23) | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f9** | 200 | 648 | 857 | 993 | 1065 | 1138 | 1185 | 15/15 | **f21** | 130 | 2236 | 4392 | 4487 | 4618 | 5074 | 11329 | 8/15 |
| CMSA | **2.3**(2) | 25(13) | 24(10) | 22(12) | 22(10) | 21(10) | 21(10) | 15/15 | CMSA | 9.5(18) | 23(41) | 20(55) | 20(14) | 19(22) | 17(23) | 7.8(12) | 12/15 |
| Thres | 4.4(0.8) | **18**(8) | **19**(11) | **18**(7) | **17**(6) | **17**(14) | **17**(7) | 15/15 | Thres | **5.9**(7) | **17**(15) | **15**(19) | **15**(10) | **14**(28) | **13**(20) | **5.9**(6) | 13/15 |
| Diag | 4.9(2) | 35(17) | 31(29) | 29(20) | 27(15) | 27(14) | 26(17) | 15/15 | Diag | 21(0.6) | 19(13) | 20(35) | 20(27) | 19(69) | 18(17) | 8.0(13) | 12/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f10** | 1835 | 2172 | 2455 | 2728 | 2802 | 4543 | 4739 | 15/15 | **f22** | 98 | 2839 | 6353 | 6620 | 6798 | 8296 | 10351 | 6/15 |
| CMSA | **6.5**(3) | **7.6**(3) | **7.7**(3) | **7.5**(2) | **7.6**(3) | **4.9**(2) | **4.9**(2) | 15/15 | CMSA | **25**(61) | **6.3**(8) | 13(16) | 12(23) | 12(25) | 10(11) | 8.1(12) | 13/15 |
| Thres | 6.6(2) | 8.3(2) | 8.2(0.9) | 7.8(1) | 8.0(1) | 5.2(1) | 5.3(1) | 15/15 | Thres | 30(40) | 8.8(6) | 13(13) | 12(11) | 12(35) | 10(13) | 8.2(8) | 13/15 |
| Diag | 14(4) | 14(3) | 14(3) | 13(2) | 13(2) | 8.4(2) | 8.3(2) | 15/15 | Diag | 60(377) | 8.9(8) | **8.8**(16) | **8.8**(12) | **8.8**(20) | **7.8**(10) | **6.5**(11) | 14/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f11** | 266 | 1041 | 2602 | 2954 | 3338 | 4092 | 4843 | 15/15 | **f23** | 2.8 | 915 | 16425 | 1.8e5 | 2.0e5 | 2.1e5 | 2.1e5 | 15/15 |
| CMSA | **14**(3) | 6.2(2) | 3.2(1.0) | 3.4(0.9) | 3.5(1) | 3.3(0.6) | 3.0(1) | 15/15 | CMSA | 1.5(1) | 391(326) | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Thres | 17(4) | **6.1**(2) | **3.0**(0.7) | **3.0**(1) | **3.0**(1) | **2.9**(0.8) | **2.7**(0.7) | 15/15 | Thres | 2.0(3) | 313(575) | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Diag | 84(47) | 30(10) | 13(3) | 12(3) | 11(4) | 10(2) | 9.0(2) | 15/15 | Diag | **1.5**(2) | **165**(103) | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ | $\Delta f_\text{opt}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f12** | 515 | 896 | 1240 | 1390 | 1569 | 3660 | 5154 | 15/15 | **f24** | 98761 | 1.0e6 | 7.5e7 | 7.5e7 | 7.5e7 | 7.5e7 | 7.5e7 | 1/15 |
| CMSA | **4.2**(10) | **10**(11) | **13**(10) | **15**(15) | **16**(10) | **8.8**(4) | **7.8**(4) | 15/15 | CMSA | **4.8**(5) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Thres | 13(14) | 24(14) | 24(11) | 24(11) | 24(10) | 12(5) | 12(4) | 15/15 | Thres | 6.3(12) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |
| Diag | 14(39) | 29(26) | 33(19) | 35(6) | 34(7) | 17(3) | 15(6) | 15/15 | Diag | 6.4(7) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *2e5* | 0/15 |

TABLE III

EXPECTED RUNNING TIME (ERT IN NUMBER OF FUNCTION EVALUATIONS) DIVIDED BY THE RESPECTIVE BEST ERT MEASURED DURING BBOB-2009 IN DIMENSION 20. THE ERT AND IN BRACES, AS DISPERSION MEASURE, THE HALF DIFFERENCE BETWEEN 90 AND 10%-TILE OF BOOTSTRAPPED RUN LENGTHS APPEAR FOR EACH ALGORITHM AND TARGET, THE CORRESPONDING BEST ERT IN THE FIRST ROW. THE DIFFERENT TARGET $\Delta f$-VALUES ARE SHOWN IN THE TOP ROW. #SUCC IS THE NUMBER OF TRIALS THAT REACHED THE (FINAL) TARGET $f_{\mathrm{opt}} + 10^{-8}$. THE MEDIAN NUMBER OF CONDUCTED FUNCTION EVALUATIONS IS ADDITIONALLY GIVEN IN *italics*, IF THE TARGET IN THE LAST COLUMN WAS NEVER REACHED. ENTRIES, SUCCEEDED BY A STAR, ARE STATISTICALLY SIGNIFICANTLY BETTER (ACCORDING TO THE RANK-SUM TEST) WHEN COMPARED TO ALL OTHER ALGORITHMS OF THE TABLE, WITH $p = 0.05$ OR $p = 10^{-k}$ WHEN THE NUMBER $k$ FOLLOWING THE STAR IS LARGER THAN 1, WITH BONFERRONI CORRECTION BY THE NUMBER OF INSTANCES.

| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
|---|---|---|---|---|---|---|---|---|
| **f1** | 43 | 43 | 43 | 43 | 43 | 43 | 43 | 15/15 |
| CMSA | 4.9(2) | 10(2) | 15(2) | 19(2) | 25(2) | 34(2) | 45(2) | 15/15 |
| Thres | **4.9**(1) | 9.5(1) | 14(2) | 18(2) | **23**(3) | 33(3) | **42**(3) | 15/15 |
| Diag | 4.9(1) | **9.2**(2) | **13**(1) | **18**(2) | 23(3) | **33**(2) | 42(4) | 15/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f2** | 385 | 386 | 387 | 388 | 390 | 391 | 393 | 15/15 |
| CMSA | 173(32) | 240(38) | 265(33) | 273(38) | 277(34) | 285(31) | 293(27) | 15/15 |
| Thres | 154(41) | 212(41) | 245(25) | 259(29) | 265(23) | 273(26) | 282(29) | 15/15 |
| Diag | **96**(21)$^{\star 4}$ | **113**(14)$^{\star 4}$ | **122**(10)$^{\star 4}$ | **126**(9)$^{\star 4}$ | **128**(7)$^{\star 4}$ | **130**(6)$^{\star 4}$ | **130**(8)$^{\star 4}$ | 15/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f3** | 5066 | 7626 | 7635 | 7637 | 7643 | 7646 | 7651 | 15/15 |
| CMSA | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |
| Thres | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |
| Diag | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f4** | 4722 | 7628 | 7666 | 7686 | 7700 | 7758 | 1.4e5 | 9/15 |
| CMSA | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |
| Thres | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |
| Diag | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f5** | 41 | 41 | 41 | 41 | 41 | 41 | 41 | 15/15 |
| CMSA | **12**(3) | **15**(3) | **15**(6) | **15**(6) | **15**(6) | **15**(6) | **15**(7) | 15/15 |
| Thres | 13(5) | 17(13) | 18(12) | 18(11) | 18(6) | 18(11) | 18(4) | 15/15 |
| Diag | 14(8) | 17(8) | 18(8) | 18(8) | 18(9) | 18(7) | 18(10) | 15/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f6** | 1296 | 2343 | 3413 | 4255 | 5220 | 6728 | 8409 | 15/15 |
| CMSA | **1.5**(1) | **2.5**(2) | **4.6**(4) | 12(22) | 34(77) | **80**(138) | 331(279) | 2/15 |
| Thres | 3.2(3) | 4.4(4) | 5.0(3) | **7.9**(19) | **21**(65) | 83(117) | 324(190) | 2/15 |
| Diag | 17(33) | 26(32) | 32(37) | 51(28) | 64(69) | 135(279) | **204**(252) | 3/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f7** | 1351 | 4274 | 9503 | 16523 | 16524 | 16524 | 16969 | 15/15 |
| CMSA | 622(740) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |
| Thres | 2059(1547) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |
| Diag | **339**(291) | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |

| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
|---|---|---|---|---|---|---|---|---|
| **f8** | 2039 | 3871 | 4040 | 4148 | 4219 | 4371 | 4484 | 15/15 |
| CMSA | **11**(4) | **30**(32) | **31**(7) | **31**(52) | **31**(50) | **31**(28) | **31**(47) | 13/15 |
| Thres | 15(12) | 33(55) | 34(77) | 34(50) | 34(25) | 34(25) | 34(44) | 13/15 |
| Diag | 28(13) | 82(36) | 84(104) | 85(125) | 86(121) | 86(47) | 85(48) | 10/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f9** | 1716 | 3102 | 3277 | 3379 | 3455 | 3594 | 3727 | 15/15 |
| CMSA | 17(12) | 40(68) | 41(96) | 41(31) | 41(5) | 40(31) | 40(4) | 13/15 |
| Thres | **15**(8) | **20**(6) | **22**(8) | **23**(4) | **23**(8) | **23**(2) | **23**(5) | 15/15 |
| Diag | 36(7) | 52(4) | 54(36) | 56(5) | 57(5) | 57(4) | 57(31) | 14/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f10** | 7413 | 8661 | 10735 | 13641 | 14920 | 17073 | 17476 | 15/15 |
| CMSA | 10(5) | 11(2) | 9.2(2) | **7.8**(2) | **7.3**(1) | **6.7**(1) | **6.8**(0.6) | 15/15 |
| Thres | **8.4**(3) | **10**(3) | **9.1**(1) | 7.8(1) | 7.4(0.9) | 6.8(0.7) | 6.8(0.6) | 15/15 |
| Diag | 33(3) | 31(2) | 26(2) | 20(2) | 19(1) | 17(2) | 17(1) | 15/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f11** | 1002 | 2228 | 6278 | 8586 | 9762 | 12285 | 14831 | 15/15 |
| CMSA | **12**(2)$^{\star 2}$ | **7.5**(1) | **3.1**(0.4) | **2.6**(0.3) | **2.6**(0.3) | 2.5(0.5) | 2.5(0.3) | 15/15 |
| Thres | 15(1) | 8.1(0.6) | 3.2(0.4) | 2.6(0.2) | 2.6(0.4) | **2.3**(0.4) | **2.2**(0.5) | 15/15 |
| Diag | 223(44) | 115(28) | 43(10) | 33(8) | 30(5) | 24(3) | 20(3) | 15/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f12** | 1042 | 1938 | 2740 | 3156 | 4140 | 12407 | 13827 | 15/15 |
| CMSA | **2.5**(0.2) | **10**(9) | **13**(8) | **15**(8)$^{\star}$ | **14**(6)$^{\star}$ | **5.9**(2) | **6.2**(3) | 15/15 |
| Thres | 14(25) | 21(20) | 23(9) | 25(13) | 22(5) | 8.6(3) | 8.5(3) | 15/15 |
| Diag | 18(0.2) | 96(84) | 103(70) | 123(125) | 113(73) | 59(80) | 86(109) | 5/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f13** | 652 | 2021 | 2751 | 3507 | 18749 | 24455 | 30201 | 15/15 |
| CMSA | 156(614) | 545(891) | 2037(1600) | ∞ | ∞ | ∞ | ∞ *4e5* | 0/15 |
| Thres | 156(0.6) | **227**(693) | 2037(3199) | 1598(2167) | **299**(635) | ∞ | ∞ *4e5* | 0/15 |
| Diag | **46**(0.7) | 298(346) | **946**(1745) | **1598**(1768) | ∞ | ∞ | ∞ *4e5* | 0/15 |
| $\Delta f_{\mathrm{opt}}$ | 1e1 | 1e0 | 1e-1 | 1e-2 | 1e-3 | 1e-5 | 1e-7 | #succ |
| **f14** | 75 | 239 | 304 | 451 | 932 | 1648 | 15661 | 15/15 |
| CMSA | **1.8**(1.0) | 1.9(0.4) | 2.5(0.6) | 3.2(0.3) | **5.2**(0.8)$^{\star 3}$ | **11**(1)$^{\star 2}$ | 4.2(1) | 15/15 |
| Thres | 1.9(1) | **1.8**(0.4) | **2.3**(0.3) | **3.0**(0.4) | 6.8(0.4) | 14(1) | **3.7**(0.8) | 15/15 |
| Diag | 1.8(1) | 1.9(1) | 2.4(0.2) | 3.2(0.4) | 15(5) | 107(42) | 17(8) | 14/15 |

[20] T. J. Fisher and X. Sun, "Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix," *Computational Statistics & Data Analysis*, vol. 55, no. 5, pp. 1909 – 1918, 2011. doi: http://dx.doi.org/10.1016/j.csda.2010.12.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167947310004743

[21] X. Chen, Z. Wang, and M. McKeown, "Shrinkage-to-tapering estimation of large covariance matrices," *Signal Processing, IEEE Transactions on*, vol. 60, no. 11, pp. 5640–5656, 2012. doi: 10.1109/TSP.2012.2210546

[22] T. Cai and W. Liu, "Adaptive thresholding for sparse covariance matrix estimation," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 672–684, 2011.

[23] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 4, pp. 603–680, 2013.

[24] R. Ros and N. Hansen, "A simple modification in cma-es achieving linear time and space complexity," in *Parallel Problem Solving from Nature–PPSN X*. Springer, 2008, pp. 296–305.

[25] N. Hansen, "Adaptive encoding: How to render search coordinate system invariant," in *Parallel Problem Solving from Nature – PPSN X*, ser. Lecture Notes in Computer Science, G. Rudolph, T. Jansen, N. Beume, S. Lucas, and C. Poloni, Eds. Springer Berlin Heidelberg, 2008, vol. 5199, pp. 205–214. ISBN 978-3-540-87699-1. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-87700-4_21

[26] M. Pourahmadi, *High-Dimensional Covariance Estimation: With High-Dimensional Data*. John Wiley & Sons, 2013.

[27] J. Fan, Y. Liao, and H. Liu, "An overview on the estimation of large covariance and precision matrices," arXiv:1504.02995.

[28] D. Guillot and B. Rajaratnam, "Functions preserving positive definiteness for sparse matrices," *Transactions of the American Mathematical Society*, vol. 367, no. 1, pp. 627–649, 2015.

[29] N. Hansen, A. Auger, S. Finck, and R. Ros, "Real-parameter black-box optimization benchmarking 2012: Experimental setup," INRIA, Tech. Rep., 2012. [Online]. Available: http://coco.gforge.inria.fr/bbob2012-downloads

[30] S. Finck, N. Hansen, R. Ros, and A. Auger, "Real-parameter black-box optimization benchmarking 2010: Presentation of the noiseless functions," Institute National de Recherche en Informatique et Automatique, Tech. Rep., 2010, 2009/22.

DOI: 10.15439/2015F249
ACSIS, Vol. 5

# Genetic Algorithms for Balanced Spanning Tree Problem

Riham Moharam, Ehab Morsy
Department of Mathematics, Suez Canal
University, Ismailia 41522, Egypt.
Email: {riham.moharam, ehabmorsy}@science.suez.edu.eg

Ismail A. Ismail
Department of Computer Sciences,
6 October University, Egypt.
Email: amr442-2@hotmail.com

***Abstract*—Given an undirected weighted connected graph $G = (V, E)$ with vertex set $V$ and edge set $E$ and a designated vertex $r \in V$, we consider the problem of constructing a spanning tree in $G$ that balances both the minimum spanning tree and the shortest paths tree rooted at $r$. Formally, for any two constants $\alpha, \beta \geq 1$, we consider the problem of computing an $(\alpha, \beta)$-balanced spanning tree $T$ in $G$, in the sense that, (i) for every vertex $v \in V$, the distance between $r$ and $v$ in $T$ is at most $\alpha$ times the shortest distance between the two vertices in $G$, and (ii) the total weight of $T$ is at most $\beta$ times that of the minimum tree weight in $G$. It is well known that, for any $\alpha, \beta \geq 1$, the problem of deciding whether $G$ contains an $(\alpha, \beta)$-balanced spanning tree is NP-complete [15]. Consequently, given any $\alpha \geq 1$ (resp., $\beta \geq 1$), the problem of finding an $(\alpha, \beta)$-balanced spanning tree that minimizes $\beta$ (resp., $\alpha$) is NP-complete. In this paper, we present efficient genetic algorithms for these problems. Our experimental results show that the proposed algorithm returns high quality balanced spanning trees.**

***Index Terms*—Minimum Spanning Tree, Shortest Paths Tree, Balanced Spanning Tree, Genetic Algorithms, Graph Algorithms**

## I. Introduction

LET $G = (V, E)$ be an undirected edge-weighted connected graph with vertex set $V$ and edge set $E$ such that $|V| = n$ and $|E| = m$.

For any vertex $r$ in $G$, a spanning tree $T$ rooted at $r$ is a shortest paths tree if, for every vertex $v \in V$, the distance between $r$ and $v$ in $T_s$ equals the shortest distance between the two vertices in $G$. Dijkstra's algorithm is one of the well known polynomial time algorithms for computing shortest path tree in weighted graphs [22].

Note that, there exist weighted graphs in which the total weight of a shortest path tree may be much more than that of a minimum spanning tree, and vertices that are close to the designated root can be far away from the root in a minimum spanning tree (see [15] for an illustrative example). In this paper, we aim to find a spanning tree in weighted graphs that balances a minimum spanning tree and a shortest path tree, that is, a rooted tree of total weight at most a constant times the minimum tree weight, and the distance between the root and any vertex in the tree is at most a constant times the shortest distance between the two vertices in the graph. A formal definition of the problem can be described as follows.

*Definition 1:* [15] For any $\alpha, \beta \geq 1$, a rooted spanning tree $T$ of $G$ is called $(\alpha, \beta)$-balanced spanning tree if it satisfies the following two conditions:

1) For every vertex $v$, the distance between the root and $v$ in $T$ is at most $\alpha$ times the shortest distance between the two vertices in $G$.
2) The total weight of $T$ is at most $\beta$ times the minimum tree weight in $G$.

The shortest paths tree and minimum spanning tree are widely used in network routing. In particular, the shortest paths tree minimizes the delay from the source to every destination through a routing tree, and the minimum spanning tree minimizes the total routing cost along a tree. See [7], [14], [17], [21] and the references therein. Thus, balanced spanning tree is an appropriate routing tree for networks with the above two objectives.

Given any $\alpha \geq 1$, Awerbuch et al. [4] proposed an algorithm that approximates a minimum spanning tree and a shortest paths tree in edge-weighted graphs.

This work is partially supported by Alexander von Humboldt foundation

978-83-60810-66-8/$25.00©2015, IEEE

Namely, they modified the algorithm described in [3], [9] to compute an $(\alpha, 1+\frac{4}{\alpha-1})$-balanced spanning tree in $O(m+n\log n)$ time. Afterwards, Khullar et al. [15] improved the above result and presented a constructive linear time algorithm that outputs an $(\alpha, 1+\frac{2}{\alpha-1})$-balanced spanning tree. In other words, for any $\gamma > 0$, the algorithm of Khullar et al. [15] outputs an $(1+\sqrt{2}\gamma, 1+\frac{\sqrt{2}}{\gamma})$-balanced spanning tree in linear time.

For any $\alpha, \beta \geq 1$, the problem of deciding whether $G$ contains an $(\alpha, \beta)$-balanced spanning tree is NP-complete [15]. Consequently, given any $\alpha \geq 1$, the problem of finding an $(\alpha, \beta)$-balanced spanning tree that minimizes $\beta$ is NP-complete. Analogously, given any $\beta \geq 1$, the problem of finding an $(\alpha, \beta)$-balanced spanning tree that minimizes $\alpha$ is also NP-complete. In this paper, we present efficient genetic algorithms for these two problems. Our experimental results show that the proposed algorithm returns high quality balanced spanning trees.

The rest of this paper is organized as follows. Section II reviews some results on related problems. Section III presents the proposed genetic algorithm. Section IV evaluates our algorithm by applying it to randomly generated instances of the balanced spanning tree problem. Section V makes some concluding remarks.

## II. Related Work

In this section, we present results on related problems.

Bharath-Kumar and Jaffe [5] studied the problem of finding a rooted tree in the underlying graph such that the total distances from the root to all vertices is at most a constant times the minimum total distances from the root to all vertices.

For each graph $G$, the greatest distance between any two vertices in $G$ is called the diameter of $G$. A tree in a graph $G$ is called shallow-light tree if its diameter is at most a constant (greater than or equal 1) times the diameter of $G$ and with total weight at most a constant times the minimum tree weight. Awerbuch et al. [3] proved that each graph has a shallow-light tree.

Cong et al. [9] proposed a model of timing-driven global routing for cell-based design to improve the construction of a shallow-light tree based on the idea of finding minimum spanning trees with bounded radius. They designed an algorithm to find, for any constant $\epsilon > 0$, a spanning tree with radius $(1+\epsilon)\cdot R$ (using an analog of the classical Prim's minimum spanning tree structure), where $R$ is the minimum possible tree radius. They find a smooth trade-off between the radius and the cost of the tree. Afterwards, they proposed a new method [10] to improve their previous algorithm based on a provably good algorithm that simultaneously minimizes both total weight and longest interconnection path length of the tree. More specifically, their algorithm produced a tree with radius at most $(1+\epsilon)\cdot R$ and of total weight at most $(1+\frac{2}{\epsilon})$ times the minimum tree weight.

## III. Genetic Algorithm

In this section, we propose genetic algorithms to the two variants of the balanced spanning tree problem described in Section I. To avoid duplication, we present an algorithm for only one of these problems; the other variant of the problem can be described analogously. Namely, throughout this section, we focus on the problem in which we are given a constant $\alpha \geq 1$ and the objective is to compute an $(\alpha, \beta)$- balanced spanning tree that minimizes $\beta$.

We first introduce some terminologies that will be used throughout this section. Let $G'$ be a subgraph of $G$. The sets $V(G')$ and $E(G')$ denote the set of vertices and edges of $G'$, respectively. The shortest distance between two vertices $u$ and $v$ in $G'$ is denoted by $d_{G'}(u,v)$. We use $w(G')$ to denote the sum $\sum_{e\in E(G')} w(e)$ of weights of all edges in $G'$. For two subgraphs $G_1$ and $G_2$ of $G$, let $G_1 \cup G_2$, $G_1 \cap G_2$, and $G_1 - G_2$ denote the subgraph induced by $E(G_1)\cup E(G_2)$, $E(G_1)\cap E(G_2)$, and $E(G_1)-E(G_2)$, respectively. For any edge $e$ in $G$, let $Adj(e)$ denote the set of all adjacent edges to $e$ in $G$, where two edges of $G$ are called adjacent if they share a common vertex.

### A. Algorithm Overview

The Genetic Algorithm (GA) is an iterative optimization approach based on the principles of genetics and natural selection [2]. The first step of genetic algorithms is to determine a suitable data structure to represent individual solutions (chromosomes), and then construct an initial population (first generation) of prescribed cardinality $pop-size$. A typical intermediate iteration of genetic algorithms can be outlined as follows. Starting with the current generation, we use a predefined selection technique to repeatedly choose a pair of individuals (parents) in the current generation to reproduce, with probability $p_c$, a new set of individuals (offsprings) by applying crossover operation to the selected pair. To keep an appropriate diversity among different generations, we apply mutation operation with specific probability $p_m$ to genes of individuals of the current generation to get more offsprings. A new generation is then selected from both the offspring and the current generation based on their fitness values (the more suitable solutions have more chances to reproduce). The algorithm terminates when it meets prescribed stopping criteria.

A formal description of the proposed genetic algorithm is given in Algorithm 1.

In genetic algorithms, determining representation method, population size, selection technique, crossover and mutation probabilities, and stopping criteria are crucial since they mainly affect the convergence of the algorithm (see [1], [13], [18], [19], [20]).

The rest of this section is devoted to describe steps of Algorithm 1 in details.

### *B. Representation*

Let $G = (V, E)$ be a given undirected graph such that $V = \{1, 2, \ldots, n\}$. Note that, each edge $e \in E$ with end points $i$ and $j$ can be defined by an unordered pair $\{i, j\}$, and consequently, any subgraph $G'$ of $G$ is uniquely defined by the set of unordered pairs of all its edges. In particular, any spanning tree $T$ in $G$ is induced by a set of exactly $n-1$ pairs corresponding to its edges since $T$ is a subgraph of $G$ that spans all vertices in $V$ and has no cycles. Therefore, each chromosome (spanning tree) can be represented as a sequence of ordered pairs of integers each of which represent a gene (edge) in the chromosome.

### *C. Initial Population*

Before initializing the first generation of the genetic algorithm, we have to first decide its size $pop-size$. As mentioned before, this decision on population size affect the convergence of the algorithm. In particular, small population size may lead to weak solutions, while, large population size increases the space and time complexity of the algorithm. Many literatures studied the influence of the population size to the performance of genetic algorithms (see [19] and the references therein). In this paper, we discuss the effect of the population size on the convergence time of the algorithm (cf. Section IV).

**Algorithm 1** Genetic Algorithm for the Balanced Spanning Tree Problem

```
Input: An edge-weighted graph $G$, a population size $pop - size$, a maximum number of generations
    $maxgen$, a crossover probability $p_c$, a mutation probability $p_m$, a shortest paths tree $T_s$,
    a minimum spanning tree $T_m$, and a real number $\alpha \geq 1$.
Output: An $(\alpha, \beta)$-balanced spanning tree that minimizes $\beta$.
1. Compute an initial population $I_0$ (cf. Section III-C).
2. $gen \leftarrow 1$.
3. While ($gen \leq maxgen$) do
4.    For $i = 1$ to $pop - size$ do
5.       Select a pair of chromosomes from $I_{gen-1}$ (cf. Section III-D).
6.       Apply crossover operator with probability $p_c$
         to the selected pair of chromosomes to get
         two offsprings (cf. Section III-E).
7.    Endfor
8.    For each chromosome in $I_{gen-1}$, apply mutation
      operator with probability $p_m$ to get an offspring
      (cf. Section III-F).
9.    Extend $I_{gen-1}$ with valid offsprings output from
      lines 6 and 8.
10.   Find the minimum total weight chromosome
      $T_{gen-1}$ in $I_{gen-1}$.
11.   If $gen \geq 2$ and $w(T_{gen-2}) = w(T_{gen-1}) =$
      $w(T_{gen})$, then break.
12.   Select $pop - size$ chromosomes from $I_{gen-1}$ to
      form $I_{gen}$ (cf. Section III-D).
13.   $gen \leftarrow gen + 1$.
14. Endwhile
15. Output $T_{gen}$.
```

We apply random initialization to generate an initial population. Namely, we compute each chromosome in the initial population by repeatedly applying the following simple procedure as long as the length of the chromosome (number of edges) is less than $n - 1$. Let $T$ denote the tree constructed so far by the procedure (initially, $T$ consists of a random vertex from $V(G)$). We first select a random vertex $v \notin V(T)$ from the set of the neighbors of all vertices in $T$, and then add the the edge $e = (u, v)$ to $T$, where $u$ is the neighbor of $v$ in $T$. It is easy to verify that the above procedure returns a tree after exactly $n-1$ iterations. The generated tree $T$ is added to the initial population only if it is valid, where a tree $T$ is said to be a valid chromosome if, for every vertex $v$ in $T$, it holds that $d_T(r,v) \leq \alpha \cdot d_G(r,v)$.

The above algorithm is repeated as long as the number of constructed population is less than $pop-size$.

### *D. Selection Process*

In this paper, we apply four common selection techniques: random selection, roulette wheel selection, stochastic universal sampling selection, and tournament selection. All these techniques, except the random selection, are called fitness-proportionate selection techniques since they are based on a predefined fitness function used to evaluate the quality of individual chromosomes. Here, the objective function of the underlying balanced spanning tree problem (i.e., the ratio of the the minimum weight tree to the total weight of the chromosome) is used as the fitness function of each chromosome. We assume that the same selection technique is used throughout the execution of the algorithm. The rest of this section is devoted to briefly describe these selection techniques.

**Random Selection (RS):** [2] It is the simplest selection operator, where each chromosome has the same probability to be selected. That is, from a population of size $q$, each chromosome has the chance to be chosen with probability $1/q$.

**Roulette Wheel Selection (RWS):** [2], [8] In the roulette wheel technique, the probability of selecting a chromosome is based on its fitness value. More precisely, each chromosome is selected with the probability that equals to its normalized fitness value, i.e., the ratio of its fitness value to the total fitness values of all chromosomes in the set from which it will be selected.

**Stochastic Universal Sampling Selection (SUS):** [6], [8] It is a single phase sampling; instead of a single selection pointer used in roulette wheel approach, SUS uses $h$ equally spaced pointers, where $h$ is the number of chromosomes to be selected from the underlying population. All chromosomes are represented in number line randomly and a single pointer $ptr \in (0, \frac{1}{h}]$ is generated to indicate the first chromosome to be selected. The remaining $h-1$ individuals whose fitness spans the positions of the pointers $ptr + i/h$, $i = 1, 2, \ldots, h-1$ are then chosen.

**Tournament Selection (TRWS):** [2], [6] In this approach, we first select a set of $k < pop-size$ chromosomes randomly from the current population. From the selected set, we choose the required number of chromosomes by applying the roulette wheel selection approach.

### *E. Crossover Process*

In each iteration of the algorithm we repeatedly select a pair of chromosomes (parents) from the current generation and then apply crossover operator with probability $p_c$ to the selected chromosomes to get new chromosomes (offsprings). Simulations and experimental results of the literatures show that a typical crossover probability lies between $0.75$ and $0.95$. There are two common crossover techniques: single-point crossover and multi-point crossover. Many researchers studied the influence of crossover approach and crossover probability to the efficiency of the whole genetic algorithm, see for example [17], [20] and the references therein. In this paper, we use a multi-point crossover approach by exchanging a randomly selected set of edges between the two parents. In particular, for each selected pair of chromosomes $T_1$ and $T_2$, we generate a random number $s \in (0,1]$. If $s < p_c$ holds, we apply crossover operator to $T_1$ and $T_2$ as follows.

Define the two sets $E_1 = E(T_1) - E(T_2)$ and $E_2 = E(T_2) - E(T_1)$ ($|E_1| = |E_2|$ holds). Let $t = |E_1| = |E_2|$, and generate a random number $k$ from $[1,t]$. We first choose a random subset $E_1'$ of cardinality $k$ from $E_1$, and then add $E_1'$ to $T_2$ to get a subgraph $T'$ (i.e., $T' = T_2 \cup E_1'$). Clearly, $T'$ contains $k$ cycles each of which contains a distinct edge from $E_1'$. For every edge $e = (u,v)$ in $E_1'$, we apply the following procedure to fix a cycle containing $e$. Let $\widetilde{T}$ be the current subgraph (initially, $\widetilde{T} = T'$). We first find a path $P_{\widetilde{T}}(u,v)$ between $u$ and $v$ in $\widetilde{T} - \{e\}$. We then choose an edge $\widetilde{e}$ in $P_{\widetilde{T}}(u,v)$ by applying the selection technique used in the algorithm to the set of all edges in $P_{\widetilde{T}}(u,v)$, assuming that the weight of each edge is its fitness value. Finally, we delete $\widetilde{e}$ from subgraph $\widetilde{T}$. Note that, edges of large weights in $P_{\widetilde{T}}(u,v)$ have more chances to be deleted, and hence it is more likely that the current subgraph $\widetilde{T}$ attains total weight less than that of $T' - \{e\}$. In general, it is more likely that offsprings output from the crossover operation attains total weights less than that of its parents.

Similarly, we apply the above crossover technique by interchanging the roles of $T_1$ and $T_2$ one more offspring. Finally, we add each of the resulting spanning trees to the set of generated offsprings if it is valid.

### F. Mutation Process

Mutation is a genetic operator that intends to maintain the diversity among different generations of the population by altering some random genes (edges) in a chromosome. This may allow the algorithm to get better solutions by avoiding local minimum. Many results analyzed the role of mutation operator in genetic algorithms [1], [13], [17].

For a chromosome $T$ in the current population, we apply mutation operator such that each edge in $T$ is mutated with probability $p_m$. The standard range of $p_m$ lies between $1/\ell$ and $0.5$, where $\ell$ is the length (number of edges) of the chromosome. With $p_m = 1/\ell$, at least one gene (edge) on average should mutate. On the other hand, with $p_m = 0.5$, half of the edges should mutate on average, and consequently a random offspring is generated. In particular, for each edge $e$ in $T$, we generate a random number $s \in (0, 1]$, and then mutate $e$ if $s < p_m$ holds by replacing $e$ with a random edge from $Adj(e) - E(T)$. Let $T'$ denote the subgraph obtained after applying mutation operator to $T$. If $T'$ is disconnected, then we discard it. Otherwise, $T'$ is a spanning tree.

Finally, we add the resulting spanning tree to the set of generated offsprings only if it is valid.

## IV. Experimental Results

In this section, we evaluate the proposed genetic algorithms by applying it to several random instances of both variants of the balanced spanning tree problem. In particular, we generate a random graph $G$ of $n$ nodes by applying Erdos and Renyi [11] approach in which an edge is independently included between each pair of nodes of $G$ with probability $p$. Here, we generate random graphs with sizes 6, 10, 15, and 20, and a randomly chosen probability $p$. Moreover, all edge weights of the generated graphs are set to random integers from the range $[1, 100]$.

For each of the generated graphs, we apply the proposed algorithm with different selection techniques and different values of $\alpha$ (resp., $\beta$). We set the population size $pop-size = 30$, the maximum number of iterations the genetic algorithm executes $maxgen = 300$, the crossover probability $p_c = 0.9$, and the mutation probability $p_m = 0.01$. The algorithm terminates if either the number of iterations exceeds $maxgen$ or the solution does not change for three consecutive iterations. All obtained solutions are compared with the corresponding optimal solutions obtained by considering all possibilities of valid spanning trees in the underlying graphs. There are many algorithms for finding all spanning trees in undirected graphs, see for example [12].

We discuss the effect of the values of $\alpha$ (resp., $\beta$) on the convergence of the algorithm. It is seen that the running time of the algorithm decreases as the the value of $\alpha$ (resp., $\beta$) increases, see Figures 1-2 (resp. Figures 5-6).

We also study the effect of the population size $pop-size$ to the convergence of the algorithm. From the experimental results, we observe that a constant fraction of the number of nodes $n$ is an appropriate value for the population size, see Figures 3-4 and Figures 7-8.

All results presented in this section were performed in MATLAB R2014b on a computer powered by a core i7 processor and 16 GB RAM.

### A. Minimizing $\beta$

In this section, we present our experimental results for the problem in which $\alpha$ is given and the objective is to minimize $\beta$. We apply our algorithm with different values of $\alpha$ from the range $[1, 2]$. The results of applying our genetic algorithm to random graphs with sizes $n = 6$, $n = 10$, $n = 15$, and $n = 20$, are shown in Table I, Table II, Table III, and Table IV, respectively. In particular, Tables I-IV compare the values of $\beta$ returned by the algorithm with the corresponding optimal value of $\beta$. It is seen that the proposed algorithm outputs optimal balanced spanning tree for all the instances the algorithm applies to.

TABLE I
Values of $\beta$ corresponding to a random graph with $n = 6$.

| $\beta$ <br> $\alpha$ | **$\beta$-Optimal** | **$\beta$-RS** | **$\beta$-RWS** | **$\beta$-SUS** | **$\beta$-TRWS** |
|---|---|---|---|---|---|
| **1.1** | 1.142 | 1.142 | 1.142 | 1.142 | 1.142 |
| **1.2** | 1.047 | 1.142 | 1.142 | 1.047 | 1.142 |
| **1.3** | 1.047 | 1.047 | 1.047 | 1.047 | 1.047 |
| **1.4** | 1 | 1 | 1 | 1 | 1 |

TABLE II
VALUES OF $\beta$ CORRESPONDING TO A RANDOM GRAPH WITH $n = 10$.

| $\beta$ / $\alpha$ | $\beta$-Optimal | $\beta$-RS | $\beta$-RWS | $\beta$-SUS | $\beta$-TRWS |
|---|---|---|---|---|---|
| **1.1** | 1.021 | 1.021 | 1.021 | 1.021 | 1.021 |
| **1.2** | 1.010 | 1.021 | 1.010 | 1.010 | 1.010 |
| **1.3** | 1 | 1 | 1 | 1 | 1 |

TABLE III
VALUES OF $\beta$ CORRESPONDING TO A RANDOM GRAPH WITH $n = 15$.

| $\beta$ / $\alpha$ | $\beta$-Optimal | $\beta$-RS | $\beta$-RWS | $\beta$-SUS | $\beta$-TRWS |
|---|---|---|---|---|---|
| **1.1** | 1.025 | 1.074 | 1.074 | 1.025 | 1.025 |
| **1.2** | 1.025 | 1.074 | 1.062 | 1.025 | 1.025 |
| **1.3** | 1.025 | 1.074 | 1.062 | 1.025 | 1.025 |
| **1.4** | 1.012 | 1.049 | 1.049 | 1.012 | 1.012 |
| **1.5** | 1 | 1.049 | 1.049 | 1 | 1 |

TABLE IV
VALUES OF $\beta$ CORRESPONDING TO A RANDOM GRAPH WITH $n = 20$.

| $\beta$ / $\alpha$ | $\beta$-Optimal | $\beta$-RS | $\beta$-RWS | $\beta$-SUS | $\beta$-TRWS |
|---|---|---|---|---|---|
| **1.1** | 1.161 | 1.161 | 1.161 | 1.161 | 1.161 |
| **1.2** | 1.104 | 1.106 | 1.104 | 1.104 | 1.104 |
| **1.3** | 1.104 | 1.104 | 1.104 | 1.104 | 1.104 |
| **1.4** | 1.104 | 1.104 | 1.104 | 1.104 | 1.104 |
| **1.5** | 1.078 | 1.078 | 1.078 | 1.078 | 1.078 |
| **1.6** | 1.078 | 1.078 | 1.078 | 1.078 | 1.078 |
| **1.7** | 1.054 | 1.063 | 1.054 | 1.054 | 1.054 |
| **1.8** | 1.054 | 1.063 | 1.054 | 1.054 | 1.054 |
| **1.9** | 1.054 | 1.063 | 1.054 | 1.054 | 1.054 |
| **2** | 1.009 | 1.054 | 1.009 | 1.009 | 1.009 |

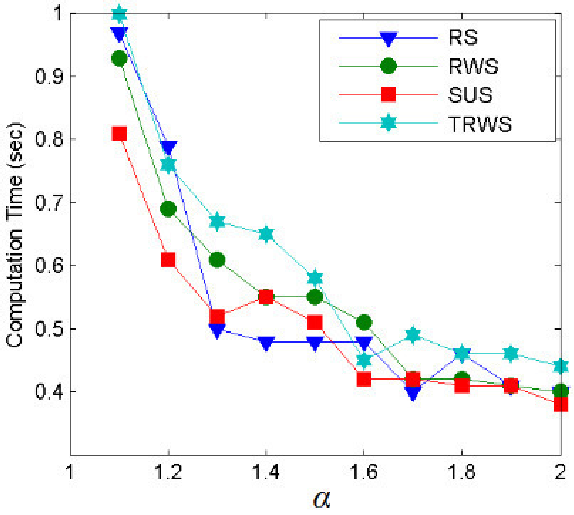


Fig. 1. The influence of $\alpha$ on the running time of the algorithm ($n = 15$)

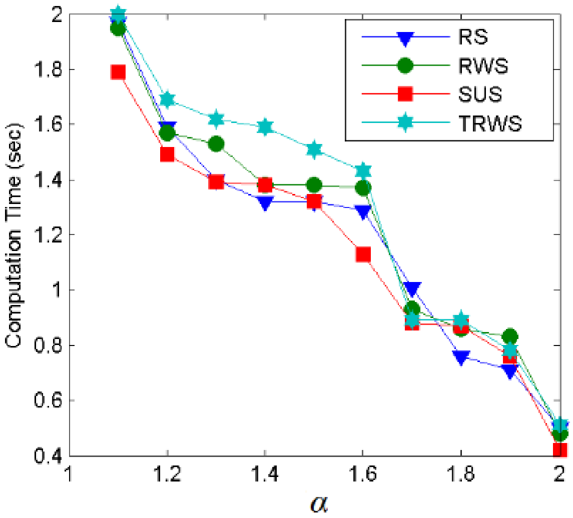


Fig. 2. The influence of $\alpha$ on the running time of the algorithm ($n = 20$)

Figures 1-2 show the influence of the value of $\alpha$ and the used selection technique on the execution time of the algorithm for graphs of $n = 15$ and $n = 20$, respectively. It is expected that the number of valid offsprings obtained in each iteration increases by relaxing the value of $\alpha$, and consequently it is more likely that the execution time of the algorithm decreases as the value of $\alpha$ increases.

We also evaluate the influence of the population size on the convergence of the proposed genetic algorithm. For a graph of $n$ nodes, we apply the algorithm with population sizes $n/3$, $2n/3$, $n$, $4n/3$, and $2n$. Figures 3-4 illustrate the running time of the algorithm applied to graphs of sizes $n = 15$, and $n = 20$, respectively, for fixed value $\alpha = 1.3$. We observe that the execution time of the algorithm increases as the population size increases, and the algorithm attains the least running time when the population size is set to a constant fraction of the graph size $n$.

### B. *Minimizing $\alpha$*

In this section, we present our experimental results for the problem in which $\beta$ is given and the objective is to

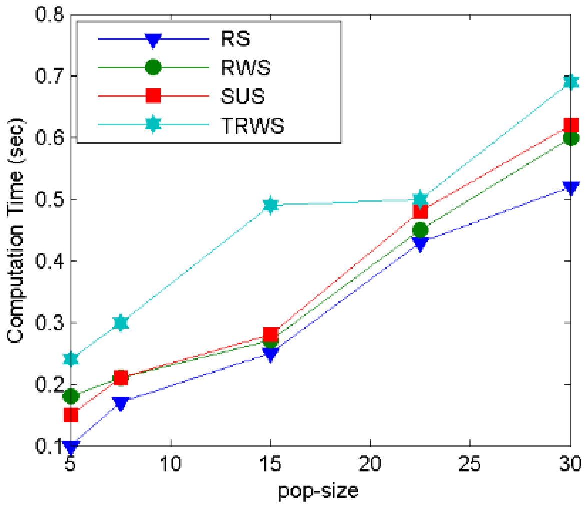


Fig. 3. The influence of $pop - size$ on the running time of the algorithm ($n = 15$)

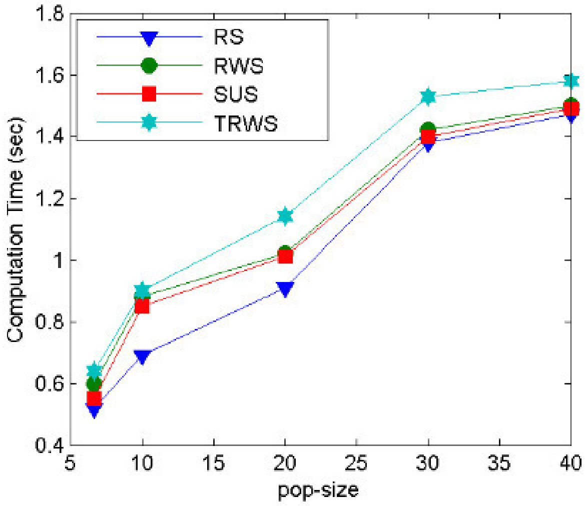


Fig. 4. The influence of $pop - size$ on the running time of the algorithm ($n = 20$)

TABLE V
VALUES OF $\alpha$ CORRESPONDING TO A RANDOM GRAPH WITH $n = 6$.

| $\alpha$ / $\beta$ | $\alpha$-Optimal | $\alpha$-RS | $\alpha$-RWS | $\alpha$-SUS | $\alpha$-TRWS |
|---|---|---|---|---|---|
| **1.1** | 1.3571 | 1.3571 | 1.3571 | 1.3571 | 1.3571 |
| **1.2** | 1.285 | 1.285 | 1.285 | 1.285 | 1.285 |
| **1.3** | 1.285 | 1.285 | 1.285 | 1.285 | 1.285 |
| **1.4** | 1 | 1 | 1 | 1 | 1 |

TABLE VI
VALUES OF $\alpha$ CORRESPONDING TO A RANDOM GRAPH WITH $n = 10$.

| $\alpha$ / $\beta$ | $\alpha$-Optimal | $\alpha$-RS | $\alpha$-RWS | $\alpha$-SUS | $\alpha$-TRWS |
|---|---|---|---|---|---|
| **1.1** | 1.238 | 1.238 | 1.238 | 1.238 | 1.238 |
| **1.2** | 1.096 | 1.096 | 1.096 | 1.096 | 1.096 |
| **1.3** | 1 | 1 | 1 | 1 | 1 |

TABLE VII
VALUES OF $\alpha$ CORRESPONDING TO A RANDOM GRAPH WITH $n = 15$.

| $\alpha$ / $\beta$ | $\alpha$-Optimal | $\alpha$-RS | $\alpha$-RWS | $\alpha$-SUS | $\alpha$-TRWS |
|---|---|---|---|---|---|
| **1.1** | 1.173 | 1.208 | 1.173 | 1.173 | 1.173 |
| **1.2** | 1.157 | 1.157 | 1.157 | 1.157 | 1.157 |
| **1.3** | 1 | 1 | 1 | 1 | 1 |

TABLE VIII
VALUES OF $\alpha$ CORRESPONDING TO A RANDOM GRAPH WITH $n = 20$.

| $\alpha$ / $\beta$ | $\alpha$-Optimal | $\alpha$-RS | $\alpha$-RWS | $\alpha$-SUS | $\alpha$-TRWS |
|---|---|---|---|---|---|
| **1.1** | 1.444 | 1.444 | 1.444 | 1.444 | 1.444 |
| **1.2** | 1.112 | 1.444 | 1.112 | 1.112 | 1.112 |
| **1.3** | 1.112 | 1.444 | 1.112 | 1.112 | 1.112 |
| **1.4** | 1.048 | 1.112 | 1.048 | 1.048 | 1.048 |
| **1.5** | 1.048 | 1.048 | 1.048 | 1.048 | 1.048 |
| **1.6** | 1.048 | 1.048 | 1.048 | 1.048 | 1.048 |
| **1.7** | 1 | 1 | 1 | 1 | 1 |

minimize $\alpha$. We apply our algorithm with different values of $\beta$. The results of applying our genetic algorithm to random graphs with sizes $n = 6$, $n = 10$, $n = 15$, and $n = 20$, are shown in Table V, Table VI, Table VII, and Table VIII, respectively. Also, for this problem, the proposed algorithm outputs optimal balanced spanning tree for all instances the algorithm applies to.

Figures 5-6 show the influence of the value of $\beta$ and the used selection technique to the execution time of the algorithm for graphs of $n = 15$ and $n = 20$, respectively. Most probably, the number of valid offsprings obtained in each iteration increases by relaxing the value of $\beta$, and consequently it is more likely that the execution time of the algorithm decreases as the value of $\beta$ increases.

We evaluate the influence of the population size on the convergence of the proposed genetic algorithm. For a graph of $n$ nodes, we apply the algorithm with population sizes $n/3$, $2n/3$, $n$, $4n/3$, and $2n$. Figures 7-8 illustrate the running time of the algorithm applied

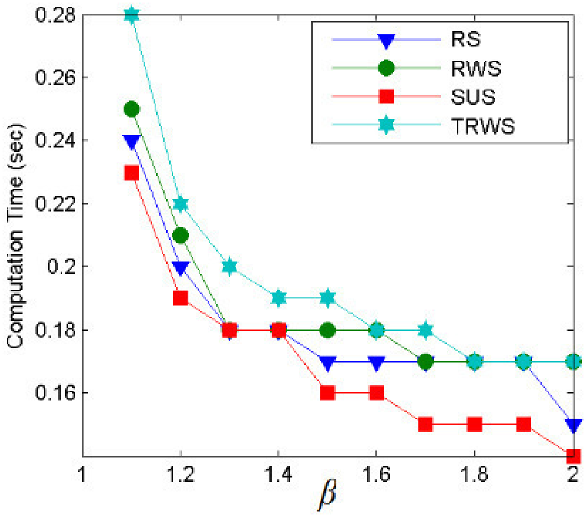


Fig. 5. The influence of $\beta$ on the running time of the algorithm ($n = 15$)

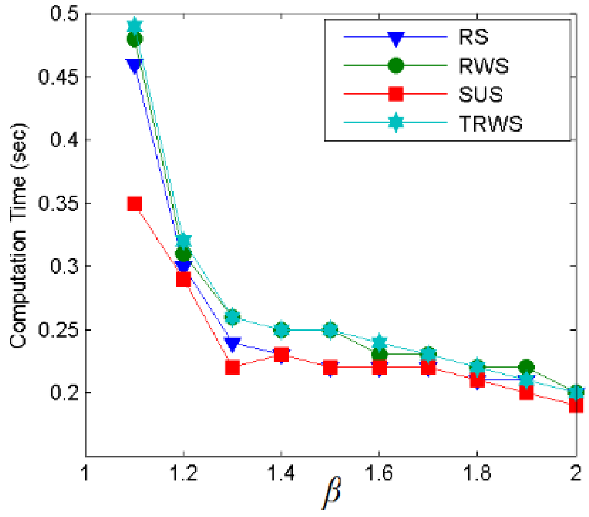


Fig. 6. The influence of $\beta$ on the running time of the algorithm ($n = 20$)

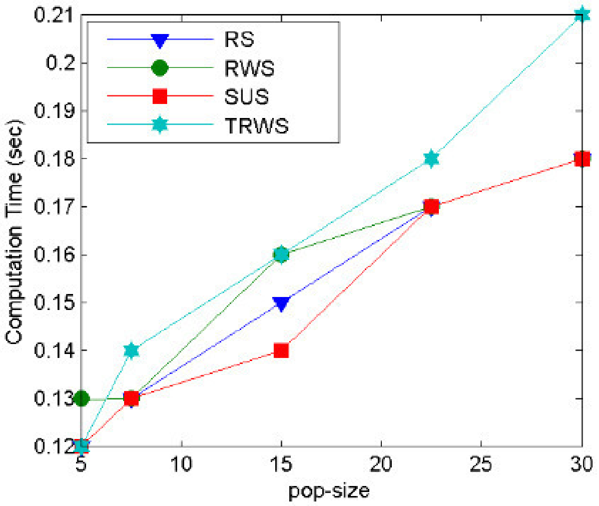


Fig. 7. The influence of $pop - size$ on the running time of the algorithm ($n = 15$)

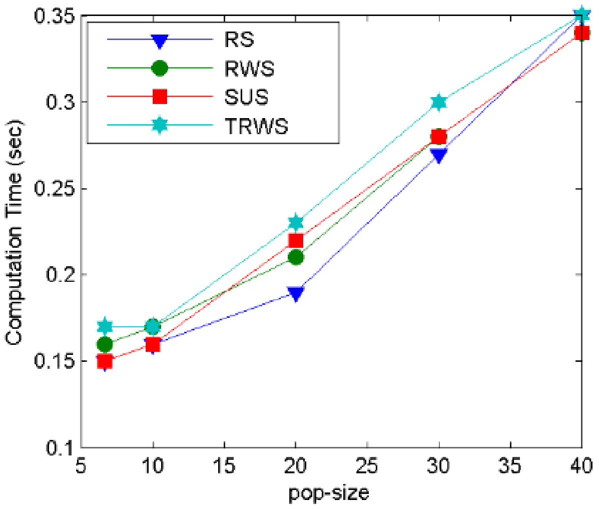


Fig. 8. The influence of $pop - size$ on the running time of the algorithm ($n = 20$)

to graphs of sizes $n = 15$, and $n = 20$, respectively, for fixed value $\beta = 1.3$. We observe that the algorithm attains the least running time when the population size is set to a constant fraction of the graph size $n$.

## V. Conclusion

In this paper, we have studied the problem of finding a tree that balances a shortest path tree and a minimum spanning tree in undirected edge-weighted graphs. In particular, we have focused on two NP-complete variants of the problem in which we are given a bound on how far the required tree is from the shortest path tree (resp., minimum spanning tree), and the objective is to find the closest tree to the minimum spanning tree (resp., shortest path tree) under this bound. We have designed genetic algorithms for these problems. We have evaluated our algorithm by applying it to random graph instances. The algorithm outputs optimal balanced spanning tree for all instances it has been applied to. It will be interesting to relax our model to balanced subgraphs instead of balanced trees, that is, the problem of finding a minimum weight subgraph such that the distance between any two vertices $u$ and $v$ in the subgraph is at most a given constant times the shortest distance between the two vertices in the underlying graph (this problem is known as $t$-spanner problem in the literatures). See [23], [24], [25], [26] and the references therein.

## References

[1] O. Abdoun, J. Abouchabaka, and C. Tajani, Analyzing the Performance of Mutation Operators to Solve the Travelling Salesman Problem, CoRR abs/1203.3099, 2012.

[2] Andris P. Engelbrecht, Computational Intelligence: an introduction, John Wiley & Sons, 2007.

[3] B. Awerbuch, A. Baratz, and D. Peleg, Cost-sensetive analysis of communication protocols, Proc. on Principles of Distributed Computing, pp. 177-187, 1990.

[4] B. Awerbuch, A. Baratz, and D. Peleg, Efficient broadcast and light-weight spanners, Manuscript, 1991.

[5] K. Bharath-Kumar and J. M. Jaffe, Routing to multiple destinations in computer networks, IEEE Transactions on Communications 31 (3), pp. 343-351, 1983.

[6] T. Blickle and L. Thiele, A comparison of Selection Schemes Used in Genetic Algorithms (Technical Report No. 11), Swiss Federal Institute of Technology (ETH) Zurich, Computer Engineering and Communications Networks Lab (TIK), 1995.

[7] R Campos and M Ricardo, A fast algorithm for computing minimum routing cost spanning trees, Computer Networks 52 (17), pp. 3229-3247, 2008.

[8] A. Chipperfield, P. Fleming, H. Pohlheim, and C. Fonseca, The Matlab Genetic Algorithm User's Guide, UK SERC, 1994.

[9] J. Cong, A. Kahng, G. Robins, M.Sarrafzadeh, and C. K. Wong, Performance-driven global routing for cell based IC's, Proc. IEEE Intl. Conference on Computer Design, pp. 170-173, 1991.

[10] J. Cong, A. Kahng, G. Robins, M.Sarrafzadeh, and C. K. Wong, Provably good performance-driven global routing, IEEE Transaction on CAD, pp. 739-752, 1992.

[11] P. Erdos and A. Renyi, On Random Graphs, Publ. Math, 290, 1959.

[12] H. N. Gabow and E. W. Myers, Finding all spanning trees of directed and undirected graphs, SIAM Journal on Computing, 7, pp. 280-287, 1978.

[13] J. Hesser and R. Männer, Towards an Optimal Mutation Probability for Genetic Algorithms, Proceedings of 1st workshop in Parallel problem solving from nature, pp. 23-32, 1991.

[14] G. Huang, X. Li, and J. He, Dynamic Minimal Spanning Tree Routing Protocol for Large Wireless Sensor Networks. In Proceedings of the 1st IEEE Conference on Industrial Electronics and Applications, Singapore, pp. 1-5, 2006.

[15] S. Khullar, B. Raghavachari, and N. Young, Balancing minimum spanning trees and shortest-path trees, Algorithmica 14, pp. 305-322, 1995.

[16] E. Kreyszig, Advanced Engineering Mathematics, John Wiley & Sons, 2011.

[17] C. Li, H. Zhang, B. Hao, and J. Li, A Survey on Routing Protocols for Large-Scale Wireless Sensor Networks. Sensors 11, pp. 3498-3526, 2011.

[18] W-Y. LIN, W-Y. LEE, and T-P. Hong, Adapting Crossover and Mutation Rates in Genetic Algorithms, the Sixth Conference on Artificial Intelligence and Applications, Kaohsiung, Taiwan, 2001.

[19] O. Roeva, S. Fidanova, and M. Paprzycki, Influence of the Population Size on the Genetic Algorithm Performance in Case of Cultivation Process Modelling. In the Proceedings of the Federated Conference on Computer Science and Information Systems pp. 371-376, 2013.

[20] K. Vekaria and C. Clack, Selective Crossover in Genetic Algorithms: An Empirical Study, volume 1498 of Lecture Notes in Computer Science, pp. 438-447, 1998.

[21] B. Xiao, Q. ZhuGe, and E. H.-M. Sha, Minimum Dynamic Update for Shortest Path Tree Construction, Global Telecommunications Conference, San Antonio, TX, pp. 126-130, 2001.

[22] B. Ye We and K. Chao, Spanning Trees and Optimization Problems, Chapman & Hall, 2004.

[23] M. Sigurd and M. Zachariasen, Construction of Minimum-Weight Spanners, Springer, Verlag Berlin Heidelberg, pp. 797-808, 2004.

[24] A. M. Farley, D. Zappala A. Proskurowski and K. Windisch, Spanners and Message Distribution in Networks, Dicrete Applied Mathematics, pp. 159-171, 2004.

[25] J. Gudmundsson, C. Levcopoulos and G. Narasimhan, Fast Greedy Algorithms for Constructing Sparse Geometric Spanners, SIAM Journal on Computing, pp. 1479-1500, 2002.

[26] G. Navarro, R. Paredes, and E. Chavez, t-Spanners as a Data Structure for Metric Space Searching, International Symposium on String Processing and Information Retrieval, SPIRE, LNCS 2476, pp. 298-309, 2002.

DOI: 10.15439/2015F296
ACSIS, Vol. 5

# Learning Fuzzy Cognitive Maps using Structure Optimization Genetic Algorithm

Katarzyna Poczęta, Alexander Yastrebov
Kielce University of Technology,
al. Tysiąclecia Państwa Polskiego 7
25-314 Kielce, Poland
Email: {k.piotrowska, a.jastriebow}@tu.kielce.pl

Elpiniki I. Papageorgiou
Center for Research and Technology Hellas, CERTH
6th km Charilaou-Thermi Rd., Thermi 57001, Greece
and
Technological Educational Institute (T.E.I.)
of Central Greece
3rd Km Old National Road Lamia-Athens,
Lamia 35100, Greece
Email: epapageorgiou@iti.gr, epapageorgiou@teilam.gr

***Abstract*—Fuzzy cognitive map (FCM) is a soft computing methodology that allows to describe the analyzed problem as a set of nodes (concepts) and connections (links) between them. In this paper a new Structure Optimization Genetic Algorithm (SOGA) for FCMs learning is presented for modeling complex decision support systems. The proposed approach allows to automatic construct and optimize the FCM model on the basis of historical multivariate time series. The SOGA defines a new learning error function with an additional penalty for highly complexity of FCM understood as a large number of concepts and a large number of connections between them. The aim of this study is the analysis of usefulness of the Structure Optimization Genetic Algorithm for fuzzy cognitive maps learning. Comparative analysis of the SOGA with other well-known FCM learning algorithms (Real-Coded Genetic Algorithm and Multi-Step Gradient Method) was performed on the example of prediction of rented bikes count. Simulations were done with the ISEMK (Intelligent Expert System based on Cognitive Maps) software tool. The obtained results show that the use of SOGA allows to significantly reduce the structure of the FCM model by selecting the most important concepts and connections between them.**

***Index Terms*—Fuzzy Cognitive Maps, Structure Optimization Genetic Algorithm, Real-Coded Genetic Algorithm, Multi-Step Gradient Method**

## I. Introduction

FUZZY cognitive map (FCM) [16] is a soft computing methodology combining the advantages of fuzzy logic and artificial neural networks. It allows to visualize and analyze problem as a set of nodes (concepts) and links (connections between them). One of the most important aspect connected with the use of fuzzy cognitive maps is their ability to learn on the basis of historical data [22]. Supervised [14], [15], [25] and population-based [1], [7], [8], [20], [28] methods allow to evaluate the weights of connections.

Fuzzy cognitive maps are an effective tool for modeling dynamic decision support systems [18]. They were applied to many different areas, such as prediction of pulmonary infection [19], scenario planning for the national wind energy sector [2] or integrated waste management [4]. Carvalho discussed possible use of FCM as tool for modeling and simulating complex social, economic and political systems [5]. The use of FCMs as pattern classifiers is presented in [23]. An innovative method for forecasting artificial emotions and designing an affective decision system on the basic of fuzzy cognitive map is proposed in [26]. The application of fuzzy cognitive maps to univariate time series modeling is discussed in [13], [12], [17]. Prediction of work of complex and imprecise systems on the basis of FCM is described in [27].

In practical applications to solve certain classes of problems (e.g. data analysis, prediction or diagnosis), finding the most significant concepts and connections plays an important role. It can be based on expert knowledge at all stages of analysis: designing the structure of the FCM model, determining the weights of the relationships and selecting input data. Supervised and population-based algorithms allow the automatic construction of fuzzy cognitive map on the basis of data selected by the experts or all available input data. However, modeling of complex systems can be difficult task due to the large amount of the information about analyzed problem. Fuzzy cognitive maps with the large number of concepts and connections between them can be difficult to interpret and impractical to use as the number of parameters to be established grows quadratically with the size of the FCM model [13]. In [12] nodes selection criteria for FCM designed to model univariate time series are proposed. Also some simplifications strategies by posteriori removing nodes and weights are presented [13].

In [21] the Structure Optimization Genetic Algorithm allowing selection of the crucial connections is proposed. In this paper, we present an extension of this algorithm that allows to significantly reduce the size of FCM by automatic selection not only the most important connections but also the most important concepts from all possible nodes. The proposed approach enables fully automatic construction of the FCM model by selection of crucial concepts and determining the relationships between them on the basis of available historical data. The SOGA is compared with well-known methods: the Multi-Step Gradient Method (MGM) [14] and the Real-Coded Genetic Algorithm (RCGA) [28] on the example of system for prediction of count of rented bikes. Learning and

978-83-60810-66-8/$25.00©2015, IEEE

testing of FCMs are based on historical multivariate time series, taken from the UCI Machine Learning Repository [6]. Simulations are accomplished with the use of the developed ISEMK (Intelligent Expert System based on Cognitive Maps) software tool [24].

The aims of this paper are:

- to introduce a new FCM learning algorithm that allows to automatic optimize the structure of fuzzy cognitive map by finding the most important concepts and connections between them on the basic of historical data,
- to perform a comparative analysis of the proposed method of FCM learning with other well-known algorithms on the example of system for prediction of count of rented bikes.

The paper is organized as follows. In Section II fuzzy cognitive maps are described. Section III introduces the proposed method for fuzzy cognitive maps learning. Section IV describes the developed software tool ISEMK. Section V presents selected results of simulation analysis of the proposed approach. The last Section contains a summary of the paper.

## II. Fuzzy Cognitive Maps

The structure of FCM is based on a directed graph:

$$< X, W > , \tag{1}$$

where $X = [X_1, ..., X_n]^T$ is the set of the concepts significant for the analyzed problem, $W$ is weights matrix, $W_{j,i}$ is the weight of the connection between the $j$-th concept and the $i$-th concept, taking on the values from the range $[-1, 1]$. Value of $-1$ means full negative influence, 1 denotes full positive influence and 0 means no causal effect [16].

Concepts obtain values in the range between $[0, 1]$ so they can be used in time series prediction. The values of concepts can be calculated according to the formula:

$$X_i(t+1) = F\left(X_i(t) + \sum_{j \neq i} W_{j,i} \cdot X_j(t)\right) , \tag{2}$$

where $t$ is discrete time, $t = 0, 1, 2, ..., T$, $T$ is end time of simulation, $X_i(t)$ is the value of the $i$-th concept, $i = 1, 2, ..., n$, $n$ is the number of concepts, $F(x)$ is a transformation function, which can be chosen in the form:

$$F(x) = \frac{1}{1 + e^{-cx}} , \tag{3}$$

where $c > 0$ is a parameter.

Fuzzy cognitive map can be automatic constructed with the use of supervised and population-based learning algorithms. In the next section, selected methods of FCMs learning are described.

## III. Fuzzy Cognitive Maps Learning

The aim of the FCM learning process is to estimate the weights matrix $W$. In the paper a new population-based approach for fuzzy cognitive maps learning is analyzed. Performance of the Structure Optimization Genetic Algorithm is compared with the Real-Coded Genetic Algorithm and the Multi-Step Gradient Method. Description of these methods is presented below.

### A. Real-Coded Genetic Algorithm

Real-Coded Genetic Algorithm defines each chromosome as a floating-point vector, expressed as follows [28]:

$$W' = [W_{1,2}, ..., W_{1,n}, W_{2,1}, W_{2,3}, ..., W_{2,n}, ..., W_{n,n-1}]^T , \tag{4}$$

where $W_{j,i}$ is the weight of the connection between the $j$-th and the $i$-th concept.

Each chromosome in the population is decoded into a candidate FCM and its quality is evaluated on the basis of a fitness function according to the objective [9]. The aim of the analyzed learning process is to optimize the weights matrix with respect to the prediction accuracy. Fitness function can be described as follows:

$$fitness_p(J_l) = \frac{1}{a \cdot J(l) + 1} , \tag{5}$$

where $a$ is a parameter, $a > 0$, $p$ is the number of chromosome, $p = 1, ..., P$, $P$ is the population size, $l$ is the number of population, $l = 1, ..., L$, $L$ is the maximum number of populations, $J(l)$ is the learning error function, described as follows:

$$J(l) = \frac{1}{(T-1)n_o} \sum_{t=1}^{T-1} \sum_{i=1}^{n_o} (Z_i^o(t) - X_i^o(t))^2 , \tag{6}$$

where $t$ is discrete time of learning, $T$ is the number of the learning records, $Z(t) = [Z_1(t), ..., Z_n(t)]^T$ is the desired FCM response for the initial vector $Z(t-1)$, $X(t) = [X_1(t), ..., X_n(t)]^T$ is the FCM response for the initial vector $Z(t-1)$, $n$ is the number of the all concepts, $n_o$ is the number of the output concepts, $X_i^o(t)$ is the value of the $i$-th output concept, $Z_i^o(t)$ is the reference value of the $i$-th output concept.

Each population is assigned a probability of reproduction. According to the assigned probabilities parents are selected and new population of chromosomes is generated. Chromosomes with above average fitness tend to receive more copies than those with below average fitness [9]. The basic operator of selection is a roulette wheel method. For each chromosome in population the probability of including a copy of such chromosome into the next population can be calculated according to the formula [11]:

$$P(p) = \frac{fitness_p(J_l)}{\sum\limits_{i=1}^{P} fitness_i(J_l)} , \tag{7}$$

where $p$ is the number of the chromosome, $P$ is the population size.

The population is mapped onto a roulette wheel, where each chromosome $p$ is represented by a space that proportionally corresponds to $P(p)$. In the analysis a more effective ranking method of selection was used [3]. Selecting a copy of the chromosome into the next population is based on ranking by fitness [9].

The crossover operator is a method for sharing information between parents to form new chromosomes. It can be applied to random pairs of chromosomes and the likelihood of crossover depends on probability defined by the crossover probability $P_c$. The popular crossover operators are the single-point crossover and the uniform crossover [11].

The mutation operator modifies elements of a selected chromosome with a probability defined by the mutation probability $P_m$. The use of mutation prevents the premature convergence of genetic algorithm to suboptimal solutions [11]. In the analysis Mühlenbein's and random mutation were used. To ensure the survival of the best chromosome in the population, elite strategy was applied. It retains the best chromosome in the population [9].

The learning process stops when the maximum number of populations $L$ is reached or the condition (8) is met.

$$fitness_{best}(J_l) > fitness_{max}\ , \tag{8}$$

where $fitness_{best}(J_l)$ is the fitness function value for the best chromosome, $fitness_{max}$ is a parameter.

### B. Structure Optimization Genetic Algorithm

In this paper a new Structure Optimization Genetic Algorithm is proposed, which allows to select the most important for prediction task concepts and connections between them. SOGA defines each chromosome as a floating-point vector type (4) and a binary vector expressed as follows:

$$C' = [C_1, C_2, ..., C_n]^T\ , \tag{9}$$

where $C_i$ is the information about including the $i$-th concept to the candidate FCM model, whereas $C_i$ =1 means that the candidate FCM model contains the $i$-th concept, $C_i$ =0 means that the candidate FCM model does not contain the $i$-th concept.

The quality of each population is calculated based on an original fitness function, described as follows:

$$fitness_p(J'_l) = \frac{1}{a \cdot J'(l) + 1}\ , \tag{10}$$

where $a$ is a parameter, $a > 0$, $p$ is the number of the chromosome, $l$ is the number of population, $l = 1, ..., L$, $L$ is the maximum number of populations, $J'(l)$ is the new learning error function with an additional penalty for highly complexity of FCM understood as a large number of concepts and non-zero connections between them [10]:

$$J'(l) = J(l) + b_1 \cdot \frac{n_r}{n^2} \cdot J(l) + b_2 \cdot \frac{n_c}{n} \cdot J(l)\ , \tag{11}$$

where $t$ is discrete time of learning, $T$ is the number of the learning records, $b_1, b_1$ are the parameters, $b_1 > 0$, $b_2 > 0$, $n_r$ is the number of the non-zero weights of connections, $n_c$ is the number of the concepts in the candidate FCM model, $n$ is the number of the all possible concepts, $J(l)$ is the learning error function type (11).

Fig. 1 illustrates the steps of the learning and analysis of the FCM in modeling prediction systems with the use of population-based algorithms (SOGA and RCGA).

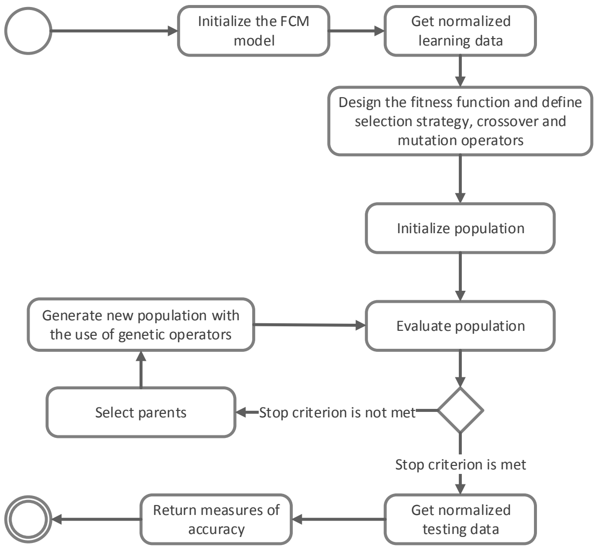

Initialize the FCM model
Get normalized learning data
Design the fitness function and define selection strategy, crossover and mutation operators
Initialize population
Evaluate population
Generate new population with the use of genetic operators
Select parents
Stop criterion is not met
Stop criterion is met
Get normalized testing data
Return measures of accuracy

Fig. 1. Activity diagram for population-based learning algorithm

In the paper the proposed algorithm was compared with the Real-Coded Genetic Algorithm and also with supervised learning based on Multi-Step Gradient Method.

### C. Multi-Step Gradient Method

Multi-step algorithms of FCM learning are some kind of generalization of known one-step methods. Effectiveness of these methods in modeling of decision support systems was presented in [15], [25]. Multi-step supervised learning based on gradient method is described by the equation [14]:

$$W_{j,i}(t+1) = P_{[-1,1]}(\sum_{k=0}^{m_1} \alpha_k \cdot W_{j,i}(t-k)+ \sum_{l=0}^{m_2}(\beta_l \cdot \eta_l(t) \cdot (Z_i(t-l) - X_i(t-l)) \cdot y_{j,i}(t-l)))\ , \tag{12}$$

where $\alpha_k, \beta_l, \eta_l$ are learning parameters, $k = 1, ..., m_1; l = 1, ..., m_2$, $m_1, m_2$ are the number of the steps of the method, $t$ is a time of learning, $t = 0, 1, ..., T-1$, $T$ is end time of learning, $X_i(t)$ is the value of the $i$-th concept, $Z_i(t)$ is the reference value of the $i$-th concept, $y_{j,i}(t)$ is a sensitivity function, $P_{[-1,1]}(x)$ is an operator of design for the set [-1,1], described as follows:

$$P_{[-1,1]}(x) = \frac{1 - e^{-x}}{1 + e^{-x}}\ , \tag{13}$$

Sensitivity function $y_{j,i}(t)$ is described by the equation:

$$\begin{aligned} y_{j,i}(t+1) = (y_{j,i}(t) + X_j(t)) \cdot F'(X_i(t) \\ + \textstyle\sum_{j \neq i} W_{j,i} \cdot X_j(t))\ , \end{aligned} \tag{14}$$

where $F'(x)$ is derivative of the stabilizing function.

Termination criterion can be expressed by the formula:

$$J(t) = \frac{1}{n} \sum_{i=1}^{n} (Z_i(t) - X_i(t))^2 < e\ , \tag{15}$$

where $e$ is a level of error tolerance.

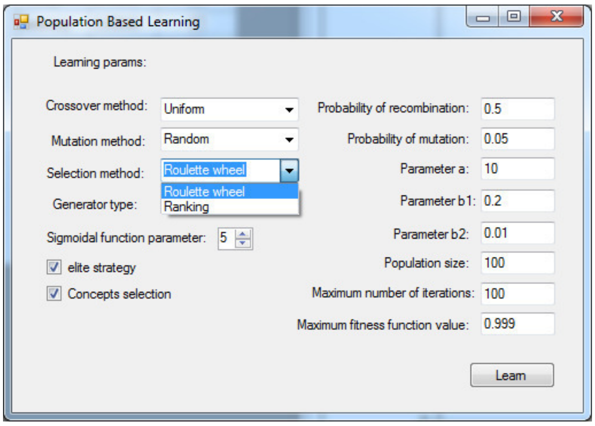
Simulation analysis of the presented algorithms performance was done with the use of ISEMK software tool. The basic features of ISEMK are described below.

## IV. ISEMK Software Tool

ISEMK is a universal tool for modeling decision support systems based on FCMs [24]. It allows to:

- initialize the structure of the FCM model historical data (reading from .data files),
- visualize the structure of the FCM model,
- learn the FCM model based on Multi-Step Gradient Method and historical data (reading from .data files),
- learn the FCM model with the use of population-based learning algorithms (RCGA, SOGA) and historical data (reading from .data files),
- test the accurace of the learned FCMs operation based on historical data (reading from .data files) by calculating Mean Squared Error measure,
- export the results of learning and testing to .csv files,
- visualize the results of learning and testing in the form of charts.

Figure 2 shows an exemplary initialization of SOGA. Figure 3 shows an exemplary visualization of testing of the learned FCM operation.


Population Based Learning

Learning params:

Crossover method: Uniform

Mutation method: Random

Selection method: Roulette wheel

Generator type:

Sigmoidal function parameter: 5

☑ elite strategy

☑ Concepts selection

Probability of recombination: 0.5

Probability of mutation: 0.05

Parameter a: 10

Parameter b1: 0.2

Parameter b2: 0.01

Population size: 100

Maximum number of iterations: 100

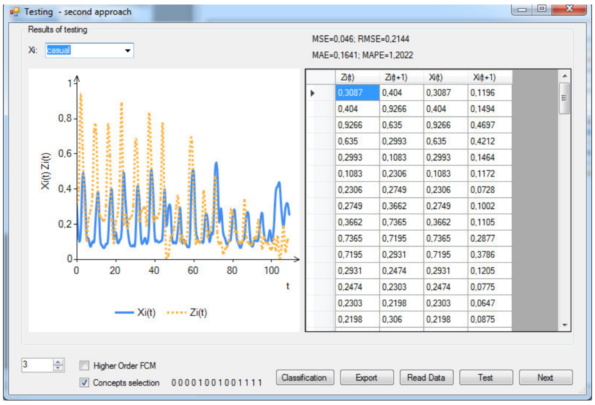
Maximum fitness function value: 0.999

Fig. 2. Exemplary visualization of population-based learning results

## V. Simulation Results

To evaluate the proposed Structure Optimization Genetic Algorithm, historical data taken from the UCI Machine Learning Repository [6] were used. The dataset contains bike sharing counts aggregated on daily basis (731 days) and has the following fields:

- season (1:springer, 2:summer, 3:fall, 4:winter),
- year (0: 2011, 1:2012),
- month ( 1 to 12),
- hour (0 to 23),
- holiday : weather day is holiday or not,
- weekday : day of the week,
- working day : if day is neither weekend nor holiday is 1, otherwise is 0,


Testing - second approach

Results of testing

Xi: casual

MSE=0,046; RMSE=0,2144

MAE=0,1641; MAPE=1,2022

| Zi(t) | Zi(t+1) | Xi(t) | Xi(t+1) |
|---|---|---|---|
| 0,3087 | 0,404 | 0,3087 | 0,1196 |
| 0,404 | 0,9266 | 0,404 | 0,1494 |
| 0,9266 | 0,635 | 0,9266 | 0,4697 |
| 0,635 | 0,2993 | 0,635 | 0,4212 |
| 0,2993 | 0,1083 | 0,2993 | 0,1464 |
| 0,1083 | 0,2306 | 0,1083 | 0,1172 |
| 0,2306 | 0,2749 | 0,2306 | 0,0728 |
| 0,2749 | 0,3662 | 0,2749 | 0,1002 |
| 0,3662 | 0,7365 | 0,3662 | 0,1105 |
| 0,7365 | 0,7195 | 0,7365 | 0,2877 |
| 0,7195 | 0,2931 | 0,7195 | 0,3786 |
| 0,2931 | 0,2474 | 0,2931 | 0,1205 |
| 0,2474 | 0,2303 | 0,2474 | 0,0775 |
| 0,2303 | 0,2198 | 0,2303 | 0,0647 |
| 0,2198 | 0,306 | 0,2198 | 0,0875 |

3

□ Higher Order FCM

☑ Concepts selection 0 0 0 0 1 0 0 1 0 0 1 1 1 1

Fig. 3. Exemplary visualization of testing of learned FCM

- weather situation (1: Clear, Few clouds, Partly cloudy, Partly cloudy, 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog),
- temperature (normalized temperature in Celsius),
- feeling temperature (normalized feeling temperature in Celsius),
- humidity (normalized humidity),
- wind speed (normalized wind speed),
- casual (count of casual users),
- registered (count of registered users),
- count (count of total rented bikes including both casual and registered).

The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA (http://capitalbikeshare.com/system-data). The corresponding weather and seasonal information (http://www.freemeteo.com) were added. Normalization of the available data in the $[0, 1]$ range is needed in order to use the FCM model. Conventional min-max normalization (16) can be used.

$$f(x) = \frac{x - min}{max - min} , \tag{16}$$

where $x$ is an input numeric value, $min$ is the minimum of the dataset, $max$ is the maximum of the dataset.

The aim of the study is one-step-ahead prediction of daily count of rented bikes based on the current values, environmental and seasonal settings. The dataset was divided into two subsets: learning (621 records) and testing (110 records) data. The learning process was accomplished with the use of Multi-Step Gradient Method, Real-Coded Genetic Algorithm and Structure Optimization Genetic Algorithm. Mean Squared Error for the output concepts (17) was used to estimate the performance of the FCM learning algorithms.

$$MSE = \frac{1}{n_o(T-1)} \sum_{t=1}^{T-1} \sum_{i=1}^{n_o} \left( Z_i^o(t) - X_i^o(t) \right)^2 , \tag{17}$$

where $t$ is time of testing, $t = 0, 1, ..., T-1$, $T$ is the number of the test records, $Z(t) = [Z_1(t), ..., Z_n(t)]^T$ is the desired FCM response for the initial vector $Z(t-1)$, $X(t) = [X_1(t), ..., X_n(t)]^T$ is the FCM response for the initial vector $Z(t-1)$, $X_i^o(t)$ is the value of the $i$-th output concept, $Z_i^o(t)$ is the reference value of the $i$-th output concept, $n_o$ is the number of the output concepts.

The learning process was accomplished for various learning parameters, determined using experimental trial and error method. Optimal parameters of all analyzed algorithms were chosen based on minimization of the objective function, described as follows:

$$f_c = f(MSE, n_r, n_c) = 10000MSE + n_r + 10n_c \ , \quad (18)$$

where $n_r$ is the number of the non-zero weights of the connections, $n_c$ is the number of the concepts.

The best results of all three approaches were selected. Fig. 4 presents the structure of the FCM learned with the use of MGM (for the following parameters: $m_1 = 1$, $m_2 = 0$, $\alpha_0 = 1.5$, $\alpha_1 = -0.5$, $\beta_0 = 50$, $\lambda_0 = 100$, $e = 0.001$, $c = 3$). Table I shows the weights matrix for this map. Fig. 5 presents the structure of the FCM learned with the use of RCGA (for the following parameters: $P = 100$, $L = 200$, ranking selection, uniform crossover, Mühlenbein's mutation, $a = 10$, $h_{max} = 0.999$, $c = 3$). Table II shows the weights matrix for this map. Fig. 6 presents the structure of the FCM learned with the use of SOGA (for the following parameters: $P = 100$, $L = 500$, ranking selection, uniform crossover, random mutation, $a = 100$, $h_{max} = 0.999$, $c = 5$, $b_1 = 0.1$, $b_2 = 0.01$). Table III shows the weights matrix for this map.

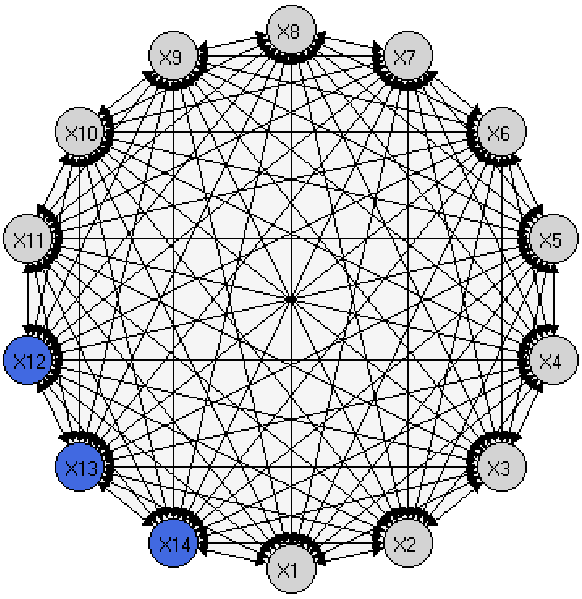


Fig. 4. The structure of the FCM learned with the use of MGM, where: $X_1$ – season, $X_2$ – year, $X_3$ – month, $X_4$ – holiday, $X_5$ – day of the week, $X_6$ – working day, $X_7$ – weather situation, $X_8$ – temperature, $X_9$ – feeling temperature, $X_{10}$ – humidity, $X_{11}$ – wind speed, $X_{12}$ – count of casual users, $X_{13}$ – count of registered users, $X_{14}$ – count of total rented bikes

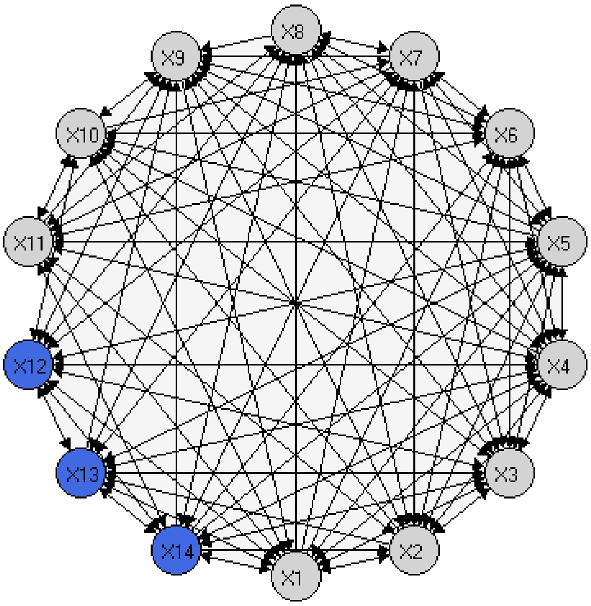


Fig. 5. The structure of the FCM learned with the use of RCGA, where: $X_1$ – season, $X_2$ – year, $X_3$ – month, $X_4$ – holiday, $X_5$ – day of the week, $X_6$ – working day, $X_7$ – weather situation, $X_8$ – temperature, $X_9$ – feeling temperature, $X_{10}$ – humidity, $X_{11}$ – wind speed, $X_{12}$ – count of casual users, $X_{13}$ – count of registered users, $X_{14}$ – count of total rented bikes

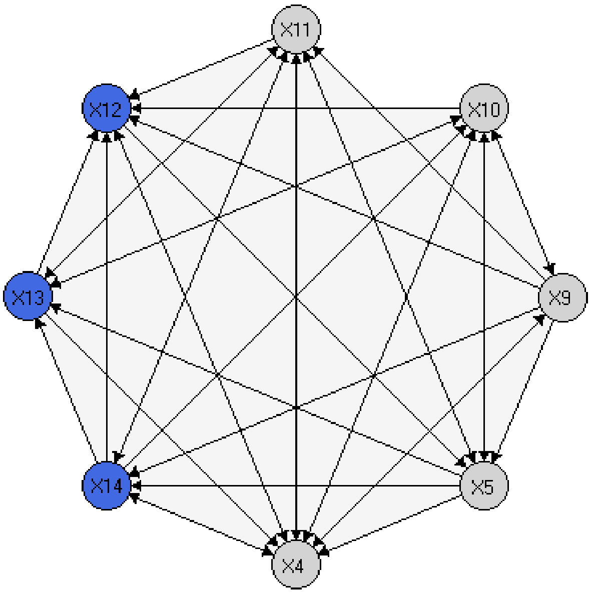


Fig. 6. The structure of the FCM learned with the use of SOGA, where: $X_4$ – holiday, $X_5$ – day of the week, $X_9$ – feeling temperature, $X_{10}$ – humidity, $X_{11}$ – wind speed, $X_{12}$ – count of casual users, $X_{13}$ – count of registered users, $X_{14}$ – count of total rented bikes

Figures 7-9 show the exemplary results of testing of the learned FCMs operation. Table IV shows selected results of the comparative analysis of the Multi-Step Gradient Method, the Real-Coded Genetic Algorithm and the Structure Optimization Genetic Algorithm. Ranking selection and uniform crossover were used.

TABLE I
EXEMPLARY WEIGHTS MATRIX FOR THE MAP LEARNED WITH THE USE OF MGM

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 0 | -0.02 | -0.01 | -0.12 | -0.06 | 0.06 | -0.12 | -0.07 | -0.06 | -0.08 | -0.09 | -0.11 | -0.01 | -0.03 |
| $X_2$ | -0.07 | 0 | -0.07 | -0.14 | -0.06 | -0.03 | -0.13 | -0.07 | -0.09 | -0.08 | -0.11 | -0.11 | -0.07 | -0.03 |
| $X_3$ | -0.08 | -0.01 | 0 | -0.12 | -0.09 | 0.05 | -0.12 | -0.07 | -0.07 | -0.08 | -0.09 | -0.13 | -0.02 | -0.03 |
| $X_4$ | -0.05 | -0.02 | -0.04 | 0 | 0.08 | 0.05 | -0.06 | -0.02 | 0.07 | 0.02 | -0.06 | -0.08 | 0.08 | 0.07 |
| $X_5$ | -0.12 | 0.07 | -0.12 | -0.11 | 0 | -0.04 | -0.13 | -0.11 | -0.13 | -0.11 | -0.11 | -0.17 | -0.1 | -0.09 |
| $X_6$ | -0.04 | 0.08 | -0.08 | -0.13 | -0.08 | 0 | -0.11 | -0.04 | -0.09 | -0.07 | -0.09 | 0.11 | -0.12 | -0.05 |
| $X_7$ | -0.06 | -0.02 | -0.05 | -0.08 | 0.05 | 0.08 | 0 | -0.01 | -0.02 | -0.09 | -0.07 | -0.09 | 0.04 | 0.05 |
| $X_8$ | -0.06 | 0.04 | -0.06 | -0.13 | 0.03 | -0.03 | -0.12 | 0 | -0.08 | -0.08 | -0.09 | -0.11 | -0.03 | -0.03 |
| $X_9$ | -0.07 | 0.03 | -0.05 | -0.12 | -0.06 | 0.06 | -0.12 | -0.09 | 0 | -0.08 | -0.09 | -0.12 | 0 | -0.03 |
| $X_{10}$ | -0.04 | 0.07 | -0.04 | -0.13 | 0.01 | -0.01 | -0.12 | -0.06 | -0.07 | 0 | -0.09 | -0.08 | -0.03 | -0.02 |
| $X_{11}$ | -0.05 | -0.01 | -0.04 | -0.08 | 0.04 | 0.06 | -0.08 | 0 | 0.03 | 0.02 | 0 | -0.07 | 0.04 | 0.04 |
| $X_{12}$ | -0.06 | 0 | 0.01 | -0.12 | 0.04 | 0.11 | -0.11 | -0.08 | 0.06 | -0.05 | -0.09 | 0 | 0.11 | 0.04 |
| $X_{13}$ | -0.07 | 0.06 | -0.07 | -0.13 | -0.09 | -0.08 | -0.12 | -0.08 | -0.09 | -0.06 | -0.1 | -0.1 | 0 | -0.07 |
| $X_{14}$ | -0.07 | 0.06 | -0.07 | -0.13 | -0.08 | 0.03 | -0.12 | -0.08 | -0.08 | -0.06 | -0.1 | -0.12 | -0.06 | 0 |

TABLE II
EXEMPLARY WEIGHTS MATRIX FOR THE MAP LEARNED WITH THE USE OF RCGA

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 0 | -0.99 | 0 | 0.99 | -0.4 | 0 | 0 | 0 | 0.85 | 0.58 | 0.37 | 0.03 | 0.1 | 0.01 |
| $X_2$ | -0.31 | 0 | 0.11 | -0.7 | -0.22 | -0.77 | 0.41 | -0.99 | 0.93 | -0.31 | 0 | 0 | 0.01 | 0.05 |
| $X_3$ | -0.4 | 0 | 0 | 0.16 | 0.65 | -0.73 | -0.07 | 0 | 0.17 | 0.34 | -0.94 | 0.09 | -0.03 | 0 |
| $X_4$ | 0 | -0.19 | 0.91 | 0 | -0.43 | -0.05 | 0.21 | 0.03 | 0.98 | 0.82 | 0 | 0 | 0 | -0.79 |
| $X_5$ | -0.11 | 0 | 0.41 | 0 | 0 | 0 | 0 | 0.2 | -0.35 | 0.08 | 0 | 0.1 | -0.1 | -0.08 |
| $X_6$ | 0.55 | 0.93 | -0.72 | -0.15 | 0.86 | 0 | -0.73 | -0.52 | 0.59 | -0.01 | 0 | 0.16 | 0 | 0 |
| $X_7$ | 0 | 0.23 | 0.64 | -0.19 | -0.27 | 0 | 0 | 0.45 | 0.5 | 0 | 0 | 0.11 | 0.1 | 0 |
| $X_8$ | -0.81 | 0 | 0.65 | 0.27 | 0.89 | -0.57 | -0.39 | 0 | 0.58 | 0.47 | 0.09 | -0.4 | 0 | 0.24 |
| $X_9$ | 0.83 | -0.1 | 0 | 0.42 | 0 | -0.27 | -0.68 | 0.01 | 0 | 0.08 | 0 | 0 | -0.68 | 0.24 |
| $X_{10}$ | 0.91 | 0 | 0 | 0 | 0.14 | -0.86 | 0.18 | -0.92 | 0 | 0 | 0 | -1 | -0.28 | -0.87 |
| $X_{11}$ | -0.15 | 0 | 0 | 0.49 | 0.21 | -0.58 | 0 | 0.77 | 0 | -0.94 | 0 | 0 | -0.55 | -0.75 |
| $X_{12}$ | 0.59 | 0.63 | -0.5 | 0 | -0.2 | 0.08 | 0.33 | 0 | -0.27 | 0 | 0 | 0 | 0.42 | -0.11 |
| $X_{13}$ | 0 | 0 | 0.88 | 0 | -0.77 | 0.8 | -0.47 | -0.64 | 0 | 0.45 | 0.51 | 0 | 0 | 0 |
| $X_{14}$ | -0.53 | 0.6 | 0.04 | 0.03 | -0.46 | 0 | 0.32 | -0.27 | -0.6 | 0 | 0.85 | 0 | 0.03 | 0 |

TABLE III
EXEMPLARY WEIGHTS MATRIX FOR THE MAP LEARNED WITH THE USE OF SOGA

| | $X_4$ | $X_5$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ |
|---|---|---|---|---|---|---|---|---|
| $X_4$ | 0 | 0 | -0.95 | 0.27 | 0.85 | -0.58 | 0 | -0.6 |
| $X_5$ | -0.51 | 0 | 0 | 0.9 | -0.13 | 0 | -0.38 | -0.01 |
| $X_9$ | 0.89 | 0.68 | 0 | 0.16 | -0.76 | -0.36 | 0 | -0.77 |
| $X_{10}$ | 0.79 | 0.11 | -0.99 | 0 | 0 | -0.15 | -0.12 | 0 |
| $X_{11}$ | 0.72 | 0.16 | 0 | 0 | 0 | -0.63 | -0.41 | -0.51 |
| $X_{12}$ | 0.19 | 0.55 | 0 | 0 | 0 | 0 | 0 | 0 |
| $X_{13}$ | 0.97 | 0 | 0 | -0.22 | 0.38 | -0.57 | 0 | 0 |
| $X_{14}$ | -0.35 | 0 | 0 | 0.66 | 0.56 | 0.25 | -0.29 | 0 |

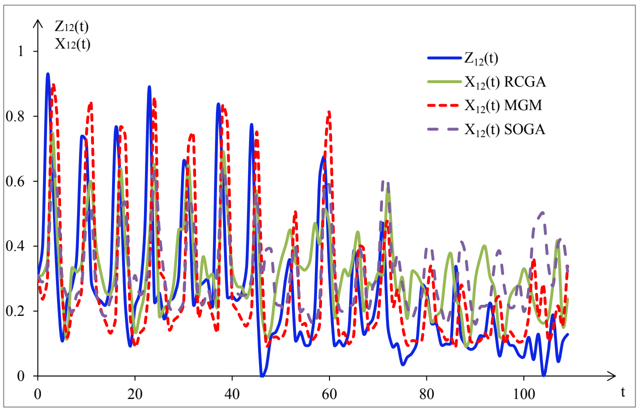


Fig. 7. Obtained values $X_{12}(t)$ and the desired values $Z_{12}(t)$ during testing

The lowest values of the objective function were obtained for the FCMs learned with the use of the Structure Optimization Genetic Algorithm. The minimum of $f_c = 471$ was achieved for the following parameters: $P = 100$, $L = 200$, ranking selection, uniform crossover, Mühlenbein's mutation, $a = 10$, $h_{max} = 0.999$, $c = 3$, $b_1 = 0.1$, $b_2 = 0.01$. Moreover, the second case of SOGA gives value of $MSE$ lowest than Multi-Step Gradient Method and very similar to Real-Coded Genetic Algorithm. Also, from the fig. 6, we can say that bike-sharing rental process is highly related to holiday, day of the week, feeling temperature, humidity and wind speed. The results show the superiority of the SOGA. It allows to significantly simplify the FCM model by selecting the most important for prediction task concepts (e.g. 8 out of 14 possible) and connections (e.g. 41 out of 182 possible) with keeping the low values of $MSE$.

TABLE IV
CHOSEN RESULTS OF ANALYSIS OF THE MGM, RCGA, SOGA

| Method | Learning parameters | $MSE$ | $n_r$ | $n_c$ | $f_c$ |
|---|---|---|---|---|---|
| MGM | $m_1 = 1$, $m_2 = 0$,<br>$\alpha_0 = 1.4$, $\alpha_1 = -0.4$,<br>$\beta_0 = 10$,<br>$\lambda_0 = 100$,<br>$e = 0.001$, $c = 5$ | 0.0429 | 174 | 14 | 743 |
| MGM | $m_1 = 1$, $m_2 = 0$,<br>$\alpha_0 = 1.5$, $\alpha_1 = -0.5$,<br>$\beta_0 = 50$,<br>$\lambda_0 = 100$,<br>$e = 0.001$, $c = 3$ | 0.0418 | 179 | 14 | 737 |
| MGM | $m_1 = 1$, $m_2 = 0$,<br>$\alpha_0 = 1.5$, $\alpha_1 = -0.5$,<br>$\beta_0 = 10$,<br>$\lambda_0 = 10$,<br>$e = 0.001$, $c = 5$ | 0.043 | 181 | 14 | 751 |
| RCGA | $P = 100$, $L = 500$,<br>$a = 100$,<br>$fitness_{max} = 0.999$,<br>$c = 5$, random mutation,<br>$P_c = 0.8$, $P_m = 0.2$ | 0.0366 | 141 | 14 | 647 |
| RCGA | $P = 100$, $L = 200$,<br>$a = 10$,<br>$fitness_{max} = 0.999$,<br>$c = 3$, Mühlenbein's mutation,<br>$P_c = 0.5$, $P_m = 0.1$ | 0.0333 | 127 | 14 | 601 |
| RCGA | $P = 100$, $L = 500$,<br>$a = 100$,<br>$fitness_{max} = 0.999$,<br>$c = 3$,random mutation,<br>$P_c = 0.5$, $P_m = 0.1$ | 0.0533 | 135 | 14 | 808 |
| SOGA | $P = 100$, $L = 500$,<br>$a = 100$, $b_1$=0.2, $b_2$=0.05,<br>$fitness_{max} = 0.999$,<br>$c = 5$, random mutation,<br>$P_c = 0.8$, $P_m = 0.2$ | 0.0463 | 37 | 8 | 580 |
| SOGA | $P = 100$, $L = 200$,<br>$a = 10$, $b_1$=0.1, $b_2$=0.01,<br>$fitness_{max} = 0.999$,<br>$c = 3$, Mühlenbein's mutation,<br>$P_c = 0.5$, $P_m = 0.1$ | 0.035 | 41 | 8 | **471** |
| SOGA | $P = 100$, $L = 500$,<br>$a = 100$, $b_1$=0.2, $b_2$=0.005,<br>$fitness_{max} = 0.999$,<br>$c = 3$,random mutation,<br>$P_c = 0.5$, $P_m = 0.1$ | 0.046 | 19 | 6 | 539 |

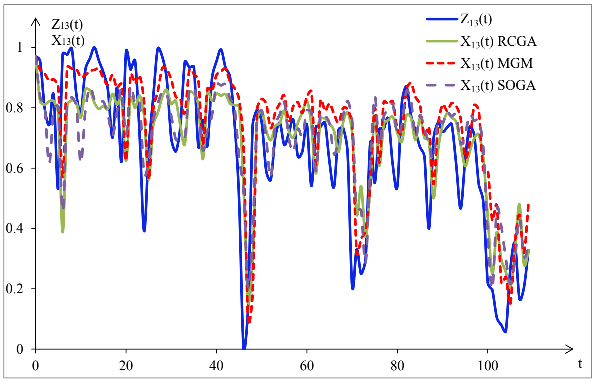


Fig. 8. Obtained values $X_{13}(t)$ and the desired values $Z_{13}(t)$ during testing

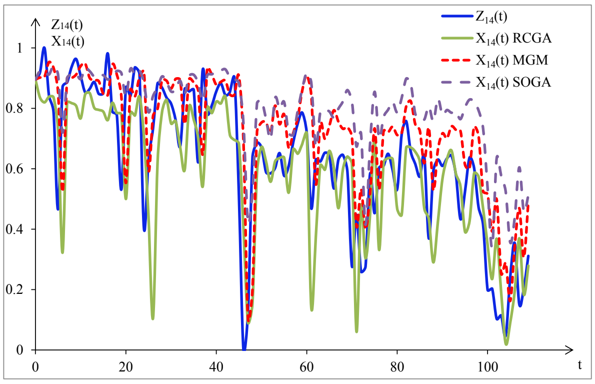


Fig. 9. Obtained values $X_{14}(t)$ and the desired values $Z_{14}(t)$ during testing

## VI. CONCLUSION

In this paper we present a new approach for fuzzy cognitive maps learning allowing selection of the most important for the analyzed tasks concepts and connections between them. The proposed Structure Optimization Genetic Algorithm is described together with well-known methods for FCM learning. Comparative analysis of SOGA, RCGA and MGM was performed on the example of prediction of count of rented bikes. Simulation research was done in ISEMK. Selected results of simulation analysis of the developed algorithms performance are presented. The obtained results show that the proposed approach can significantly reduce the size of FCM by selecting the most important concepts and connections between them. The SOGA algorithm seems to be promising and effective method for modeling complex decision support systems based on fuzzy cognitive maps. There are plans of further analysis of the use of the Structure Optimization Genetic Algorithm.

## REFERENCES

[1] S. Ahmadi, N. Forouzideh, S. Alizadeh, and E. Papageorgiou, "Learning Fuzzy Cognitive Maps using Imperialist Competitive Algorithm," *Neural Computing and Applications*, in press, http://dx.doi.org/10.1007/s00521-014-1797-4.

[2] M. Amer, A. J. Jetter, and T. U. Daim, "Scenario planning for the national wind energy sector through Fuzzy Cognitive Maps," *Proceedings of PICMET'13*, 2013, pp. 2153–2162.

[3] J. Arabas, *Lectures on genetic algorithms*, WNT, Warsaw, 2001.

[4] A. Buruzs, M. F. Hatwágner, A. Torma, and L. T. Kóczy, "Expert Based System Design for Integrated Waste Management," *International Scholarly and Scientific Research & Innovation*, vol. 8, no. 12, 2014, pp. 685–693.

[5] J. P. Carvalho, "On the semantics and the use of fuzzy cognitive maps and dynamic cognitive maps in social sciences," *Fuzzy Sets and Systems*, vol. 214, 2013, pp. 6–19, http://dx.doi.org/10.1016/j.fss.2011.12.009.

[6] H. Fanaee-T and J. Gama, "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence*, Springer Berlin Heidelberg 2013, pp. 1–15, http://dx.doi.org/10.1007/s13748-013-0040-3.

[7] W. Froelich, E.I. Papageorgiou, M. Samarinasc, and K. Skriapasc, "Application of evolutionary fuzzy cognitive maps to the long-term prediction of prostate cancer," *Applied Soft Computing*, vol. 12, 2012, pp. 3810–3817, http://dx.doi.org/10.1016/j.asoc.2012.02.005.

[8] W. Froelich and J. Salmeron, "Evolutionary learning of fuzzy grey cognitive maps for the forecasting of multivariate, interval-valued time series," *International Journal of Approximate Reasoning*, vol. 55, 2014, pp. 1319–1335, http://dx.doi.org/10.1016/j.ijar.2014.02.006.

[9] D. B. Fogel, *Evolutionary Computation. Toward a new philosophy of machine inteligence* 3rd edition, John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

[10] L. Grad, “An example of feed forward neural network structure optimisation with genetic algorithm,” *BIULETYN INSTYTUTU AUTOMATYKI I ROBOTYKI*, no. 23, 2006, pp. 31–41.

[11] F. Herrera, M. Lozano, and J. L. Verdegay, “Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioural Analysis,” *Artificial Intelligence Review*, vol. 12, 1998, pp. 265–319, http://dx.doi.org/10.1023/A:1006504901164.

[12] W. Homenda, A. Jastrzebska, and W. Pedrycz, “Nodes Selection Criteria for Fuzzy Cognitive Maps Designed to Model Time Series,” *Advances in Intelligent Systems and Computing*, vol. 323, 2015, pp. 859–870, http://dx.doi.org/10.1007/978-3-319-11310-4_75.

[13] W. Homenda, A. Jastrzebska, and W. Pedrycz, “Time Series Modeling with Fuzzy Cognitive Maps: Simplification Strategies. The Case of a Posteriori Removal of Nodes and Weights,” *Lecture Notes in Computer Science LNCS*, vol. 8838, 2014, pp. 409–420, http://dx.doi.org/10.1007/978-3-662-45237-0_38.

[14] A. Jastriebow and K. Piotrowska, “Simulation analysis of multistep algorithms of relational cognitive maps learning,” in: A. Yastrebov, B. Kuźmińska-Sołśnia and M. Raczyńska (Eds.) *Computer Technologies in Science, Technology and Education*, Institute for Sustainable Technologies - National Research Institute, Radom, 2012, pp. 126–137.

[15] A. Jastriebow and K. Poczęta, “Analysis of multi-step algorithms for cognitive maps learning,” *BULLETIN of the POLISH ACADEMY of SCIENCES TECHNICAL SCIENCES*, vol. 62, Issue 4, 2014, pp. 735–741, http://dx.doi.org/10.2478/bpasts-2014-0079.

[16] B. Kosko, “Fuzzy cognitive maps,” *International Journal of Man-Machine Studies*, vol. 24, no.1, 1986, pp. 65–75, http://dx.doi.org/10.1016/S0020-7373(86)80040-2.

[17] W. Lu, W. Pedrycz, X. Liu, J. Yang, and P. Li, “The modeling of time series based on fuzzy information granules,” *Expert Systems with Applications*, vol. 41, 2014, pp. 3799–3808, http://dx.doi.org/10.1016/j.eswa.2013.12.005.

[18] E. I. Papageorgiou , “Fuzzy Cognitive Maps for Applied Sciences and Engineering From Fundamentals to Extensions and Learning Algorithms,” *Intelligent Systems Reference Library*, vol. 54, Springer Verlag, 2014.

[19] E. I. Papageorgiou and W. Froelich, “Multi-step prediction of pulmonary infection with the use of evolutionary fuzzy cognitive maps,” *Neurocomputing*, vol. 92, 2012, pp. 28–35, http://dx.doi.org/10.1016/j.neucom.2011.08.034.

[20] E. I. Papageorgiou, K. E. Parsopoulos, C. D. Stylios, P. P. Groumpos, and M. N. Vrahtis, “Fuzzy Cognitive Maps Learning Using Particle Swarm Optimization,” *Journal of Intelligent Information Systems*, 25:1, 2005, pp. 95–121, http://dx.doi.org/10.1007/s10844-005-0864-9.

[21] E. I. Papageorgiou, K. Poczęta and C. Laspidou, “Application of Fuzzy Cognitive Maps to Water Demand Prediction,” *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Istanbul, Turkey, in press.

[22] E. I. Papageorgiou and J. L. Salmeron, “A Review of Fuzzy Cognitive Maps Research during the last decade,” *IEEE Transactions on Fuzzy Systems*, vol.21 , Issue: 1, 2013, pp. 66–79, http://dx.doi.org/10.1109/TFUZZ.2012.2201727.

[23] G. A. Papakostas, D. E. Koulouriotis, A. S. Polydoros, and V. D. Tourassis, “Towards Hebbian learning of Fuzzy Cognitive Maps in pattern classification problems," *Expert Systems with Applications*, vol. 39, 2012, pp. 10620–10629, http://dx.doi.org/10.1016/j.eswa.2012.02.148.

[24] K. Piotrowska, “Intelligent expert system based on cognitive maps,” *STUDIA INFORMATICA*, vol. 33, no 2A (105), 2012, pp. 605–616.

[25] K. Poczęta and A. Yastrebov, “Analysis of Fuzzy Cognitive Maps with Multi-Step Learning Algorithms in Valuation of Owner-Occupied Homes,” *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Beijing, China, 2014, pp.1029–1035, http://dx.doi.org/10.1109/FUZZ-IEEE.2014.6891587.

[26] J. L. Salmeron, “Fuzzy cognitive maps for artificial emotions forecasting," *Applied Soft Computing*, vol. 12, 2012, pp. 3704–3710, http://dx.doi.org/10.1016/j.asoc.2012.01.015.

[27] G. Słoń, “Application of Models of Relational Fuzzy Cognitive Maps for Prediction of Work of Complex Systems," *Lecture Notes in Artificial Intelligence LNAI*, vol. 8467, Springer Verlag, 2014, pp. 307–318, http://dx.doi.org/10.1007/978-3-319-07173-2_27.

[28] W. Stach, L. Kurgan, W. Pedrycz, and M. Reformat, “Genetic learning of fuzzy cognitive maps,” *Fuzzy Sets and Systems*, vol. 153, no. 3, 2005, pp. 371–401, http://dx.doi.org/10.1016/j.fss.2005.01.009.

DOI: 10.15439/2015F305
ACSIS, Vol. 5

# Innovative GPU accelerated algorithm for fast minimum convex hulls computation

Artem Potebnia
Email: potebnia@mail.ua
Kyiv National Taras Shevchenko University
in Kyiv, Ukraine

Sergiy Pogorilyy
Email: sdp@univ.net.ua
Kyiv National Taras Shevchenko University
in Kyiv, Ukraine

***Abstract*—Innovative algorithm for forming graph minimum convex hulls using the GPU is proposed. High speed and linear complexity of this method are achieved by distribution of the graph's vertices into separate units and their filtering. The key factor for improving the performance of innovative algorithm is the massively-parallel implementation of local hulls formation using video accelerators. A computational process is controlled by means of auxiliary matrices. A number of experimental studies of the algorithm have been carried out, and its suitability for application in the hull processing for large-scale problems has been demonstrated. The speed of the new method is 10 – 20 times higher compared to using functions of the professional mathematical package Wolfram Mathematica.**

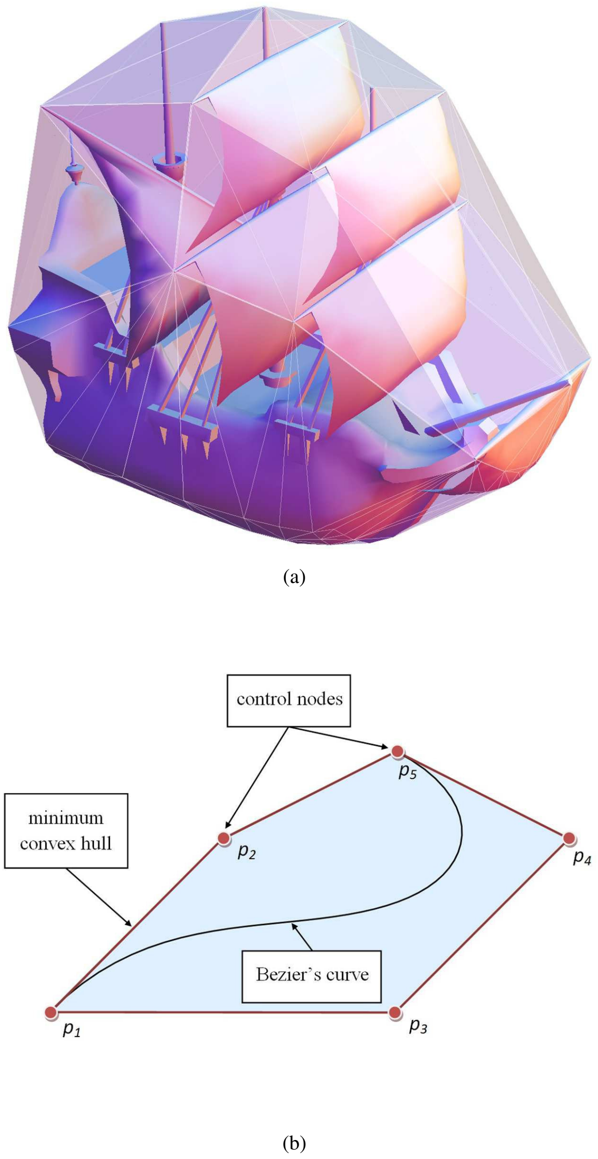
***Index Terms*—Minimum convex hull, CPU+GPU hybrid system, GPGPU technology, High-performance computing, CUDA, Graph**

## I. Introduction

FINDING the minimum convex hull (MCH) of the graph's vertices is a fundamental problem in many areas of modern research [7]. The solution of this task involves the formation of the minimum convex set containing all the nodes present in the graph (Fig. 1a). It is known that MCH is a common tool in computer-aided design and computer graphics packages [21]. For example, Bezier's curves used in *Adobe Photoshop*, *GIMP* and *CorelDraw* for modeling smooth lines fully lie in the convex hull of their control nodes (Fig. 1b). This feature greatly simplifies finding the points of intersection between curves and allows their transformation (moving, scaling, rotating, etc.) by appropriate control nodes [23]. The formation of some fonts and animation effects in the *Adobe Flash* package also uses splines composed of quadratic Bezier's curves [8].

It should be noted that convex hulls are used in *Geographical Information Systems* and routing algorithms in determining the optimal ways for avoiding obstacles. The paper [1] offers the methods for solving complex optimization problems using them.

Last decades are associated with rapid data volume growth in research processed by the information systems [19]. According to IBM, about 15 petabytes of new information are created daily in the world [14]. Therefore, in modern science, there is a separate area called *Big Data* related to the study of large data sets [13]. However, most of the known algorithms for MCH construction have time complexity $O(nlogn)$, making them useless when forming solutions to large-scale graphs. Therefore, there is a need to develop efficient algorithms with the complexity close to linear $O(n)$.



(a)

(b)

Fig. 1. Examples of the minimum convex hulls

978-83-60810-66-8/$25.00©2015, IEEE

TABLE I
COMPARISON OF THE COMMON ALGORITHMS FOR MCH CONSTRUCTION

| **Algorithm** | **Complexity** | **Parallel versions** | **Multidimensional cases** |
|---|---|---|---|
| Jarvis's march | $O(nh)$, where $h$ is a number of points on MCH | + | + |
| Graham's Scan | $O(nlogn)$ | - | - |
| QuickHull | $O(nlogn)$, in the worst case – $O(n^2)$ | + | + |
| Divide and Conquer | $O(nlogn)$ | + | + |

It is known that Wolfram Mathematica is one of the most powerful mathematical tools for the high performance computing. Features of this package encapsulate a number of algorithms and, depending on the input parameters of the problem, select the most productive ones [20]. Therefore, Wolfram Mathematica 9.0 is used to track the performance of the algorithm proposed in this article.

In recent years, CPU+GPU hybrid systems (GPGPU technology) allowing for a significant acceleration of computations have become widespread. Unlike CPU, consisting of several cores, the graphics processor is a multicore structure and the number of its components is measured in hundreds [16]. In this case, the sequential steps of algorithm are executed on the CPU, while its parallel parts are implemented on the GPU [20]. For example, the latest generation of NVIDIA Fermi GPUs contains 512 computing cores, allowing for the introduction of new algorithms with large-scale parallelism [9]. Thus, the usage of NVIDIA GPU ensures the conversion of standard workstations to powerful supercomputers with cluster performance [17].

The paper goal is to develop a high-speed algorithm for finding the minimum convex hulls using GPU. The time complexity of the proposed method is close to linear. The optimal values of algorithm's parameters, with which the usage of GPU resources is the most effective, are established in this paper.

## II. A REVIEW OF ALGORITHMS FOR FINDING THE MINIMUM CONVEX HULLS

Despite intensive research, which lasted for the past 40 years, the problem of developing efficient algorithms for MCH formation is still open. The main achievement is the development of numerous methods based on the extreme points determination of the original graph and the link establishment among them [6]. These techniques include the *Jarvis's march* [15], *Graham's Scan* [11], *QuickHull* [4], *Divide and Conquer* algorithm and many others. The main features of their practical usage are given in Table I.

For parallelization the *Divide and Conquer* algorithm is the most suitable. It provides a random division of the original vertex set into subsets, formation of partial solutions and their connection to the general hull [21]. Although the hull connection phase has linear complexity, it leads to a significant slowdown of the algorithm, and as a result, to the unsuitability of its application in the hull processing for large-scale graphs.

Chan's algorithm, which is a combination of slower algorithms, has the lowest time complexity $O(nlogh)$. However, it can work by the known number of vertices contained in the hull [3]. Therefore, currently, its usage in practice is limited [5].

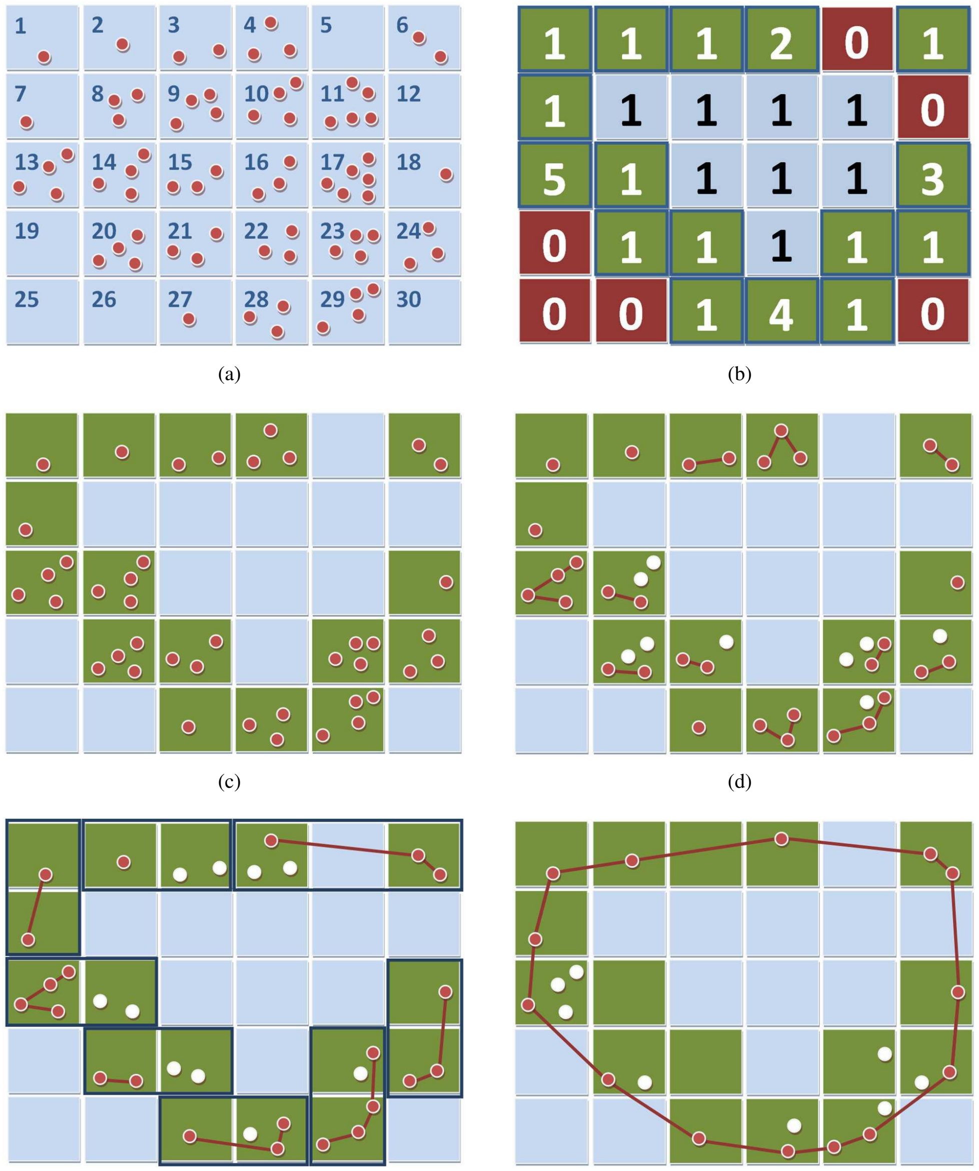
Study [2] gives a variety of acceleration tools for known MCH formation algorithms by cutting off the graph's vertices falling inside an octagon or rectangle and appropriate reducing the dimensionality of the original problem. The paper [12] suggests numerous methods of convex hull approximate formation, which have linear complexity. Such algorithms are widely used for tasks where speed is a critical parameter. But linearithmic time complexity of the fastest exact algorithms demonstrates the need for the introduction of new high-speed methods of convex hulls formation for large-scale graphs.

## III. INNOVATIVE ALGORITHM FOR FORMATION OF THE CONVEX HULLS

We shall consider non-oriented graph $G = (V, E)$. The new algorithm provides a division of the original graph's vertex set into a set of output units $U = \langle U_1, U_2, ..., U_n \rangle$, $U_i \subseteq V$. However, unlike the *Divide and Conquer* method, this division is not random, but it is based on the spatial distribution of vertices. All nodes of the graph should be distributed by the formed subsets, i.e. $\bigcup_{i=1}^{n} U_i = V$. This allows the presence of empty units, which do not contain vertices. Additionally, the condition of orthogonality division is met, i.e. one vertex cannot be a part of the different blocks: $U_i \cap U_j = \varnothing$, $\forall i \neq j$. Fig. 2a shows an example of division taking into account the above requirements.

The next stage of the proposed algorithm involves the formation of an auxiliary matrix based on the distribution of nodes by units. The purpose of this procedure is the primary filtration of the graph's vertices, which provides a significant decrease in the original problem dimensionality. In addition, the following matrices define the sets of blocks for the calculation in the subsequent stages of the algorithm and the sequence of their connection to the overall result. An auxiliary matrix formation involves the following operations:

1) Each block of the original graph must be mapped to one cell of the supporting matrix. Accordingly, the dimension of this matrix is $n \times m$, where $n$ and $m$ are the numbers of blocks allocated by the relevant directions.
2) The following operations provide the necessary coding of matrix's cells. Thus, the value of cell $c_{i,j}$ is zero if the corresponding block $U_{i,j}$ of original graph contains no vertices. Coding of blocks that contain extreme nodes (the highest, rightmost, lowest and leftmost points) of



(a) (b)

(c) (d)

(e) (f)

Fig. 2. Example of the algorithm execution

a given set is important for the algorithm. Appropriate cells are filled with numbers from 2 to 5. Other units that are filled, and contain no extreme peaks, shall be coded with ones in auxiliary matrix.

3) Further, primary filtration of allocated blocks is carried out using the filled matrix. Empty subsets thus shall be excluded from consideration. Blocks containing extreme vertices shall determine the graph division into parts for which the filtration procedure is applied. We shall consider the example of the block selection for the section limited with cells 2 – 3. If $c_{i,j} = 2$, then the next non-zero cell is searched by successive increasing of $j$. In their absence, the next matrix's row $i + 1$ is reviewed. Selection of blocks is completed, if the value of another chosen cell is $c_{i,j} = 3$. The study of the other graph's parts is based on a similar principle.

Partial solutions are formed for selected blocks. Such operations require the formation of fragments rather than full-scale hulls that provides secondary filtration of the graph's vertices. The last step of the algorithm involves the connection of partial solutions to the overall result. Thus the sequential merging of local fragments is done on a principle similar to *Jarvis's march*. It should be noted that at this stage filtration mechanism leads to a significant reduction in the dimensionality of the original problem. Therefore, when processing the hulls for large graphs, combination operations constitute about 0.1% of the algorithm total operation time.

We shall consider the example of this algorithm execution. Let the set of the original graph's vertices have undergone division into 30 blocks (Fig. 2a). Auxiliary matrix calculated for this case is given in Fig. 2b. After application of primary filtration, only 57% of the graph's nodes were selected for investigation at the following stages of the algorithm (Fig. 2c). The next operations require the establishment of local hulls (Fig. 2d) and their aggregations are given in Fig. 2e. After performing of pairwise connections, this operation is applied repeatedly until a global convex hull is obtained (Fig. 2f).

## IV. The development of hybrid CPU-GPU algorithm

We know that the video cards have much greater processing power compared to the central processing elements. GPU computing cores work simultaneously, enabling to use them to solve problems with the large volume of data. CUDA (*Compute Unified Device Architecture*), the technology created by NVIDIA, is designed to increase the productivity of conventional computers through the usage of video processors computing power [22].

CUDA architecture is based on SIMD (*Single Instruction Multiple Data*) concept, which provides the possibility to process the given set of data via one function. Programming model provides for consolidation of threads into blocks, and blocks – into a grid, which is performed simultaneously. Accordingly, the key to effective usage of GPU hardware capabilities is algorithm parallelization into hundreds of blocks performing independent calculations on the video card [24].

It is known that GPU consists of several clusters. Each of them has a texture unit and two streaming multiprocessors, each containing 8 computing devices and 2 superfunctional units [10]. In addition, multiprocessors have their own distributed memory resources (16 KB) that can be used as a programmable cache to reduce delays in data accessing by computing units [22]. From these features of CUDA architecture, it may be concluded that it is necessary to implement massively-parallel parts of the algorithm on the video cards, while sequential instructions must be executed on the CPU. Accordingly, the stage of partial solutions formation is suitable for implementation on the GPU since the operations for each of the numerous blocks are carried out independently.

It is known that function designed for executing on the GPU is called a kernel. The kernel of the innovative algorithm contains a set of instructions to create a local hull of any selected subset. In this case, distinguishing between the individual subtasks is realized only by means of the current thread's number. Thus, the developed hybrid algorithm has the following execution stages:

1) Auxiliary matrix is calculated on the CPU. The program sends cells' indexes that have passed the primary filtration procedure and corresponding sets of vertices to the video card.
2) Based on the received information, particular solutions are formed on the GPU, recorded to its global memory and sent to the CPU.
3) Further, the procedure of their merging is carried out and the overall result is obtained.

It should be noted that an important drawback of hybrid algorithms is the need to copy data from the CPU to the GPU and vice versa, which leads to significant time delays [16], [18]. Communication costs are considerably reduced by means of filtration procedure.

When developing high-performance algorithms for the GPU it is important to organize the correct usage of the memory resources. It is known that data storage in the global video memory is associated with significant delays in several hundred GPU cycles. Therefore, in the developed algorithm, the global memory is used only as a means of communication between the processor and video card. The results of intermediate calculations for each of the threads are recorded in the shared memory, access speed of which is significantly higher and is equal to 2 – 4 cycles.

## V. Experimental studies of the proposed algorithm

In the current survey, experimental tests were run on a computer system with an Intel Core i7-3610QM processor (2.3 GHz), 8 GB RAM and DDR3-1600 NVIDIA GeForce GT 630M video card (2GB VRAM). This graphics accelerator contains 96 CUDA kernels, and its clock frequency is 800 MHz.

It is known that the number of allocated blocks increases linearly with enhancing of processed graphs dimensionality. The complexity of calculating the relevant auxiliary matrices grows by the same principle. The stages of multi-step filtration and local hulls construction provide a significant simplification of final connection procedure. Thus, the complexity of the developed algorithm is linear $O(n)$ for uniformly distributed data.

MCH instances composed of all graph's vertices are the worst for investigation. In this case, the filtration operations do not provide the required acceleration and the algorithm complexity is equal to $O(nlogn)$. However, these examples have purely theoretical significance and almost never occur in practice.

Fig. 3 shows the dependence of the innovative algorithm execution time on the graph dimensionality and the number of vertices in the selected blocks. These results confirm the linear complexity of the proposed method. In addition, it is important to set the optimal dimensionality of the subsets allocated in the original graph. A selection of smaller blocks (up to 1000 nodes) leads to a dramatic increase in the algorithm operation time.

This phenomenon is caused by the significant enhancing of the auxiliary matrices dimensionality, making it difficult to control the computing process (Fig. 4). Per contra, the allocation of large blocks (over 5000 vertices) is associated with the elimination of the massive parallel properties, mismanagement of the video card resources, and as a consequence, increasing of the algorithm execution time. Thus, the highest velocity of the proposed method is observed for intermediate values of

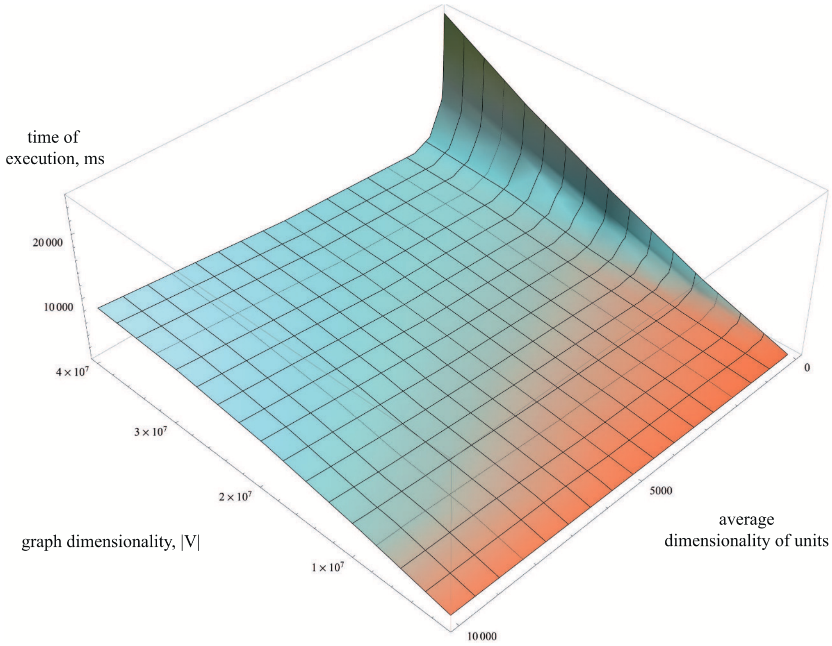


Fig. 3. Dependence of the innovative algorithm performance on the graph dimensionality and the number of vertices in the selected blocks

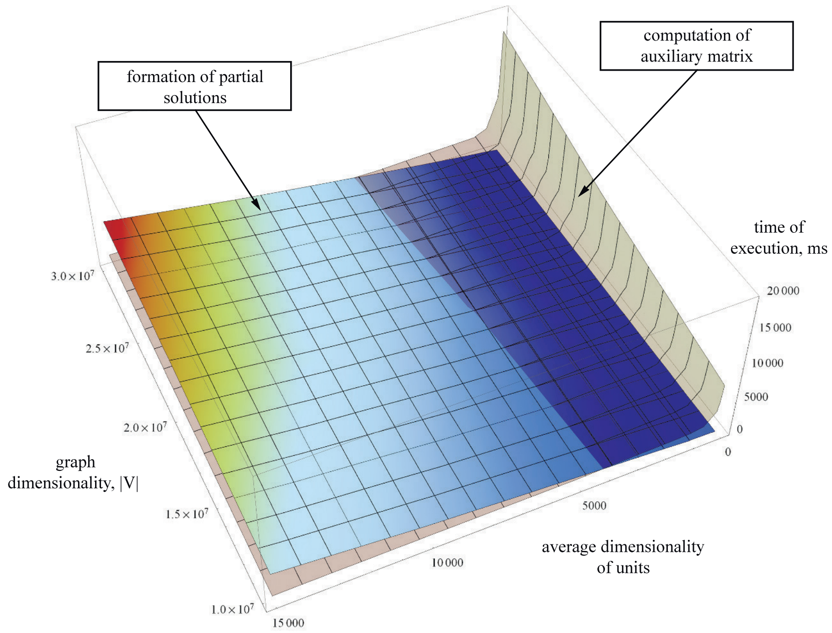
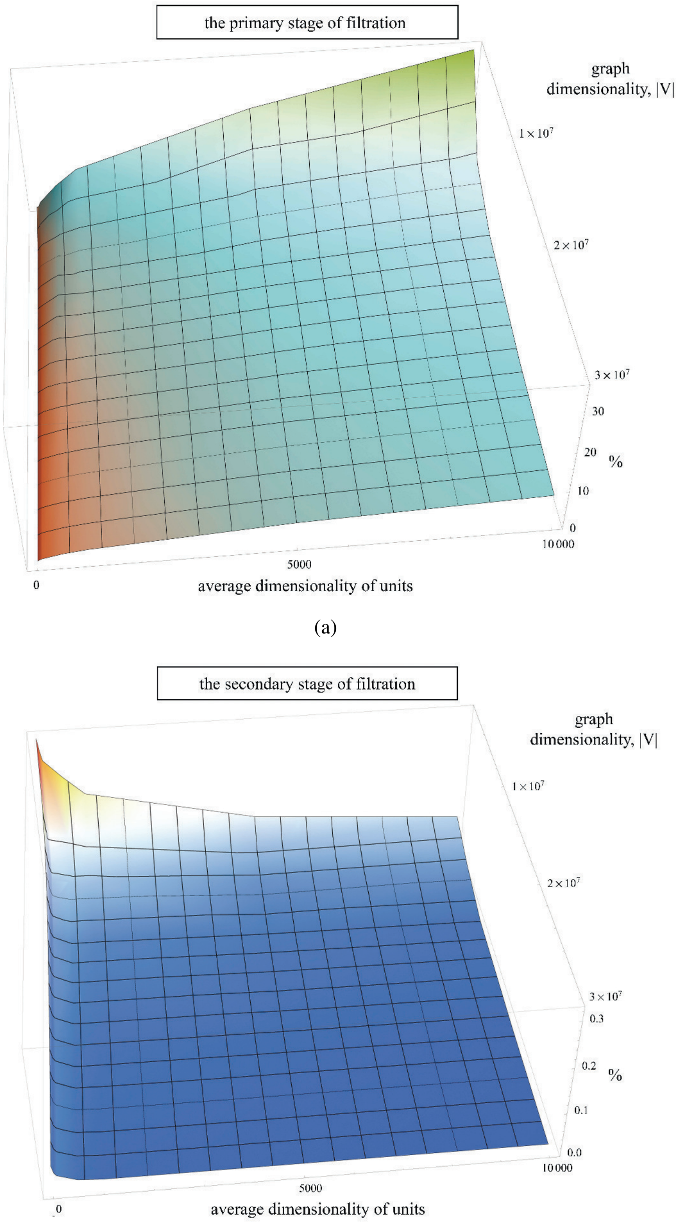


Fig. 4. Dependences of the various stages performance on the graph dimensionality and the number of vertices in the selected blocks

the blocks dimensionality (1000 – 5000 vertices). In this case, auxiliary matrices are relatively small, and the second stage of the algorithm preserves the properties of massive parallelism.

One of the most important means to ensure the algorithm's high performance is the multi-step filtration of the graph's vertices. Fig. 5a shows the dependence of the primary selection quality on the dimensionality of the original problem and allocated subsets. These results show that such filtration is the most efficient with the proviso that the graph's vertices are distributed into small blocks. Furthermore, the number of selected units increases with the raising of the problem's size, providing rapid solutions to graphs of extra large dimensionality. By virtue of a riddance from the discarded blocks, the following operations of the developed algorithm are applied only to 1 – 3% of the initial graph's vertices.

However, the results of the secondary filtration (Fig. 5b) are the opposite. In this case, the highest quality of the selection is obtained on the assumption that the original



(a)

(b)

Fig. 5. The influence of filtration procedure over the reduction in the problem's size

vertices are grouped into large subsets. Withal, the secondary filtration is much slower than the primary procedure, so the most effective selection occurs at intermediate values of the blocks dimensionality. As a result of these efforts, only 0.05 – 0.07% of the initial graph's vertices are involved in the final operations of the proposed algorithm.

In order to determine the efficiency of the developed algorithm, its execution time has been compared with the built-in tools of the mathematical package *Wolfram Mathematica 9.0*. All choice paired comparison tests were conducted for randomly generated graphs. The MCH formation in *Mathematica* package is realized by the instrumentality of *ConvexHull[]* function, while the *Timing[]* expression is used to measure the obtained performance. The results of the performed comparison are given in Fig. 6. They imply that the new algorithm computes the hulls up to 10 – 20 times faster than *Mathematica's* standard features.

## VI. Conclusions

The paper suggests an innovative algorithm for finding the minimum convex hulls, which is based on the GPGPU technology and uses graphic accelerators. Unlike its predecessors, this algorithm is adapted to fast solving of the large-scale problems and, therefore, is suitable for using with respect to *Big Data*

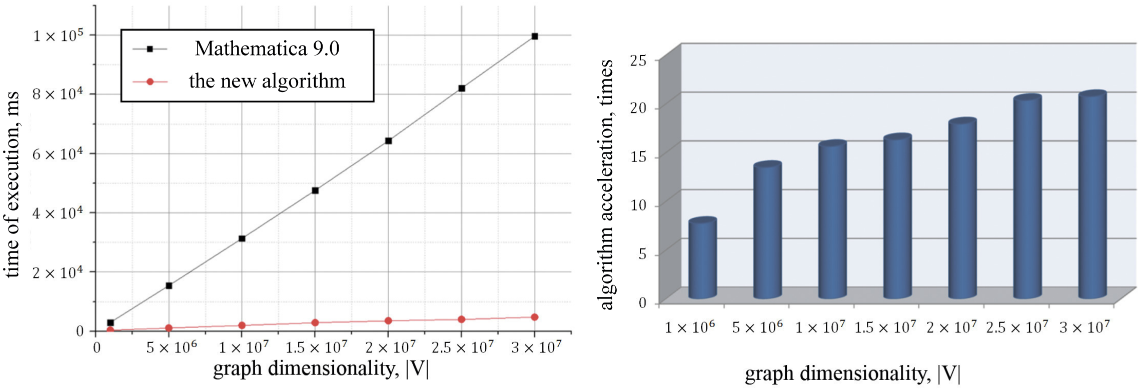


Fig. 6. A performance comparison between the new algorithm and built-in tools of the mathematical package Wolfram Mathematica 9.0

direction. Compared to the classical methods it has a number of the following benefits:

1) **High speed of operation and linear complexity.** Algorithm performance is increased by the steps of vertices' distribution into blocks, filtration and using of the auxiliary matrices. As a result, the speed of the new method is 10 – 20 times higher in contrast to the usage of professional mathematical package *Mathematica*.
2) **Massive parallelism.** Formation of partial hulls is carried out independently, which contributes to the implementation of these calculations by using graphics processors.
3) **The ability of hulls' dynamic adjustment.** When adding new vertices to the initial set, calculations are executed only for units that have undergone modification. These operations require only local updating of the convex hulls, because the results for intact parts of the original graph are invariable.
4) **The ability of generalization for the multidimensional problem instances.** In these cases, the selected subsets are the $n$-dimensional cubes to which operations of the developed method are applied.

These advantages may be the basis for the inclusion of the innovative algorithm in the professional mathematical packages to promote the high-performance computations among their users. A further direction of research is related to the development of hybrid CPU+GPU versions of this algorithm for complex systems with many processors and video cards.

## References

[1] Aardal K., Van Hoesel S., Polyhedral Techniques in Combinatorial Optimization I: Theory. Statistica Neerlandica 50, pp. 3–26, 1995. DOI: 10.1111/j.1467-9574.1996.tb01478.x.

[2] Akl S.G., Toussaint G.T., A fast convex hull algorithm. Information processing letters 7, pp. 219–222, 1978.

[3] Allison D.C.S., Noga M.T., Some performance tests of convex hull algorithms. BIT 24(1), pp. 2–13, 1984. DOI: 10.1007/BF01934510.

[4] Barber C.B., Dobkin D.P., Huhdanpaa H., The Quickhull Algorithm for Convex Hulls, ACM Trans. Math. Softw. 22(4), pp. 469–483, 1996. DOI: 10.1145/235815.235821.

[5] Chan T.M., Optimal output-sensitive convex hull algorithms in two and three dimensions. Discrete & Computational Geometry 16, pp. 361–368, 1996. DOI: 10.1007/BF02712873.

[6] Cormen T.H., Leiserson C.E., Rivest R.L., Stein C., Introduction to Algorithms, Second Edition. Section 33.3: Finding the convex hull. MIT Press, pp. 947–957, 2001.

[7] De Berg M., Cheong O., van Kreveld M., Overmars M., Computational Geometry: Algorithms and Applications. Springer-Verlag, Heidelberg, 2008. DOI: 10.1007/978-3-540-77974-2

[8] Duncan M., Applied Geometry for Computer Graphics and CAD. Springer-Verlag, London, 2005. DOI: 10.1007/b138823

[9] Glaskowsky P.N., NVIDIA's Fermi: The First Complete GPU Computing Architecture. NVIDIA, 2009.

[10] Govindaraju N.K., Larsen S., Gray J., Manocha D., A memory model for scientific algorithms on graphics processors. Proceedings of the ACM/IEEE conference on Supercomputing, 2006. DOI: 10.1109/SC.2006.2.

[11] Graham R.L., An efficient algorithm for determining the convex hull of a finite planar set, Info. Proc. Lett. 1(1), pp. 132–133, 1972. DOI: 10.1016/0020-0190(72)90045-2.

[12] Hossain M., Amin M., On Constructing Approximate Convex Hull. American Journal of Computational Mathematics 3(1A), pp. 11–17, 2013. DOI: 10.4236/ajcm.2013.31A003.

[13] IBM: Analytics: The real-world use of big data, 2013.

[14] IBM: Storage strategies that deliver business value, 2011.

[15] Jarvis R.A., On the identification of the convex hull of a finite set of points in the plane, Info. Proc. Lett. 2(1), pp. 18–21, 1973. DOI: 10.1016/0020-0190(73)90020-3.

[16] Lee C., Ro W.W., Gaudiot J.-L., Boosting CUDA Applications with CPU-GPU Hybrid Computing. International Journal of Parallel Programming 42(2), pp. 384–404, 2014. DOI: 10.1007/s10766-013-0252-y.

[17] Nickolls J., Dally W., The GPU computing era. Micro IEEE 30(2), pp. 56–69, 2010. DOI: 10.1109/MM.2010.41.

[18] Novakovic V., Singer S., A GPU-based hyperbolic SVD algorithm. BIT 51(4), pp. 1009–1030, 2011. DOI: 10.1007/s10543-011-0333-5.

[19] Pogorilyy S.D., Potebnia A.V., Formation and investigation of Kruskal's algorithm parallel scheme for shared memory systems. Scientific Papers of Donetsk National Technical University "Informatics, Cybernetics and Computer Science" 16(204), pp. 82–89, 2012. DOI: 10.5281/zenodo.16440 (in Ukrainian).

[20] Potebnia A.V., Pogorilyy S.D., Exploration of data coding methods in wireless computer networks. Proceedings of the Fourth International Conference on Theoretical and Applied Aspects of Cybernetics (TAAC), Kyiv, pp. 17–31, 2014. DOI: 10.13140/RG.2.1.3186.3844.

[21] Preparata F.P., Shamos M.I., Computational Geometry: An Introduction. Springer-Verlag, New York, 1985. DOI: 10.1007/978-1-4612-1098-6.

[22] Sanders, J., Kandrot, E., CUDA by Example: An Introduction to General-Purpose GPU Programming. Addison-Wesley Professional, 2010.

[23] Sederberg T. W., Computer aided geometric design course notes, 2011. Available: http://tom.cs.byu.edu/ 557/text/cagd.pdf.

[24] Sklodowski, P., Zorski, W., Movement Tracking in Terrain Conditions Accelerated with CUDA. Proceedings of the Federated Conference on Computer Science and Information Systems, FedCSIS 2014, Warsaw, Poland, pp. 709–717. DOI: 10.15439/2014F282.

DOI: 10.15439/2015F260
ACSIS, Vol. 5

# The density Turán problem for some 3-uniform unihypercyclic linear hypergraphs. An efficient testing algorithm

Halina Bielak
Institute of Mathematics
Maria Curie Skłodowska University
Pl. M. Curie-Skłodowskiej 5
20-031 Lublin, Poland
Email: hbiel@hektor.umcs.lublin.pl

Kamil Powroźnik
Institute of Mathematics
Maria Curie Skłodowska University
Pl. M. Curie-Skłodowskiej 5
20-031 Lublin, Poland
Email: kamil.pawel.powroznik@gmail.com

***Abstract*—Let $\mathcal{H} = (V, \mathcal{E})$ be a 3-uniform linear hypergraph with one hypercycle $\mathcal{C}_3$. We consider a blow-up hypergraph $\mathcal{B}[\mathcal{H}]$.**

**We are interested in the following problem. We have to decide whether there exists a blow-up hypergraph $\mathcal{B}[\mathcal{H}]$ of the hypergraph $\mathcal{H}$, with hyperedge densities satisfying special conditions, such that the hypergraph $\mathcal{H}$ appears in a blow-up hypergraph as a transversal. We present an efficient algorithm to decide whether a given set of hyperedge densities ensures the existence of a 3-uniform linear hypergraph $\mathcal{H}$ with hypercycle $\mathcal{C}_3$ in the blow-up hypergraph $\mathcal{B}[\mathcal{H}]$.**

**Moreover, we state some relations between roots of the multivariate matching polynomial and the inhomogeneous density Turán problem.**

***Index Terms*—blow-up hypergraph; density; Turán density problem; unicyclic hypergraph.**

## I. Introduction

Let $\mathcal{H} = (V, \mathcal{E})$ be a simple, connected and finite hypergraph with the vertex set $V$ and hyperedge set $\mathcal{E}$ (see [2]). Turán [13] stated the first results in extremal graph theory. Then many authors extended this subject and formulated similar and new Turán density problems. Many interesting results for some families of simple graphs were published in [1], [6], [7], [9], [11], [12] and [14] obtained.

In this paper we present some algorithm for testing whether a hypergraph with a given set of hyperedge densities is a factor (a transversal) of a blow-up hypergraph for some unihypercyclic hypergraphs. Our algorithm has the time complexity at most $\mathcal{O}(n^2)$, where $n$ is the number of hyperedges of the hypergraph.

Ealier Csikvári and Nagy [8] discovered an interesting algorithm for testing whether a tree with a given set of edge densities is a factor of a blow-up graph. Some generalization of their algorithm is presented in [4]. In this paper we extend this ideas to create a respective algorithm for the family of 3-uniform linear unihypercyclic hypergraphs with a hypercycle $\mathcal{C}_3$.

First we define some notions and notations. Other definitions one can find in [2], [5] and [10].

A hypergraph $\mathcal{H}$ is called *linear* if any two hyperedges intersect in at most one vertex. A hypergraph $\mathcal{H}$ is called $r$-*uniform* if each hyperedge consists of $r$ vertices.

A subhypergraph $\mathcal{P}_t$ of $\mathcal{H}$ is called *a linear hyperpath of length* $t$ if the hyperedges of $\mathcal{P}_t$ can be labelled by $e_i, 0 \leq i \leq t-1$ such that the sequence $(e_0, e_1, .., e_{t-1})$ satisfies the condition: $|e_i \cap e_j| = 1$ if and only if $|i-j| = 1$ and $e_i \cap e_j = \emptyset$ if and only if $|i-j| > 1$, where $e_i \in \mathcal{E}(\mathcal{H})$ (see Fig. 1(a)).

A subhypergraph $\mathcal{C}_t$ of $\mathcal{H}$, $t \geq 3$, is called *a linear hypercycle of length* $t$ if the hyperedges of $\mathcal{C}_t$ can be labelled by $e_i, 0 \leq i \leq t-1$ such that the sequence $(e_0, e_1, .., e_{t-1})$ satisfies the condition: $|e_i \cap e_j| = 1$ if and only if $|i-j| = 1$ or $i = 0$ and $j = t-1$ and $e_i \cap e_j = \emptyset$, $i \neq j$, in the opposite case, where $e_i \in \mathcal{E}(\mathcal{H})$ (see Fig. 2).

A 3-*uniform linear unihypercyclic hypergraph* $\mathcal{H}$ is a connected linear 3-uniform hypergraph with one hypercycle $\mathcal{C}_3$ (see Fig. 3).

The degree of the vertex $v$ in the hypergraph $\mathcal{H}$ is the number of hyperedges containing this vertex. Each vertex of degree 1 in a hypergraph $\mathcal{H}$ is called *the leaf*. We say that the hypergraph $\mathcal{H}$ is $r$-*regular* if each vertex of $\mathcal{H}$ has degree $r$. A hyperedge $e \in \mathcal{E}(\mathcal{H})$ is called *a pendant hyperedge* if it contains exactly one vertex of degree $> 1$.

A set $S \subset V(\mathcal{H})$ is called *the independent vertex set* if the subhypergraph of $\mathcal{H}$ induced by $S$ has empty set of hyperedges. The set $M \subseteq \mathcal{E}(\mathcal{H})$ is called *the matching* (or *independent hyperedge set*) in the hypergraph $\mathcal{H}$ if the subhypergraph of $\mathcal{H}$ induced by $M$ is 1-regular.

Let $\mathcal{H}$ be a 3-uniform linear hypergraph. For each vertex $i \in V(\mathcal{H})$ we associate *a cluster* $A_i$, as a set of new, independent vertices.

For a hypergraph $\mathcal{H}$ we define *a blow-up hypergraph* $\mathcal{B}[\mathcal{H}]$ of the hypergraph $\mathcal{H}$ as follows. First we replace each vertex $i \in V(\mathcal{H})$ by a cluster $A_i$ and next we create some hyperedges between the clusters $A_i$, $A_j$ and $A_k$ if $\{i, j, k\}$ is a hyperedge in $\mathcal{H}$, $i, j, k \in V(\mathcal{H})$. Equivalently each hyperedge in $\mathcal{B}[\mathcal{H}]$ has exactly one vertex from the clusters.

978-83-60810-66-8/$25.00©2015, IEEE

For any three clusters we define *a density* between them by the following formula

$$d(A_i, A_j, A_k) = \frac{e(A_i, A_j, A_k)}{|A_i||A_j||A_k|}, \tag{1}$$

where $e(A_i, A_j, A_k)$ denotes the number of hyperedges with one element of each of the clusters $A_i$, $A_j$ and $A_k$.

The hypergraph $\mathcal{H}$ is *a transversal* of $\mathcal{B}[\mathcal{H}]$ if $\mathcal{H}$ is a subhypergraph of $\mathcal{B}[\mathcal{H}]$ such that we have a homomorphism

$$\phi : V(\mathcal{H}) \to V(\mathcal{B}[\mathcal{H}])$$

for which $\phi(i) \in A_i$ for all $i \in V(\mathcal{H})$. Other terminology: $\mathcal{H}$ is *a factor* of $\mathcal{B}[\mathcal{H}]$ (see Fig. 1(b)).

A hyperedge $e = \{i, j, k\}$ of the hypergraph $\mathcal{H}$ we denote shortly by $e = ijk$.

*The homogeneous density Turán problem for* 3*-uniform linear hypergraphs* can be defined as follows. Let us determine the critical hyperedge density, denoted by $d_{crit}(\mathcal{H})$, which ensures the existence of the subhypergraph $\mathcal{H}$ of $\mathcal{B}[\mathcal{H}]$ as a transversal. Precisely, assume that all hyperedges $e = \{i, j, k\}$ in the hypergraph $\mathcal{H}$ satisfy the condition

$$d(A_i, A_j, A_k) > d_{crit}(\mathcal{H}),$$

where $i, j, k \in V(\mathcal{H})$. Then, no matter how we construct the blow-up hypergraph $\mathcal{B}[\mathcal{H}]$, it contains the hypergraph $\mathcal{H}$ as a transversal. On the other words, for any value $d < d_{crit}(\mathcal{H})$ there exists a blow-up hypergraph $\mathcal{B}[\mathcal{H}]$ such that

$$d(A_i, A_j, A_k) > d$$

for all hyperedges $ijk \in \mathcal{E}(\mathcal{H})$ which does not contain $\mathcal{H}$ as a transversal.

Moreover, we define *the inhomogeneous density Turán problem for* 3*-uniform linear hypergraphs* as follows. Let us assume that for every hyperedge $e \in \mathcal{E}(\mathcal{H})$ a density $\gamma_e$ is given. Now our task is to decide if the set of densities $\{\gamma_e\}_{e\in\mathcal{E}(\mathcal{H})}$ ensure the existence of the hypergraph $\mathcal{H}$ as a transversal or we can construct a blow-up hypergraph $\mathcal{B}[\mathcal{H}]$ such that

$$d(A_i, A_j, A_k) \geq \gamma_{ijk},$$

$\{i, j, k\} \in \mathcal{E}(\mathcal{H})$, but it does not induce the hypergraph $\mathcal{H}$ as a transversal.

This two problems has been studied in [8], [12] for simple graphs which are 2-uniform linear hypergraphs. We extend some of those results to 3-uniform linear hypergraphs with the hypercycle $\mathcal{C}_3$.

Let us recall the definition of *the multivariate matching polynomial* of the hypergraph.

Let $\mathcal{H}$ be a hypergraph and let $\underline{x_e}$ be the vector of variables $x_e$, for $e \in \mathcal{E}(\mathcal{H})$. We define *the multivariate matching polynomial* $F_{\mathcal{H}}(\underline{x_e}, t)$ of the hypergraph $\mathcal{H}$ as follows

$$F_{\mathcal{H}}(\underline{x_e}, t) = \sum_{M\in\mathcal{M}} \left( \prod_{e\in M} x_e \right) (-t)^{|M|}, \tag{2}$$

where the summation goes over all matchings of the hypergraph $\mathcal{H}$, including the empty matching (see Example 1 ).

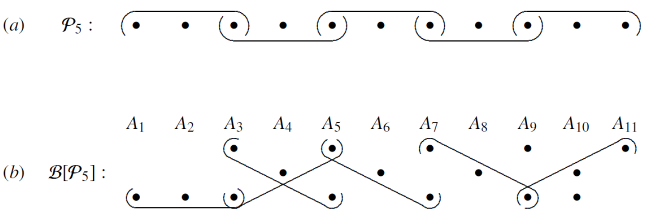


Fig. 1. A 3-uniform linear hyperpath on 5 hyperedges and a blow-up hypergraph $\mathcal{B}[\mathcal{P}_5]$ without the factor $\mathcal{P}_5$. Let $|A_i| = 1$ for $i \in \{1, 2, 4, 6, 8, 11\}$ and $|A_i| = 2$ for $i \in \{3, 5, 7, 9, 10\}$. We obtain the following densities between the clusters in $\mathcal{B}[\mathcal{P}_5]$: $d(A_1, A_2, A_3) = d(A_3, A_4, A_5) = \frac{1}{2}$, $d(A_5, A_6, A_7) = d(A_7, A_8, A_9) = d(A_9, A_{10}, A_{11}) = \frac{1}{4}$ and 0 for others. If we add the new hyperedge between clusters $A_7, A_8, A_9$, we get $d(A_7, A_8, A_9) = \frac{1}{2}$ and $\mathcal{P}_5$ as a factor.

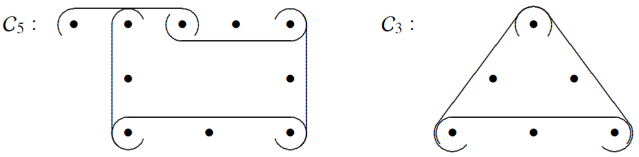


Fig. 2. A 3-uniform linear hypercycles $\mathcal{C}_5$ and $\mathcal{C}_3$.

The polynomial is the useful tool for the proofs of our results. In particular, we state some relations between roots of the multivariate matching polynomial and the inhomogeneous density Turán problem for 3-uniform linear hypergraphs with the hypercycle $\mathcal{C}_3$ which are presented in Theorem 4.

**Example 1.** *Let us consider the* 3*-uniform linear hypergraph* $\mathcal{H}$ *with* 7 *hyperedges as in Fig. 3. Assume that variables* $x_e$ *are given for hyperedges* $e \in \mathcal{E}(\mathcal{H})$ *as follows*

$$x_1 = x_4 = x_7 = 2,\ x_2 = x_6 = 1 \text{ and } x_3 = x_5 = 3.$$

*Then the multivariate matching polynomial of the hypertree* $\mathcal{H}$ *is presented below*

$$\begin{aligned} F_{\mathcal{H}}(\underline{x_e}, t) &= 1 - t(x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7) \\ &+ t^2(x_1x_5 + x_1x_6 + x_1x_7 + x_2x_4 + x_2x_5 \\ &+ x_2x_6 + x_2x_7 + x_3x_5 + x_3x_6 + x_3x_7 + \\ &x_4x_6 + x_4x_7 + x_6x_7) - t^3(x_1x_6x_7 + x_2x_4x_6 \\ &+ x_2x_4x_7 + x_2x_6x_7 + x_3x_6x_7 + x_4x_6x_7) \\ &+ t^4x_2x_4x_6x_7 = 1 - 14t + 44t^2 - 22t^3 + 4t^4. \end{aligned}$$

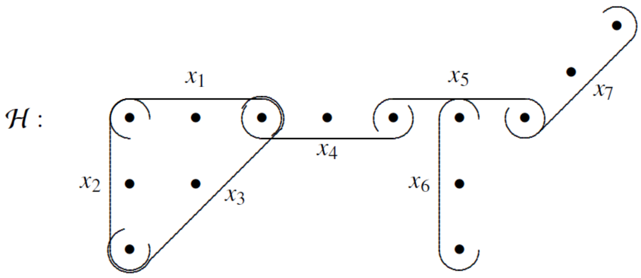


Fig. 3. A 3-uniform linear hypergraph $\mathcal{H}$ with hypercycle $\mathcal{C}_3$, where $|V(\mathcal{H})| = 14$ and $|\mathcal{E}(\mathcal{H})| = 7$, with variables $x_e$ assigned to hyperedges $e \in \mathcal{E}(\mathcal{H})$.

## II. The inhomogeneous density Turán problem for 3-uniform linear unihypercyclic hypergraphs with hypercycle $\mathcal{C}_3$

In this section we study the inhomogeneous density Turán problem for 3-uniform linear hypergraphs $\mathcal{H}$ with one hypercycle $\mathcal{C}_3$, where a hyperedge density $\gamma_e$ is given for each hyperedge $e \in \mathcal{E}(\mathcal{H})$. We extend some results presented in [8], where authors studied the inhomogeneous problem for trees and proved the following theorem.

**Theorem 1.** *(Csikvári, Nagy [8]) Let $T$ be a tree of order $n$ and let $v$ be a leaf of $T$. Assume that for each edge of $T$ a density $\gamma_e = 1 - r_e$ is given. Let $T^{'}$ be a tree obtained from $T$ by deleting the leaf $v$ and the edge $uv$, where $u$ is the unique neighbour of $v$. Let the edge densities $\gamma_e^{'}$ in $T^{'}$ be defined as follows*

$$\gamma_e^{'} = \begin{cases} \gamma_e = 1 - r_e, & \text{if } e \text{ is not incident to } u \text{ in } T', \\ 1 - \frac{r_e}{1 - r_{uv}}, & \text{if } e \text{ is incident to } u \text{ in } T'. \end{cases}$$

*Then the set of densities $\{\gamma_e\}_{e \in E(T)}$ ensures the existence of the factor $T$ if and only if all $\gamma_e^{'} \in (0,1]$ and the set of densities $\{\gamma_e^{'}\}_{e \in E(T')}$ ensures the existence of the factor $T^{'}$.*

Theorem 1 provides authors of [8] with an efficient algorithm (Algorithm $T$) to decide whether a given set of edge densities in a tree ensures the existence of a transversal or does not ensure. Their algorithm is cited on the next page.

We show that their algorithm can be extended for 3-uniform linear hypergraphs with hypercycle $\mathcal{C}_3$. This extension is presented in Algorithm $\mathcal{HC}_3$, which is presented in the second half of this paragraph.

**Proposition 1.** *The Algorithm $\mathcal{HC}_3$ stops in at most $\mathcal{O}(n^2)$ steps, where $n$ is the number of hyperedges of the input hypergraph.*

*Proof.* Execution time for checking of the property described in *Step 0* is at most $\mathcal{O}(n)$, where $n$ is the number of hyperedges of the input hypergraph. Similarly, execution time for checking the first property described in *Step 2* is at most $\mathcal{O}(n)$. In the worst case *Step 1* is executed at most $\mathcal{O}(n)$, similarly, *Step 2*, so the time complexity of our algorithm is at most $\mathcal{O}(n^2)$. □

The correctness of the Algorithm $\mathcal{HC}_3$ follows from the following theorem.

**Theorem 2.** *Let $\mathcal{H}$ be a 3-uniform linear hypergraph with the hypercycle $\mathcal{C}_3$. If $|\mathcal{E}(\mathcal{H})| > 3$ let $u, v \in V(\mathcal{H})$ be two leaves from a pendant hyperedge $e = \{u, v, w\} \in \mathcal{E}(\mathcal{H})$ for some $w \in V(\mathcal{H})$. Assume that for each hyperedge of $\mathcal{H}$ the density $\gamma_e = 1 - r_e$ is given. Let $\mathcal{H}^{'}$ be a hypergraph obtained from $\mathcal{H}$ by deleting vertices $u$ and $v$ with the hyperedge $uvw$. Let the hyperedge densities $\gamma_e^{'}$ in $\mathcal{H}^{'}$ be defined as follows*

$$\gamma_e^{'} = \begin{cases} \gamma_e = 1 - r_e, & \text{if } e \text{ is not incident to } w \text{ in } \mathcal{H}^{'}, \\ 1 - \frac{r_e}{1 - r_{uvw}}, & \text{if } e \text{ is incident to } w \text{ in } \mathcal{H}^{'}. \end{cases}$$

*If $|\mathcal{E}(\mathcal{H})| = 3$ (with hyperedge set $\mathcal{E} = \{ayb, axc, bzc\}$), then let $\mathcal{H}^{'}$ be a hypertree obtained from $\mathcal{H}$ by deleting a vertex of degree 2, say vertex $a$, with incident hyperedges $ayb$ and $axc$. $\mathcal{H}^{'}$ is a hyperpath $\mathcal{P}_{bzc}$. Let the density $\gamma_{czb}^{'}$ in $\mathcal{H}^{'}$ be defined as follows*

$$\gamma_{bzc}^{'} = 1 - \frac{r_{bzc}}{(1 - r_{ayb})(1 - r_{axc})}.$$

*Then the set of densities $\{\gamma_e\}_{e \in \mathcal{E}(\mathcal{H})}$ ensures the existence of a factor $\mathcal{H}$ if and only if all $\gamma_e^{'} \in (0,1]$ and the set of densities $\{\gamma_e^{'}\}_{e \in \mathcal{E}(\mathcal{H}')}$ ensures the existence of a factor $\mathcal{H}^{'}$.*

*Proof.* Let $\mathcal{H}$ be a 3-uniform linear hypergraph with one hypercycle $\mathcal{C}_3$ and let a density $\gamma_e = 1 - r_e$ be given for each $e \in \mathcal{E}(\mathcal{H})$.

($\Rightarrow$) First we prove the following statement: if all $\gamma_e^{'}$ are indeed densities and they ensure the existence of a factor $\mathcal{H}^{'}$, then the original densities $\gamma_e$ ensure the existence of a factor $\mathcal{H}$.

Let $\mathcal{B}[\mathcal{H}]$ be a blow-up hypergraph of the hypergraph $\mathcal{H}$ such that the density between clusters $A_i$, $A_j$ and $A_k$ is at least $\gamma_{ijk}$, where $A_i, A_j, A_k$ are clusters of the vertices and $i, j, k \in V(\mathcal{H})$. We show that $\mathcal{B}[\mathcal{H}]$ contains a factor $\mathcal{H}$.

Assume that $|\mathcal{E}(\mathcal{H})| > 3$. Let $u, v, w \in V(\mathcal{H})$ and $\{u, v, w\} \in \mathcal{E}(\mathcal{H})$, where $u, v$ are leaves. Define $R_{u,v,w}$ as the subset of $A_w$ in the following way (see Fig. 4)

$$R_{u,v,w} = \{x \in A_w \mid \exists_{u' \in A_u, v' \in A_v} \{u^{'}, v^{'}, x\} \in \mathcal{E}(\mathcal{B}[\mathcal{H}])\}.$$

Note that by (1)

$$|R_{u,v,w}| \cdot |A_u| \cdot |A_v| \geq e(R_{u,v,w}, A_u, A_v) = e(A_u, A_v, A_w)$$
$$= \gamma_{uvw} |A_u| \cdot |A_v| \cdot |A_w|.$$

Hence

$$|R_{u,v,w}| \geq \gamma_{uvw} |A_w|.$$

Now we show the lower bound for the number of hyperedges incident to $R_{u,v,w}$. Let $k, z \in V(\mathcal{H})$ such that $\{k, z, w\} \in \mathcal{E}(\mathcal{H})$. By the inclusion - exclusion formula we count the lower bound for the number of hyperedges between $R_{u,v,w}$, $A_k$ and $A_z$ as follows

$$e(R_{u,v,w}, A_k, A_z) \geq$$
$$e(A_w, A_k, A_z) - (|A_w| - |R_{u,v,w}|) \cdot |A_k| \cdot |A_z| =$$

```
Algorithm $T$
  Step 0.
  Let there be given a tree $T^0$ and edge densities $\gamma_e^0$. Set $T := T^0$ and $r_e = 1 - \gamma_e^0$.

  Step 1.
  Consider $(T, r_e)$.
    • if $|V(T)| = 2$ and $0 \leq r_e < 1$ then
        STOP: the densities $\gamma_e^0$ ensure the existence of a factor $T^0$.
    • if $|V(T)| \geq 2$ and there exists an edge for which $r_e \geq 1$ then
        STOP: the densities $\gamma_e^0$ do not ensure the existence of a factor $T^0$.

  Step 2.
  if $|V(T)| \geq 3$ and $0 \leq r_e < 1$ for all edges $e \in E(T)$ then
  DO pick a vertex $v$ of degree 1 and let $u$ be its unique neighbour. Let $T^{'} := T - v$ and
      $r_e^{'} = \begin{cases} r_e, & \text{if } e \text{ is not incident to } u, \\ \frac{r_e}{1-r_{uv}}, & \text{if } e \text{ is incident to } u. \end{cases}$
  Jump to Step 1 with $(T, r_e) := (T^{'}, r_e^{'})$.
```

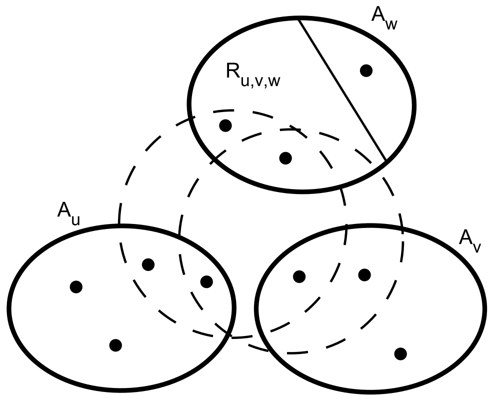


Fig. 4. Clusters $A_u, A_v$ and $A_w$ (bold line) with some hyperedge (broken line) and the set $R_{u,v,w}$

$$|R_{u,v,w}| \cdot |A_k| \cdot |A_z| + (\gamma_{wkz} - 1) \cdot |A_w| \cdot |A_k| \cdot |A_z| \geq$$

$$|R_{u,v,w}| \cdot |A_k| \cdot |A_z| + \frac{1}{\gamma_{uvw}}(\gamma_{wkz} - 1) \cdot |R_{u,v,w}| \cdot |A_k| \cdot |A_z| =$$

$$\left(1 - \frac{r_{wkz}}{1 - r_{uvw}}\right) \cdot |R_{u,v,w}| \cdot |A_k| \cdot |A_z| =$$

$$\gamma_{wkz}^{'} \cdot |R_{u,v,w}| \cdot |A_k| \cdot |A_z|.$$

Now, by deleting the vertex sets $A_u$, $A_v$ and $A_w \backslash R_{u,v,w}$ from $\mathcal{B}[\mathcal{H}]$, we obtain a hypergraph which is a blow-up hypergraph of $\mathcal{H}^{'}$ with the hyperedge densities ensuring the existence of the factor $\mathcal{H}^{'}$.

Moreover, by the definition of $R_{u,v,w}$ the factor $\mathcal{H}^{'}$ can be extended to a factor $\mathcal{H}$.

Now let us assume that $|\mathcal{E}(\mathcal{H})| = 3$, i.e. a hypergraph $\mathcal{H}$ is isomorphic to the hypercycle $\mathcal{C}_3$. Let $\mathcal{E}(\mathcal{H}) = \{ayb, axc, bzc\}$, where vertices $a, b, c$ have degree equal to 2 and vertices $x, y, z$ have degree equal to 1. Let $A_a$ be a cluster of the vertex $a$. Define sets $R_{a,y,b}$ and $R_{a,x,c}$ in the following way (see Fig. 5 )

$$R_{a,y,b} = \{v \in A_b \ |\exists_{a' \in A_a, y' \in A_y} \{a^{'}, y^{'}, v\} \in \mathcal{E}(\mathcal{B}[\mathcal{H}])\},$$

$$R_{a,x,c} = \{v \in A_c \ |\exists_{a' \in A_a, x' \in A_x} \{a^{'}, x^{'}, v\} \in \mathcal{E}(\mathcal{B}[\mathcal{H}])\}.$$

Note that by (1)

$$|R_{a,y,b}| \cdot |A_a| \cdot |A_y| \geq e(R_{a,y,b}, A_a, A_y) =$$

$$e(A_b, A_a, A_y) = \gamma_{ayb}|A_a| \cdot |A_y| \cdot |A_b|$$

and

$$|R_{a,x,c}| \cdot |A_a| \cdot |A_x| \geq e(R_{a,x,c}, A_a, A_x) =$$

$$e(A_c, A_a, A_x) = \gamma_{axc}|A_a| \cdot |A_x| \cdot |A_c|.$$

Hence we have the following lower bounds for the cardinalities of $R_{a,y,b}$ and $R_{a,x,c}$

$$|R_{a,y,b}| \geq \gamma_{ayb}|A_b|$$

```
Algorithm $\mathcal{HC}_3$ (for 3-uniform linear hypergraph with a hypercycle $\mathcal{C}_3$)
  Input: a 3-uniform linear hypergraph $\mathcal{H}$ with one hypercycle $\mathcal{C}_3$ with the set of hyperedge densities $\{\gamma_e\}_{e\in\mathcal{E}(\mathcal{H})}$.
  Output: a boolean value
      $D = \begin{cases} TRUE, & \text{the densities } \gamma_e \text{ ensure the existence of a factor } \mathcal{H}, \\ FALSE, & \text{the densities } \gamma_e \text{ does not ensure the existence of a factor } \mathcal{H}. \end{cases}$
  Consider a weighted hypergraph $(\mathcal{H}, r_e)$, where $r_e = 1 - \gamma_e$.

  Step 0.
  if $|E(\mathcal{H})| \ge 1$ and there exists a hyperedge $e \in \mathcal{E}(\mathcal{H})$ for which $r_e \ge 1$ then
      $D := FALSE$; STOP;

  Step 1.
  if $|E(\mathcal{H})| = 1$ (means $\mathcal{H}$ is a hyperpath $\mathcal{P}_1$) and $0 \le r_e < 1$ then
      $D := TRUE$; STOP;

  Step 2.
  if $|E(\mathcal{H})| > 3$ then
      pick two leaves $u, v$ from a pendant hyperedge $f = \{u, v, w\} \in \mathcal{E}(\mathcal{H})$. Let
      $\mathcal{H}' = (V(\mathcal{H}) - \{u, v\}, \mathcal{E}(\mathcal{H}) - \{\{u, v, w\}\})$ and for each hyperedge $e \in \mathcal{E}(\mathcal{H}')$ set
          $r_e^{'} = \begin{cases} r_e, & \text{if } e \cap f = \emptyset, \\ \frac{r_e}{1-r_{uvw}}, & \text{if } e \cap f = \{w\}; \end{cases}$

  if $|E(\mathcal{H})| = 3$ ($\mathcal{E}(\mathcal{H}) = \{ayb, axc, bzc\}$) then
      pick vertex of degree equal to 2, say vertex $a$, and let $\mathcal{H}' = (V(\mathcal{H}) - \{a, y, x\}, \mathcal{E}(\mathcal{H}) - \{ayb, axc\})$. For hyperedge
      $e = bzc \in \mathcal{E}(\mathcal{H}')$ set
          $r_e^{'} = r_{bzc}^{'} = \frac{r_{bzc}}{(1 - r_{ayb})(1 - r_{axc})};$

  if $r_e^{'} \ge 1$ for some hyperedge $e \in \mathcal{E}(\mathcal{H}')$ then
      $D := FALSE$; STOP;

  Go to Step 1 with $(\mathcal{H}, r_e) := (\mathcal{H}^{'}, r_e^{'})$.
```

and
$$|R_{a,x,c}| \ge \gamma_{axc}|A_c|.$$

Next let us show how many hyperedges are incident to the sets $R_{a,y,b}$ and $R_{a,x,c}$. By the inclusion - exclusion formula we count the lower bound for the number of hyperedges between $R_{a,y,b}$ and $R_{a,x,c}$

$$e(R_{a,y,b}, R_{a,x,c}, A_z) \ge e(A_b, A_c, A_z) -$$
$$(|A_b| - |R_{a,y,b}|) \cdot |A_c| \cdot |A_z| - (|A_c| - |R_{a,x,c}|) \cdot |A_b| \cdot |A_z| +$$
$$(|A_b| - |R_{a,y,b}|)(|A_c| - |R_{a,x,c}|) \cdot |A_z| = |R_{a,y,b}| \cdot |R_{a,x,c}| \cdot |A_z| +$$
$$(\gamma_{bcz} - 1) \cdot |A_b| \cdot |A_c| \cdot |A_z| \ge |R_{a,y,b}| \cdot |R_{a,x,c}| \cdot |A_z| +$$
$$(\gamma_{bcz} - 1) \frac{1}{\gamma_{ayb}} \frac{1}{\gamma_{axc}} \cdot |R_{a,y,b}| \cdot |R_{a,x,c}| \cdot |A_z| =$$
$$\left(1 - \frac{r_{bzc}}{(1 - r_{ayb})(1 - r_{axc})}\right) \cdot |R_{a,y,b}| \cdot |R_{a,x,c}| \cdot |A_z| =$$
$$\gamma_{bzc}^{'} |R_{a,y,b}| \cdot |R_{a,x,c}| \cdot |A_z|.$$

Now, by deleting the vertex sets $A_a$, $A_b \backslash R_{a,y,b}$ and $A_c \backslash R_{a,x,c}$ from $\mathcal{B}[\mathcal{H}]$, we obtain a hypergraph which is a blow-up hypergraph of $\mathcal{C}^{'} = \mathcal{P}_1$, where $V(\mathcal{P}_1) = \{b, z, c\}$, with the hyperedge density ensuring the existence of the factor $\mathcal{P}_2$.

Moreover, by the definition of $R_{a,y,b}$ and $R_{a,x,c}$ the factor $\mathcal{P}_1$ can be extended to a factor $\mathcal{C}_3$.

($\Leftarrow$) Note that if $\gamma_{wkz}^{'} < 0$ then $\gamma_{wkz} + \gamma_{uvw} < 1$. So there exists a construction of blow-up hypergraph which does not induce the linear hyperpath $\mathcal{P}_2$ with the consecutive vertices $u, v, w, k, z$ and hyperedges $\{u, v, w\}$, $\{w, k, z\}$, where $i \in A_i$

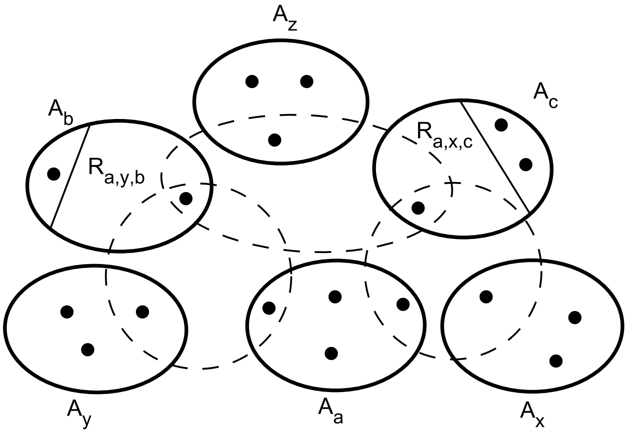


Fig. 5. Clusters $A_a, A_x$, $A_y$, $A_b$ and $A_c$ (bold line) with some hyperedges (broken line) and the sets $R_{a,y,b}$ and $R_{a,x,c}$.

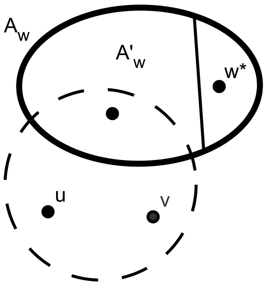


Fig. 6. We assume that $\mathcal{B}^{'}[\mathcal{H}^{'}]$ is without a factor $\mathcal{H}'$. The construction of the blow-up hypergraph $\mathcal{B}[\mathcal{H}]$ without factor $\mathcal{H}$ for the case where vertices $u, v$ are leaves in $\mathcal{H}$ and $\mathcal{H}' = (V(\mathcal{H}) - \{u, v\}, \mathcal{E}(\mathcal{H}) - \{uvw\})$. The cluster $A_w'$ is in $\mathcal{B}^{'}[\mathcal{H}^{'}]$. Let $A_w = \{w^*\} \cup A_w^{'}$, $A_u = \{u\}$ and $A_v = \{v\}$ be clusters in $\mathcal{B}[\mathcal{H}]$. Bold line - cluster, broken line - hyperedge.

for $i \in \{u, v, w, k, z\}$ in this case. Therefore, if some $\gamma_{wkz}^{'} < 0$ then there exists a construction for a blow-up hypergraph of the hypertree $\mathcal{H}$ without a factor $\mathcal{H}$.

Next assume that all the $\gamma_e^{'}$ are proper densities, but there is a construction of a blow-up hypergraph, say $\mathcal{B}^{'}[\mathcal{H}^{'}]$, with hyperedge densities at least $\gamma_e^{'}$, but which does not induce a factor $\mathcal{H}^{'}$. Thus we construct a blow-up hypergraph $\mathcal{B}[\mathcal{H}]$ of the hypertree $\mathcal{H}$ not inducing $\mathcal{H}$. We consider two cases. First, let $|\mathcal{E}(\mathcal{H})| > 3$ and $v, u$ be two leaves of $\mathcal{H}$ such that $uvw \in \mathcal{E}(\mathcal{H})$ for some vertex $w \in V(\mathcal{H})$. Let $\mathcal{H}' = (V(\mathcal{H}) - \{u, v\}, \mathcal{E}(\mathcal{H}) - \{uvw\})$. Set $A_w = \{w^*\} \cup A_w^{'}$, $A_u = \{u\}$ and $A_v = \{v\}$. We create hyperedges $uvw$ for all $w \in A_w^{'}$ but do not create $uvw^*$ without changing densities in $\mathcal{B}^{'}[\mathcal{H}^{'}]$ and with an appropriate density $\gamma_{uvw}$ (see Fig. 6 ).

Now assume that $\mathcal{H} = \mathcal{C}_3$ with hyperedge set $\mathcal{E}(\mathcal{H}) = \{ayb, axc, bzc\}$. Let $\mathcal{H}' = (V(\mathcal{H}) - \{a, y, c\}, \mathcal{E}(\mathcal{H}) - \{ayb, axc\})$, where $a$ is a vertex of $\mathcal{C}_3$ of degree 2. Set $A_b = \{b^*\} \cup A_b'$, $A_c = \{c^*\} \cup A_c'$, $A_a = \{a\}$, $A_x = \{x\}$ and $A_y = \{y\}$. We create hyperedges $ayb$ for all $b \in A_b^{'}$ and $axc$ for all $c \in A_c^{'}$ but do not create hyperedges $ayb^*$ and $axc^*$ without changing densities in $\mathcal{B}^{'}[\mathcal{H}^{'}]$ and with an appropriate densities $\gamma_{ayb}$ and $\gamma_{axc}$ (see Fig. 7). □

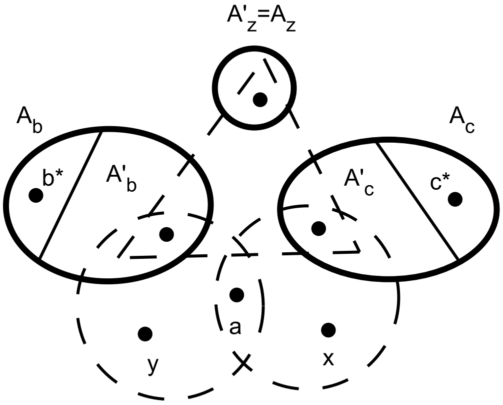


Fig. 7. We assume that $\mathcal{B}^{'}[\mathcal{H}^{'}]$ is without a factor $\mathcal{H}'$. The construction of the blow-up hypergraph $\mathcal{B}[\mathcal{H}]$ without factor $\mathcal{H} = \mathcal{C}_3$ for the case where vertex $a$ has degree 2 in $\mathcal{H}$ and $\mathcal{H}' = (V(\mathcal{H}) - \{a, x, y\}, \mathcal{E}(\mathcal{H}) - \{ayb, axc\})$. The clusters $A_b'$ and $A_c'$ are in $\mathcal{B}^{'}[\mathcal{H}^{'}]$. Let $A_b = \{b^*\} \cup A_b^{'}$, $A_c = \{c^*\} \cup A_c^{'}$, $A_a = \{a\}$, $A_x = \{x\}$ and $A_y = \{y\}$ be clusters in $\mathcal{B}[\mathcal{H}]$. Bold lines - clusters, broken lines - hyperedges.

**Example 2.** *Let us consider a* 3*-uniform linear hypergraph* $\mathcal{H}$ *with one hypercycle* $\mathcal{C}_3$*, such that* $|\mathcal{E}(\mathcal{H})| = 5$ *and* $|V(\mathcal{H})| = 10$ *, presented in Fig. 8 with two different sets of parameters* $\{r_e\}_{e\in\mathcal{E}(\mathcal{H})}$ *(in round brackets are given parameters* $r_e$ *from the second set of hyperedge densities). In Table I are presented two different sets of densities* $\{\gamma_e\}_{e\in\mathcal{E}(\mathcal{H})}$, $\gamma_e = 1 - r_e$, *and changes of parameters* $r_e$ *during the execution of the Algorithm* $\mathcal{HC}_3$.

*We are interested in whether these sets of hyperedge densities ensure an existence of the hypergraph* $\mathcal{H}$ *as a factor. To solve this problem we use Algorithm* $\mathcal{HC}_3$*. For each hyperedge* $e$ *a parameter* $r_e = 1 - \gamma_e$ *is assigned as in Fig. 8. Let run Algorithm* $\mathcal{HC}_3$*. All parameters satisfy the condition* $0 \leq r_e < 1$*, so we cut the hyperedge* $e_{fhi}$ *and modify parameters* $r_e$ *by proper formulas presented in the algorithm. We repeat this procedure untill we get a hypergraph with at least one hyperedge* $e^*$ *for which parameter* $r_{e^*} \geq 1$ *or one-hyperedge hyperpath (see Fig. 9-12). Notice that we get two different velues at the end. First set of densities* $\{\gamma_e\}$ *ensure the existence of* $\mathcal{H}$ *as a factor and the second set* $\{\gamma_e\}$ *does not ensure.*

Now we show some relations between roots of the multivariate matching polynomial and the inhomogeneous density Turán problem. This kind of relations for 3-uniform linear hypertrees have been studied in [4]. Authors of [4] obtained Theorem 3 cited for completeness of this paper. Ealier this subject has been studied by Csikvári and Nagy [8] for trees and in [3] for some family of connected unicyclic graphs.

**Theorem 3.** *(Bielak, Powroźnik [4]) Let* $\mathcal{T} = (V, \mathcal{E})$ *be a weighted* 3*-uniform linear hypertree. Let* $\gamma_e = 1 - tr_e$ *be densities assigned to each hyperedge* $e \in \mathcal{E}(\mathcal{T})$*, where* $r_e \in$

TABLE I
CHANGING OF PARAMETERS $r_e$ FOR HYPEREDGES $e \in \mathcal{E}(\mathcal{H})$ DURING THE EXECUTION OF THE ALGORITHM $\mathcal{HC}_3$ FOR TWO DIFFERENT SETS OF HYPEREDGE DENSITIES $\gamma_e$ OF THE HYPERGRAPH PRESENTED IN FIG. 8.

| $e$ | TRUE | | | | | | FALSE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | step: | (1) | (2) | (3) | (4) | (5) | step: | (1) | (2) | (3) | (4) | (5) |
| | $\gamma_e$ | $r_e$ | $r_e$ | $r_e$ | $r_e$ | $r_e$ | $\gamma_e$ | $r_e$ | $r_e$ | $r_e$ | $r_e$ | $r_e$ |
| $abc$ | 3/4 | 1/4 | 1/4 | 1/4 | 10/31 | 40/69 | 1/2 | 1/2 | 1/2 | 1/2 | 7/8 | 175/104 |
| $axy$ | 3/4 | 1/4 | 1/4 | 1/4 | 1/4 | – | 4/5 | 1/5 | 1/5 | 1/5 | 1/5 | – |
| $yzc$ | 4/5 | 1/5 | 1/5 | 1/5 | 8/31 | – | 4/5 | 1/5 | 1/5 | 1/5 | 7/20 | – |
| $cde$ | 7/8 | 1/8 | 1/8 | 9/40 | – | – | 3/4 | 1/4 | 1/4 | 3/7 | – | – |
| $efg$ | 2/3 | 1/3 | 4/9 | – | – | – | 2/3 | 1/3 | 5/12 | – | – | – |
| $fhi$ | 3/4 | 1/4 | – | – | – | – | 4/5 | 1/5 | – | – | – | – |

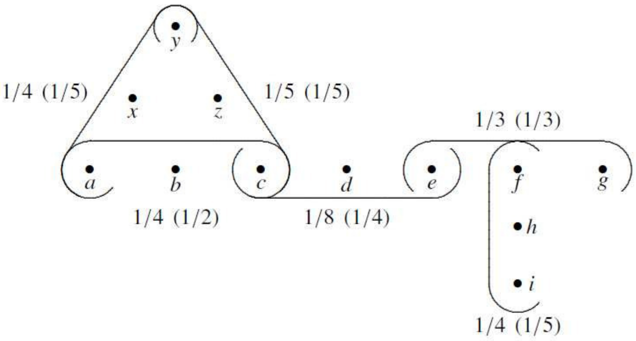


Fig. 8. The input hypergraph $\mathcal{H}$ for Algorithm $\mathcal{HC}_3$.

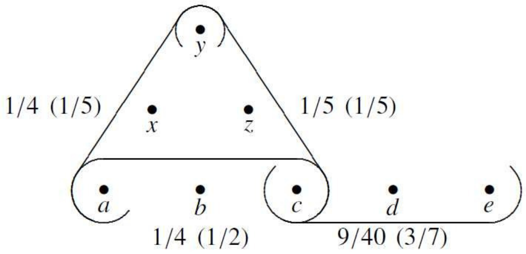


Fig. 10. The hypergraph $\mathcal{H}'$ obtained after the second execution of *Step 2* of Algorithm $\mathcal{HC}_3$ with the hypergraph $\mathcal{H}$ presented in Fig. 9, where the hyperedge $efg$ was deleted and the parameter $r_{cde}$ was modyfied according to the first conditional instruction.

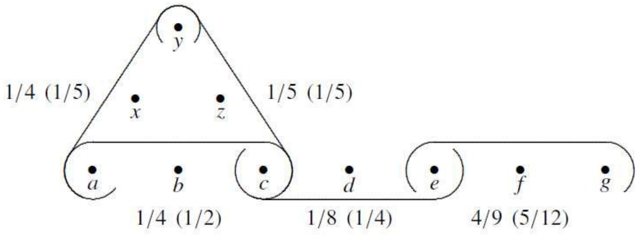


Fig. 9. The hypergraph $\mathcal{H}'$ obtained after the first execution of *Step 2* of Algorithm $\mathcal{HC}_3$ with the hypergraph $\mathcal{H}$ presented in Fig. 8, where the hyperedge $fhi$ was deleted and the parameter $r_{efg}$ was modyfied according to the first conditional instruction.

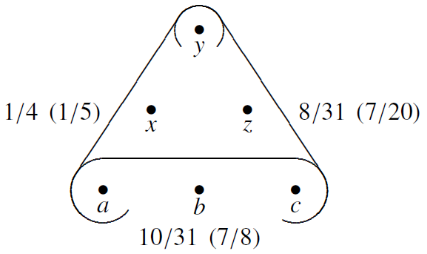


Fig. 11. The hypergraph $\mathcal{H}'$ obtained after the third execution of *Step 2* of Algorithm $\mathcal{HC}_3$ with the hypergraph $\mathcal{H}$ presented in Fig. 10, where the hyperedge $cde$ was deleted and the parameters $r_{abc}$ and $r_{yzc}$ were modyfied according to the first conditional instruction.

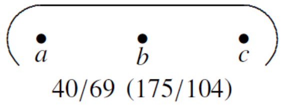


Fig. 12. The hypergraph $\mathcal{H}'$ obtained after the last execution of *Step 2* of Algorithm $\mathcal{HC}_3$ with the hypergraph $\mathcal{H}$ presented in Fig. 11, where the hyperedges $axy$ and $yzc$ were deleted and the parameter $r_{abc}$ was modyfied according to the second conditional instruction.

$[0, 1)$. *Assume that after running of Algorithm $\mathcal{T}$ we get a one-hyperedge hyperpath $\mathcal{P}_1$ with*

$$F_{\mathcal{P}_1}(\underline{r_e}, t) = 0.$$

*Then $t$ is a root of the multivariate matching polynomial $F_{\mathcal{T}}(\underline{r_e}, s)$ of the hypertree $\mathcal{T}$.*

In Theorem 4 we show similar relation between roots of the multivariate matching polynomial and the inhomogeneous density Turán problem for 3-uniform linear hypergraphs with a hypercycle $\mathcal{C}_3$.

**Theorem 4.** *Let $\mathcal{H} = (V, \mathcal{E})$ be a weighted 3-uniform linear unihypercyclic hypergraph with a hypercycle $\mathcal{C}_3$. Let $\gamma_e = 1 - tr_e$ be densities assigned to each hyperedge $e \in \mathcal{E}(\mathcal{H})$, where $r_e \in [0, 1)$. Assume that after running of Algorithm $\mathcal{HC}_3$ we get a hypercycle $\mathcal{C}_3$ with*

$$F_{\mathcal{C}_3}(\underline{r_e}, t) = 0.$$

*Then $t$ is a root of the multivariate matching polynomial $F_{\mathcal{H}}(\underline{r_e}, s)$ of the hypergraph $\mathcal{H}$.*

*Proof.* Let $\mathcal{H} = (V, \mathcal{E})$ be a weighted 3-uniform linear unihypercyclic hypergraph with hypercycle $\mathcal{C}_3$. Assume that $|\mathcal{E}(\mathcal{H})| = n$. To prove this theorem we use induction on the number of hyperedges of the hypergraph $\mathcal{H}$.

If this hypergraph consists of 3 hyperedges (i.e., $\mathcal{H}$ is isomorphic to $\mathcal{C}_3$, say with $\mathcal{E}(\mathcal{H}) = \{abc, cde, efa\}$), then

$$F_{\mathcal{H}}(\underline{r_e}, t) = 1 - t(r_{abc} + r_{cde} + r_{efa})$$

and the condition $F_{\mathcal{H}}(\underline{r_e}, t) = 0$ means that $t$ is a root of this multivariate matching polynomial of the hypergraph $\mathcal{H}$.

Assume that the statement of the theorem is true for each hypergraph on at most $n - 1$ hyperedges, where $n > 3$. Let $\mathcal{H}$ be a hypergraph with $n$ hyperedges and assume that we execute the Algorithm $\mathcal{HC}_3$ for a hyperedge $e = \{u, v, w\}$, shortly $uvw$, in the *Step 2*, where vertices $u, v$ are two leaves in $\mathcal{H}$. Let $\mathcal{H}^{'} = \mathcal{H} - \{u, v\}$ be a hypergraph obtained from hypergraph $\mathcal{H}$ by deleting $u$ and $v$ and the hyperedge $uvw$. Densities on hyperedges in hypergraph $\mathcal{H}^{'}$ are modyfied by formulas presented in Algorithm $\mathcal{HC}_3$. By executing the Algorithm $\mathcal{HC}_3$ with input $\mathcal{H}'$ we obtain a hypercycle $\mathcal{C}_3$ with $F_{\mathcal{C}_3}(\underline{r'_e}, t) = 0$. By induction we get that $F_{\mathcal{T}'}(\underline{r_e}^{'}, t) = 0$.

Now we apply the formula (2) for hypergraphs $\mathcal{H}'$ and $\mathcal{H}$.

We can expand $F_{\mathcal{H}'}$ according to whether a monomial contains $x_{wkz}$ (where $wkz \in \mathcal{E}(\mathcal{H}^{'})$) or not. Obviously, each monomial contains at most one of the variables $x_{wkz}$ where $wkz \in \mathcal{E}(\mathcal{H})$.

Thus

$$F_{\mathcal{H}'}(\underline{x_e}, s) = Q_0(\underline{x_e}, s) - \sum_{\{k,z,w\} \in \mathcal{E}(\mathcal{H}')} s x_{wkz} Q_{kz}(\underline{x_e}, s),$$

where $Q_0(\underline{x_e}, s)$ consists of those terms which do not contain $x_{wkz}$ and $-s x_{wkz} Q_{kz}(\underline{x_e}, s)$ consists of those terms which contain $x_{wkz}$ (i.e., these terms correspond to the matchings containing the hyperedge $wkz$).

We observe that

$$F_{\mathcal{H}}(\underline{x_e}, s) = (1 - s x_{uvw}) Q_0(\underline{x_e}, s) - \sum_{\{k,z,w\} \in \mathcal{E}(\mathcal{H}')} s x_{wkz} Q_{kz}(\underline{x_e}, s).$$

Since

$$0 = F_{\mathcal{H}'}(\underline{r'_e}, t) = Q_0(\underline{r_e}, t) - \sum_{\{k,z,w\} \in \mathcal{E}(\mathcal{H}')} t \frac{r_{wkz}}{1 - t r_{uvw}} Q_{kz}(\underline{r_e}, t)$$

we have

$$0 = (1 - t r_{uvw}) F_{\mathcal{H}'}(\underline{r'_e}, t) = (1 - t r_{uvw}) Q_0(\underline{r_e}, t) - \sum_{\{k,z,w\} \in \mathcal{E}(\mathcal{H}')} t r_{wkz} Q_{kz}(\underline{r_e}, t) = F_{\mathcal{H}}(\underline{r_e}, t).$$

So $t$ is a root of $F_{\mathcal{H}}(\underline{r_e}, s)$. The proof is done. □

To search more relations between roots of the multivariate matching polynomial and the inhomogeneous density Turán problem for 3-uniform linear hypergraphs with a hypercycle $\mathcal{C}_3$ we need analogue result to Lemma 1 presented below. According to our knowledge similar result for hypercycle $\mathcal{C}_3$ is not known.

**Lemma 1.** *(Bondy, et al. [6]) Let $\alpha$, $\beta$, $\gamma$ be the edge densities between the clusters of a blow-up graph of the triangle - a cycle $C_3$. If*

$$\alpha\beta + \gamma > 1, \beta\gamma + \alpha > 1, \gamma\alpha + \beta > 1,$$

*then the blow-up graph contains a triangle as a transversal.*

## III. Conclusion

In this paper we showed some results for the inhomogeneous density Turán problem of 3-uniform connected linear unihypercyclic hypergraphs with a hypercycle $\mathcal{C}_3$.

We presented Algorithm $\mathcal{HC}_3$ for testing whether the 3-uniform connected linear hypergraph $\mathcal{H}$ with a hypercycle $\mathcal{C}_3$ with a given set of hyperedge densities $\{\gamma_e\}_{e \in \mathcal{E}(\mathcal{H})}$ is a transversal of a blow-up hypergraph $\mathcal{B}[\mathcal{H}]$. The Algorithm $\mathcal{HC}_3$ has $\mathcal{O}(n^2)$ time complexity in the worst case, where $n$ is the number of hyperedges of $\mathcal{H}$. In this way we have the answer whether the hyperedge densities ensure the existence of the transversal or does not ensure.

Moreover, we stated Theorem 2 to prove the correctness of the Algorithm $\mathcal{HC}_3$.

Additionally, in Theorem 4, we stated some relation between roots of the multivariate matching polynomial and the inhomogeneous density Turán problem for 3-uniform linear hypergraphs with a hypercycle $\mathcal{C}_3$.

**Open problem:** In the future work we want to find relations between location of roots of the multivariate matching polynomial and the inhomogeneous density Turán problem for $r$-uniform linear hypergraphs with one hypercycle of length $t$, where $t \geq 3$. This problem for trees was studied in [8] and for some connected unicyclic graphs was studied in [3].

## References

[1] R. Baber, J.R. Johnson and J. Talbot, The minimal density of triangles in tripartite graphs, *LMS J. Comput. Math.*, 13 (2010), 388–413, http://dx.doi.org/10.1112/S1461157009000436.

[2] C. Berge, Graphs and hypergraphs, *Elsevier, New York*, NY, USA (1973).

[3] H. Bielak, K. Powroźnik, An efficient algorithm for the density Turán problem of some unicyclic graphs, *Annals of Computer Science and Information Systems, Proceedings of the 2014 FedCSIS*, Vol. 2 (2014), 479–486, http://dx.doi.org/10.15439/978-83-60810-58-3.

[4] H. Bielak, K. Powroźnik, An efficient algorithm for the density Turán problem of 3-uniform linear hypertrees, unpublished.

[5] B. Bollobás, Extremal Graph Theory, *Academic Press* (1978).

[6] A. Bondy, J. Shen, S. Thomassé and C. Thomassen, Density Conditions for triangles in multipartite graphs, *Combinatorica*, 26 (2006), http://dx.doi.org/10.1007/s00493-006-0009-y.

[7] W.G Brown, P. Erdös and M. Simonovits, Extremal problems for directed graphs, *Journal of Combinatorial Theory, Series B* 15 (1) (1973), 77–93, http://dx.doi.org/10.1016/0095-8956(73)90034-8.

[8] P. Csikvári and Z. L. Nagy, The density Turán Problem, *Combinatorics, Probability and Computing*, 21 (2012), 531–553, http://dx.doi.org/10.1017/S0963548312000016.

[9] Z. Füredi, Turán type problems, *Survey in Combinatorics* Vol. 166 of *London Math. Soc. Lecture Notes* (A.D. Keedwell, ed.) (1991), 253–300, http://dx.doi.org/10.1017/cbo9780511666216.010.

[10] C.D. Godsil and G. Royle, Algebraic Graph Theory, *Springer* (2001), http://dx.doi.org/10.1007/978-1-4613-0163-9.

[11] G. Jin, Complete subgraphs of $r$-partite graphs, *Combin. Probab. Comput.*, 1 (1992), 241–250, http://dx.doi.org/10.1017/s0963548300000274.

[12] Z.L. Nagy, A multipartite version of the Turán problem - density conditions and eigenvalues, *The Electronic Journal of Combinatorics*, 18 (2011), # P46.

[13] P. Turán, On an extremal problem in graph theory, *Mat. Fiz. Lapok*, 48 (1941), 436–452.

[14] R. Yuster, Independent transversal in $r$-partite graphs, *Discrete Math.*, 176 (1997), 255–261, http://dx.doi.org/10.1016/s0012-365x(96)00300-7.

DOI: 10.15439/2015F47
ACSIS, Vol. 5

# Comparative Evaluation of the Stochastic Simplex Bisection Algorithm and the SciPy.Optimize Module

Christer Samuelsson
German Research Center for Artificial Intelligence
Email: christer.samuelsson@dfki.de

***Abstract*—The stochastic simplex bisection (SSB) algorithm is evaluated against the collection of optimizers in the Python SciPy.Optimize module on a prominent test set. The SSB algorithm greatly outperforms all SciPy optimizers, save one, in exactly half the cases. It does slightly worse on quadratic functions, but excels at trigonometric ones, highlighting its multimodal prowess. Unlike the SciPy optimizers, it sustains a high success rate. The SciPy optimizers would benefit from a more informed metaheuristic strategy and the SSB algorithm would profit from quicker local convergence and better multidimensional capabilities. Conversely, the local convergence of the SciPy optimizers is impressive and the multimodal capabilities of the SSB algorithm in separable dimensions are uncanny.**

## I. Introduction

Stochastic optimization [1], i.e., using randomness to guide search, is currently popular: genetic algorithms [2], particle swarm optimization [3], and ant-colony optimization [4] are celebrity approaches. These schemes are not only applied to discrete problems, but also to continuous ones, especially for objective functions where the gradient and Hessian are not readily available, e.g., where the function value is only obtainable by expensive simulation. They are also used to optimize highly multimodal functions.

We here compare the performance of the stochastic simplex bisection (SSB) algorithm [5] with that of the optimizers of the Python SciPy.optimize module [6]. The former employs a common stochastic optimization scheme, but unlike other stochastic approaches, it applies the scheme to search space regions, rather than to individual points. The latter are high-quality, local optimizers that are available with most Python distributions. The SSB algorithm has not previously been tested against other optimizers. We here seek to rectify this.

The rest of the article is organized as follows. Section II presents the stochastic simplex bisection algorithm. Section III describes the optimizers of the Python SciPy.Optimize module. Section IV details the experimental setup, and reports and analyses the findings. It also contains a short digression on the employed success criterion and its success regions.

## II. The Stochastic Simplex Bisection Algorithm

Consider the simple problem where we wish to minimize $f(x)$, which is strictly convex on $\mathbf{R}$. Assume further that we have found $x_1 < x_3 < x_5$, where $f(x_1) \geq f(x_3) \leq f(x_5)$, i.e., that we have found an interval that has an interior point $x_3$ with a smaller function value than its end points. [1]

**Algorithm for Convex Functions**

**Choose** $x_2 \in ]x_1, x_3[$ and $x_4 \in ]x_3, x_5[$, i.e., choose interior points $x_2$ and $x_4$.
**If** $f(x_2) \leq f(x_3)$, recurse on $x_1, x_2, x_3$.
**If** $f(x_4) \leq f(x_3)$, recurse on $x_3, x_4, x_5$.
**Otherwise**, recurse on $x_2, x_3, x_4$.

We thus recurse on the subinterval that has an interior point with a smaller function value than its end points. [2]

The SSB algorithm generalizes this to non-convex functions in $n$ dimensions. It generalizes an interval to a simplex, rather than to a hyperbox, The latter has $2^n$ corners, and bisecting it requires computing the function value in $2^{n-1}$ new corners. The former has only $n+1$ corners, and bisecting it only requires calculating the function value in one new point.

**Core SSB Algorithm**

**Given** a set $\{T_k\}$ of non-overlapping simplexes that partition the original simplex $T_0$, each equipped with a positive score $s_k$.
**Select** the next simplex $T_k$ to bisect at random with probability $\frac{s_k}{\sum_{k'} s_{k'}}$.
**Select** a bisection point at random roughly in the middle of the longest edge of $T_k$.
**Replace** $T_k$ with its two offspring.

This algorithm is complete in the sense that no portion of the search space is ever discarded, and it avoids redundancy by using non-overlapping simplexes. These are two common pitfalls of stochastic optimization. The algorithm still reaps the benefits of stochastic search in exploring more promising regions earlier, in average, while granting also less promising regions a non-zero chance of being explored.

It is often desirable to start from a hyperbox. For example, constraints often take the form of bounds on the individual variables of each dimension. The tested SSB algorithm restarts the core SSB algorithm repeatedly from a hyperbox created in the previous iteration. It uses an outer loop over epochs that maintains the hyperbox, and an inner loop over rounds that implements the core SSB algorithm. Note that although partitioning a hyperbox into a set of simplexes is trivial in

[1] By strict convexity, one of the two end point values must be strictly larger, i.e., either $f(x_1) > f(x_3)$ or $f(x_3) < f(x_5)$, or both.

[2] If $f(x_2) = f(x_3)$, the algorithm could be clever and instead recurse on $x_2, x_{23}, x_3$, with $x_{23} \in ]x_2, x_3[$, where, by necessity $f(x_2) > f(x_{23}) < f(x_3)$, due to strict convexity. Similarly for $f(x_4) = f(x_3)$.

978-83-60810-66-8/$25.00©2015, IEEE

two dimensions, it is a challenging and time-consuming task in higher ones, see [7] and [8].

### A. Outer Loop: Maintaining a Hyperbox

The tested SSB algorithm consists of two nested loops. The outer loop over epochs maintains a hyperbox. Each *epoch* runs the inner loop over rounds, where each *round* consists of one simplex bisection, see Section II-B. The terms *best point*, *very best point*, and *epoch phases* will be defined shortly.

The hyperbox is modified after each epoch. If the elapsed epoch had enough best points, this is a hyperbox that contains all best points and the very best point as interior points. Otherwise, the previous hyperbox is increased in size and re-centered around the current very best point, which may have changed during the epoch. In the tested SSB algorithm, the padding is 100 percent of the interval length in each dimension, when there are enough best points, and the old interval length is quadrupled, when there aren't. It turns out that in the former case, the simple scheme of updating the lower and higher bounds of the hyperbox in each dimension, for each new best point, works well in practice.

A *best point* is any point found during the second phase of an epoch that is the best this far in that epoch. The best points thus start over each epoch. The *very best point*, on the other hand, is the globally best point found in any round of any epoch.

The *first phase* of each epoch consists of the first quarter of its rounds. The rest of its rounds constitute the *second phase.* Other choices than one quarter were tested, but found less effective, albeit one third only marginally so. Typical figures are 60 epochs of 500 rounds each, but this can be varied with the search conditions, cf. Section IV. Higher dimensions require more rounds; fewer function evaluations entail fewer epochs and much fewer rounds.

The outer loop may seem somewhat ad hoc. It captures the idea, that if there has been non-trivial local improvements, search should focus on these improvements, yet also consider the globally best point found. We note that any new best point must be a bisection point, or the midpoint of a simplex, with a lower function value than its corner points. In one dimension, these cases coincide. For convex functions, such an interval must contain the minimum, which our introductory algorithm exploits. For non-convex functions, or in several dimensions, the area surrounding such a point merits further investigation.

### B. Inner Loop: Bisecting Simplexes

The inner loop over rounds bisects simplexes. Each bisection replaces one simplex with two new ones. In the first phase, the simplexes are processed as a first-in-first-out (FIFO) queue to create an initial grid. In the second phase, the next simplex to bisect is selected randomly with probability

$$\frac{s_k}{\sum_{k'} s_{k'}}$$

where $s_k$ is the score of the $k$th simplex. Binary trees provide an efficient way of stochastic selection that allows adding and deleting scored elements. Indexing the $K$ simplexes in a binary tree yields $O(\log_2 K)$ time complexity for lookup, addition, and deletion, whereas naively using a list swells this to $O(K)$.

We define the simplex scores $s_k$ as follows. Let $\{T_k = \langle \mathbf{x}_k^{(1)}, \ldots, \mathbf{x}_k^{(n+1)} \rangle\}$ be a collection of $n$-dimensional simplexes with (dropping the index $k$ for clarity),

$$\bar{\mathbf{x}} = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{x}^{(i)} \quad ; \quad \bar{f} = \frac{f(\bar{\mathbf{x}}) + \sum_{i=1}^{n+1} f\left(\mathbf{x}^{(i)}\right)}{n+2}$$

where $\bar{\mathbf{x}}$ is the midpoint and $\bar{f}$ is the average function value over the corners and the midpoint.

$$\begin{aligned} f^- &= \min\left(f(\bar{\mathbf{x}}), \min_i f\left(\mathbf{x}^{(i)}\right)\right) \quad ; \quad \delta = \frac{\bar{f} - f^-}{4} \\ f^\star &= f^- - \delta \quad ; \quad f^\star \leftarrow \max\left(0, f^\star - f_{\mathrm{vb}}\right) \end{aligned}$$

$f^\star$ is a combined measure of the lowest function value $f^-$ and an estimate $\delta$ of how much it might potentially decrease, judging by the average value $\bar{f}$. We make $f^\star$ offset-invariant by subtracting the lowest function value $f_{\mathrm{vb}}$, in the very best point, then cap it to be at least zero.

$$\begin{aligned} s &= l \cdot \exp\left(-\lambda f^\star\right) \quad ; \quad l = \max_{ij} \left| \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \right| \\ \lambda &= \lambda_0 \cdot \max\left(1, \frac{1}{f_{\mathrm{W}} - f_{\mathrm{vb}}}\right) \end{aligned}$$

The simplex score is $l \cdot \exp(-\lambda f^\star)$, where $l$ is the length of its longest edge. $\lambda$ is $\lambda_0$ times the reciprocal of the difference between the highest function value $f_{\mathrm{W}}$ of the corners of the bounding box and the lowest function value $f_{\mathrm{vb}}$. This renders the score scale-invariant. $\frac{\lambda}{\lambda_0}$ is capped to be at least one; $\lambda_0$ defaults to one. Thus, all simplexes must be rescored whenever a new very best point is found. As this happens rather seldom, it incurs very little overhead in practice.

Each round only creates two new simplexes: the longest edge of the selected simplex is bisected. All corners remain the same, save one of the two connected by this edge. Let these corners be $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$. The edge bisection point $\bar{\mathbf{x}}'$—the new corner—is also randomized to counter-act any symmetries of the objective function $f(\mathbf{x})$ in its argument $\mathbf{x}$:

$$\bar{\mathbf{x}}' = (0.5 + \theta)\mathbf{x}^{(i)} + (0.5 - \theta)\mathbf{x}^{(j)} \quad \text{for} \quad \theta \sim \mathrm{U}(-\alpha, \alpha)$$

An empirically good choice is $\alpha = 0.05$ (max 10% randomness). $\bar{\mathbf{x}}'$ replaces $\mathbf{x}^{(i)}$ in one offspring simplex and $\mathbf{x}^{(j)}$ in the other one. Thus only three new function values need be calculated for each bisection: $f(\bar{\mathbf{x}}')$ and the function values of the midpoints of the two new simplexes.

### C. Related Work

The SSB algorithm uses a typical stochastic optimization scheme. It maintains a set of elements, each with a positive score; randomly selects some elements based on the scores; uses these elements to explore the search space, often creating new elements in the process; and updates the set of elements and their scores according to the findings. The scheme is however here applied to regions of the search space, not to points in it, as in, e.g., [9], [10], [11], [12], [13], and [14].

The simplex method [15], Chapter 9, doesn't actually use simplexes. To create new points, controlled random search [11] generates random simplexes. These may overlap and are not guaranteed to cover the search space. Nor are they subdivided. DIRECT [16] uses hyperboxes that partition the search space. It avoids the $2^n$ complexity by directional search from their midpoints, ignoring their corners. Each hyperbox potentially containing the global minimum is trisected, rather than bisected, along each dimension in turn. There is no randomized selection.

## III. The Opimizers of the SciPy.Optimize Module

The optimizers from the Python SciPy.Optimize module that we tested were code named "Nelder-Mead," "Powell," "CG," "BFGS," "L-BFGS-B," "TNC," and "SLSQP." We did not provide the gradient nor the Hessian of the objective functions. Whenever an optimizer used these, it had to estimate them itself numerically.

The SciPy module also contains the "Anneal," "COBYLA," "dogleg," and "trust-ncg" optimizers. Anneal proved too slow and performed too poorly to be included. COBYLA was not robust enough for large-scale testing. The dogleg and trust-ncg optimizers require explicit gradients, which were not provided.

Nelder-Mead uses the simplex method, see [17], [18]. Powell is a modification of Powell's algorithm, a conjugate direction method, that performs sequential one-dimensional minimizations, see [19], [20]. CG uses a nonlinear conjugate gradient method by Polak and Ribiere, a variant of the Fletcher-Reeves algorithm described in [21], pp. 120–122.

BFGS uses the quasi-Newton method of Broyden, Fletcher, Goldfarb, and Shanno, see [21], p 136, and L-BFGS-B uses the L-BFGS-B algorithm for bound constrained minimization, see [22], [23]. TNC uses a combined Newton and conjugate gradient method, aka a truncated Newton method, see [21], p 168, and [24]. SLSQP uses sequential least squares programming, see [25].

## IV. Experiments

A very reasonable baseline, seeing that the SSB algorithm essentially adds structure to randomized search, is to evaluate it against unstructured randomized search conducted by each ScyPy optimizer. In each trial, the ScyPy optimizer was repeatedly restarted in a new random point, until a limit on the total number of function evaluations had been exceeded.

### A. Experimental Setup

We tested all *two-dimensional* objective functions of Figure 2 (last page), most of which are from [26]. Function 0 comes from [5], Function 20 from [27], and Functions 21, 22, 24, and 25 are of our own design. All functions have unique global minima, except Function 0, due to symmetry in $x$ and $x + y$, and Functions 12, 14, and 17, which have four global minima, due to symmetry in $\pm x, \pm y$. Functions 16 and 17 were corrected using [28]. We did not provide the gradient nor the Hessian of these objective functions.

We investigated the frugal function evaluation scenario, where computing function values comes at a premium, and restricted the number of function evaluations to four thousand. The domain was $[-80, 120] \times [-80, 120]$, which is typically larger than that of the test set: often $[-10, 10] \times [-10, 10]$ or even $[-5, 5] \times [-5, 5]$. It was made asymmetric in $x$ and $y$, since many test functions have their global minimum in $\mathbf{x} = \mathbf{0}$. As in [5] and [28], success was defined as finding any argument with a function value within $10^{-6}$ of the known global minimal value. See Section IV-C for a discussion on this success criterion.

No attempt was made to optimize the optimizers. The SciPy optimizers were used with their default parameter settings, and the search parameters of the SSB algorithm were taken over from [5], Section 5.3: $\lambda_0$ was set ten; epoch phase one consisted of the first five of its 50 rounds in total; once the system had made 4000 function evaluations, the current epoch and the algorithm were terminated.

In each trial, the SciPy optimizer was repeatedly run starting from a randomized point and the best candidate point of the trial was updated if the new one was better. The trial continued until 4000 function evaluations had been exceeded. The best result in the trial determined success or failure.

### B. Experimental Results

Table I shows their respective success rates in 1000 trials. To analyze these results, we need to consider the nature of each test objective function. These can be classified into:

- unimodal quadratic forms, Fcns 2, 6, 8;
- oligo-modal [3] polynomials, Fcns 3, 4, 5, 10, 18;
- multimodal damped trigonometrics, Fcns 0, 1, 12–17, 25;
- mixed trigonometrics and quadratics, Fcns 9, 11, 20, 24;
- other (unimodal) functions, Fcns 7, 21, 22.

In exactly half of the cases (Functions 0, 1, 9, 11–17, and 22–25), the SSB algorithm stands head and shoulders above the competition, except for Powell's algorithm in two of these cases, namely Functions 1 and 9. *This includes all seriously multimodal functions.* Clearly, when faced with numerous local optima, the extra structure afforded by the SSB algorithm outweighs the benefit of highly accurate local search.

Conversely, it includes no unimodal or oligo-modal function, except Function 22. This stands to reason, as any adept local optimizer should succeed for these functions, even when starting from a randomized point some distance away. See Section IV-C for a discussion of Functions 2, 21, and 22.

Function 7 is very hard, and defeats all optimizers. It has a concave valley, kinks, and a very anisotropic variable coupling. The gradient is ill-defined and unbounded in the valley bottom, and especially ill-behaved in the optimum. *Memento mori.*

Viewing the results from another angle, we note that the SSB algorithm performs under 50% in only five cases of 22. In two of these, it still outperforms all other algorithms, and in a third case, all algorithms come up empty-handed. In one of the two remaining cases, Function 3, the SSB algorithm

[3] . . . apologies for mixing Greek and Latin roots. . .

| | Optimizer | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Fcn | BFGS | CG | L-BFGS-B | Simplex | Powell | SLSQP | TNC | SSB |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **27** |
| 1 | 45 | 57 | 76 | 17 | 835 | 52 | 52 | **896** |
| 2 | **1000** | **1000** | **1000** | **1000** | **1000** | **1000** | **1000** | 989 |
| 3 | 999 | **1000** | **1000** | **1000** | **1000** | 996 | 921 | 188 |
| 4 | 521 | 725 | 741 | **998** | 109 | 975 | 716 | 661 |
| 5 | 949 | 493 | **1000** | **1000** | 951 | **1000** | **1000** | 765 |
| 6 | **1000** | **1000** | **1000** | **1000** | **1000** | **1000** | **1000** | 862 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | **1000** | **1000** | **1000** | **1000** | 966 | **1000** | **1000** | 872 |
| 9 | 50 | 89 | 140 | 545 | **1000** | 149 | 34 | *897* |
| 10 | 448 | **1000** | 960 | **1000** | 984 | 920 | **1000** | 844 |
| 11 | 297 | 239 | 289 | 136 | 222 | 269 | 202 | **638** |
| 12 | 397 | 259 | 440 | 380 | 251 | 391 | 310 | **890** |
| 14 | 91 | 112 | 251 | 52 | 136 | 201 | 57 | **808** |
| 16 | 42 | 34 | 75 | 9 | 109 | 159 | 163 | **857** |
| 17 | 83 | 68 | 147 | 17 | 10 | 243 | 34 | **815** |
| 18 | 996 | **1000** | **1000** | **1000** | **1000** | **1000** | 998 | 906 |
| 20 | 125 | 156 | 160 | 975 | **1000** | 460 | 48 | 256 |
| 21 | **1000** | 750 | **1000** | 0 | **1000** | **1000** | 585 | 829 |
| 22 | 0 | 0 | 0 | 0 | 304 | 0 | 0 | **626** |
| 24 | 2 | 0 | 13 | 14 | 111 | 0 | 2 | **139** |
| 25 | 121 | 19 | 265 | 73 | 205 | 221 | 635 | **803** |

TABLE I
SUCCESS RATE IN 1000 TRIALS OF THE SCIPY OPTIMIZERS AND THE SSB ALGORITHM.

is hampered by quadratic terms and a dominating variable coupling, and it lags behind the competition.

The other remaining case is Function 20. It has dominating quadratic terms and strong trigonometric confusion terms. The SSB algorithm outperforms half of the SciPy optimizers, but not the SLSQP, Simplex, and Powell optimizers, the latter two which excel. By contrast, all SciPy optimizers perform under 50% in more than half of the cases, except Powell's algorithm, which performs under 50% in exactly half of the 22 cases.

### C. A Digression on Success Regions

The success criterion $|f(\mathbf{s}) - f(\mathbf{x}^*)| < \epsilon$, where $\mathbf{x}^*$ is a known global optimum, is certainly reasonable for practitioners, who care mainly about finding a good enough solution $\mathbf{s}$. But when evaluating optimization algorithms, one must be careful not to draw incorrect conclusions from this criterion.

For example, when comparing performances on Functions 2, 21, and 22 in Table I, one might be tempted to conclude that the discontinuity in the gradient in the global optimum caused by the absolute value does some damage to gradient-based methods, and wreaks havoc with the simplex method, while the additional concavity of the square root foils all algorithms, save Powell's method and the SSB algorithm.

This analysis fails to take into account that the success region of Function 2 is a circle with diameter $2\sqrt{\epsilon}$, while that of Function 21 is a square with diagonal $2\epsilon$, and that of Function 22 is a concave diamond shape with diagonal $2\epsilon^2$. The alchemy-like Figure 1 illustrates this for the unrealistically large $\epsilon = 0.5$. The ratios of their surface areas are $\pi\epsilon : 2\epsilon^2 : \frac{2}{3}\epsilon^4$. This makes the latter two increasingly harder to hit, which is a potentially greater difficulty for the optimizers, given strict limitations on the number of function evaluations.

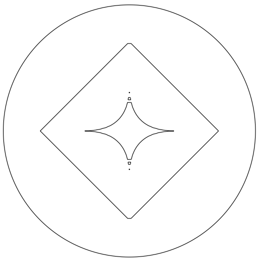


Fig. 1. $|f(x, y) - f(0, 0)| = 0.5$ for Fcns 2, 21, and 22.

## V. SUMMARY AND CONCLUSIONS

We evaluated the stochastic simplex bisection (SSB) algorithm against the optimizers of the Python SciPy.Optimize module on a prominent test set. The former employs a common stochastic optimization scheme, but unlike other stochastic approaches, it applies the scheme to search space regions, rather than to individual points. The latter are readily available, high-quality, local optimizers. This is the first evaluation of the SSB algorithm against other optimizers.

The experiments were conducted in two dimensions on a prominent test set. The domain was mostly larger, by an order of magnitude in each direction, than the domains indicated by the test set, and the number of function evaluations was limited to a few thousand. The SSB algorithm used the latter as a termination criterion and returned its very best point. Each ScyPy optimizer was repeatedly restarted in a random point, until it had exceeded the function evaluation limit, and the overall best result was returned. In both cases, if the returned solution was within $10^{-6}$ of the known global minimal value, it was deemed a successful trial. Each optimizer was tested 1000 times on each test function.

The SSB algorithm greatly outperformed all SciPy optimizers, save one, in exactly half the cases. It did slightly worse on polynomial functions, but excelled at trigonometric ones, highlighting its multimodal prowess. And unlike the SciPy optimizers, it sustained a high success rate.

We conclude that the SciPy optimizers would benefit from a more informed metaheuristic strategy and that the SSB algorithm would profit from quicker local convergence and better multidimensional capabilities. Conversely, the local convergence of the SciPy optimizers is impressive and the multimodal capabilities of the SSB algorithm in separable dimensions are uncanny.

## Acknowledgment

This work was funded by the BAULT network and the METALOGUE project. BAULT, building and using language technology, is a multidisciplinary research community at the University of Helsinki. METALOGUE is a Seventh Framework Programme collaborative project funded by the European Commission, grant agreement number 611073 (http://www.metalogue.eu).

## References

[1] M. Wahde, *Biologically Inspired Optimization Algorithms*. WIT Press, 2008.

[2] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Machine Learning*, vol. 3, no. 2-3, pp. 95–99, 1988. doi: 10.1023/A:1022602019183

[3] J. Kennedy and R. Eberhart, "Particle swarm optimization," 1995. doi: 10.1109/ICNN.1995.488968

[4] M. Dorigo, V. Maniezzo, and A. Colorni, "The ant system: Optimization by a colony of cooperating agents," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 26, no. 1, pp. 29–41, 1996. doi: 10.1109/3477.484436

[5] C. Samuelsson, "The stochastic simplex bisection algorithm," in *Procs. 15th Int.'l Conf. Computational Science*, ser. ICCS/15. Elsevier, 2015. doi: 10.1016/j.procs.2015.05.215 pp. 855–864. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915010236

[6] scipy, "scipy.optimize," http://http://docs.scipy.org/doc/scipy/reference/optimize.html#module-scipy.optimize, 2015, accessed: 2015-04-16.

[7] M. Haiman, "A simple and relatively efficient triangulation of the $n$-cube," *Discrete & Computational Geometry*, vol. 6, no. 1, pp. 287–289, 1991. doi: 10.1007/BF02574690. [Online]. Available: http://dx.doi.org/10.1007/BF02574690

[8] R. B. Hughes and M. R. Anderson, "Simplexity of the cube," *Discrete Mathematics*, vol. 158, no. 13, pp. 99 – 150, 1996. doi: http://dx.doi.org/10.1016/0012-365X(95)00075-8. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0012365X95000758

[9] Q. Duan, S. Sorooshian, and V. Gupta, "Effective and efficient global optimization for conceptual rainfall-runoff models," *Water resources research*, vol. 28, no. 4, pp. 1015–1031, 1992. doi: 10.1029/91WR02985

[10] N. Hansen and S. Kern, "Evaluating the CMA evolution strategy on multimodal test functions," in *Parallel Problem Solving from Nature VIII*, X. Yao *et al.*, Eds. Springer, 2004. doi: 10.1007/978-3-540-30217-9_29 pp. 282–291.

[11] P. Kaelo and M. M. Ali, "Some variants of the controlled random search algorithm for global optimization," *J. Optim. Theory Appl*, vol. 130, no. 2, pp. 253–264, 2006. doi: 10.1007/s10957-006-9101-0

[12] K. V. Price, R. M. Storn, and J. A. Lampinen, *Differential Evolution - A Practical Approach to Global Optimization*, ser. Natural Computing. Springer-Verlag, 2006, iSBN 540209506.

[13] A. I. Vaz and L. N. Vicente, "A particle swarm pattern search method for bound constrained global optimization," *J. of Global Optimization*, vol. 39, no. 2, pp. 197–219, Oct. 2007. doi: 10.1007/s10898-007-9133-5. [Online]. Available: http://dx.doi.org/10.1007/s10898-007-9133-5

[14] X.-S. Yang, "Firefly algorithms for multimodal optimization," in *Procs. 5th Int.'l Conf. Stochastic Algorithms: Foundations and Applications*, ser. SAGA'09. Springer-Verlag, 2009. doi: 10.1007/978-3-642-04944-6_14. ISBN 3-642-04943-5, 978-3-642-04943-9 pp. 169–178. [Online]. Available: http://dl.acm.org/citation.cfm?id=1814087.1814105

[15] N. Andréasson, A. Egrafov, and M. Patriksson, *An Introduction to Continuous Optimization*. Studentlitteratur, 2005.

[16] D. R. Jones, C. D. Perttunen, and B. E. Stuckman, "Lipschitzian optimization without the Lipschitz constant," *J. Optim. Theory Appl.*, vol. 79, no. 1, pp. 157–181, Oct. 1993. doi: 10.1007/BF00941892. [Online]. Available: http://dx.doi.org/10.1007/BF00941892

[17] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, pp. 308–313, 1965. doi: 10.1093/comjnl/7.4.308

[18] M. H. Wright, "Direct Search Methods: Once Scorned, Now Respectable," in *Numerical Analysis 1995 (Proceedings of the 1995 Dundee Biennial Conference in Numerical Analysis)*, ser. Pitman Research Notes in Mathematics, D. F. Griffiths and G. A. Watson, Eds., vol. 344. Boca Raton, Florida: CRC Press, 1996. doi: 10.1.1.47.6891 pp. 191–208.

[19] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *The Computer Journal*, vol. 7, no. 2, pp. 155–162, 1964. doi: 10.1093/comjnl/7.2.155. [Online]. Available: http://comjnl.oxfordjournals.org/content/7/2/155.abstract

[20] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd ed. New York, NY, USA: Cambridge University Press, 2007. ISBN 0521880688, 9780521880688

[21] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.

[22] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995. doi: 10.1137/0916069

[23] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, Dec. 1997. doi: 10.1145/279232.279236. [Online]. Available: http://doi.acm.org/10.1145/279232.279236

[24] S. Nash, "Newton-type minimization via the Lanczos method," *SIAM Journal on Numerical Analysis*, vol. 21, no. 4, pp. 770–788, 1984. doi: DOI:10.1137/0721052

[25] D. Kraft, *A software package for sequential quadratic programming*, ser. Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht. Wiss. Berichtswesen d. DFVLR, 1988. [Online]. Available: http://books.google.fr/books?id=4rKaGwAACAAJ

[26] Wikipedia, "Test functions for optimization," http://en.wikipedia.org/wiki/Test_functions_for_optimization.html, 2014, accessed: 2014-12-14.

[27] K. Mullen, D. Ardia, D. Gil, D. Windover, and J. Cline, "DEoptim: An R package for global optimization by differential evolution," *J. of Stat. Software*, vol. 40, no. 6, pp. 1–26, 2011. [Online]. Available: http://www.jstatsoft.org/v40/i06/

[28] A. Gavana, "Global optimization benchmarks and AMPGO," http://infinity77.net/global_optimization, 2015, accessed: 2015-01-09.

## Objective Functions

Fcns 21, 22, 24, 25 are novel; Fcn 0 from [5]; Fcn 20 from [27]; remainder from [26]. Fcns 16 and 17 corrected using [28].

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) = \prod_{k=1}^{n} f_0\left(\sum_{i=1}^{k} x_i\right) \quad \text{with} \tag{0}$$

$$f_0(x) = \sin(\sqrt{1+x^2}) \cdot \cos(2x(x+1)) \cdot \frac{\ln(1+x^2)}{\sqrt{1+x^2}}$$

$$f(\mathbf{x}) = 20\left(1-\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\right)\right) + e - \exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right) \tag{1}$$

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) = \sum_{i=1}^{n} x_i^2 \tag{2}$$

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) = \sum_{i=1}^{n-1}\left(100\cdot(x_{i+1}-x_i^2)^2 + (x_i-1)^2\right) \tag{3}$$

$$f(x,y) = (1.5-x+xy)^2 + (2.25-x+xy^2)^2 + (2.625-x+xy^3)^2 \tag{4}$$

$$\begin{aligned} f(x,y) = {} & \left(1+(x+y+1)^2(19-14x+3x^2-14y+6xy+3y^2)\right)\cdot \\ & \cdot\left(30+(2x-3y)^2(18-32x+12x^2+48y-36xy+27y^2)\right) \end{aligned} \tag{5}$$

$$f(x,y) = (x+2y-7)^2 + (2x+y-5)^2 \tag{6}$$

$$f(x,y) = 100\sqrt{|y-0.01x^2|} + 0.01|x+10| \tag{7}$$

$$f(x,y) = 0.26(x^2+y^2) - 0.48xy \tag{8}$$

$$f(x,y) = \sin^2(3\pi x) + (x-1)^2\left(1+\sin^2(3\pi y)\right) + (y-1)^2\left(1+\sin^2(2\pi x)\right) \tag{9}$$

$$f(x,y) = 2x^2 - 1.05x^4 + \frac{x^6}{6} + xy + y^2 \tag{10}$$

$$f(x,y) = -\cos(x)\cos(y)\exp\left(-(x-\pi)^2-(y-\pi)^2\right) \tag{11}$$

$$f(x,y) = -0.0001\left(|\sin(x)\sin(y)|\exp\left(\left|100-\frac{\sqrt{x^2+y^2}}{\pi}\right|\right)+1\right)^{0.1} \tag{12}$$

$$f(x,y) = -|\sin(x)\sin(y)|\exp\left(\left|1-\frac{\sqrt{x^2+y^2}}{\pi}\right|\right) \tag{14}$$

$$f(x,y) = 0.5 + \frac{\sin^2(x^2-y^2)-0.5}{1+0.001(x^2+y^2)^2} \tag{16}$$

$$f(x,y) = 0.5 + \frac{\cos^2(\sin(x^2-y^2))-0.5}{1+0.001(x^2+y^2)^2} \tag{17}$$

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) = \frac{1}{2}\sum_{i=1}^{n}\left(x_i^4-16x_i^2+5x_i\right) \tag{18}$$

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) = \sum_{i=1}^{n}\left(x_i^2+10\cdot(1-\cos(2\pi x_i))\right) \tag{20}$$

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) = \sum_{i=1}^{n}|x_i| \tag{21}$$

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) = \sum_{i=1}^{n}\sqrt{|x_i|} \tag{22}$$

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) = \sum_{i=1}^{n}|x_i| + \left(10+x_i^2\right)\cdot(1-\cos(2\pi x_i)) \tag{24}$$

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) = \sum_{i=1}^{n} 1-\text{sink}(x_i) \quad \text{with} \;\; \text{sink}(x) = \frac{\sin(x)}{x} \tag{25}$$

Fig. 2.

Fig. 2 List of objective functions

DOI: 10.15439/2015F7
ACSIS, Vol. 5

# Hybrid Metaheuristic for Portfolio Selection: Comparison with an exact solver and search space analysis

Giacomo di Tollo
Dipartimento di Economia, Universitá Ca' Foscari
Cannaregio 873, 30121-Venezia, Italia
Email: giacomo.ditollo@unive.it

***Abstract*—In this paper we use a metaheuristic approach to solve the Portfolio Selection problem, in a constrained formulation which is NP-hard and difficult to be solved by standard optimization methods. We are comparing the algorithm's performances with an exact solver and we are showing that different mathematical formulations lead to different algorithm's behaviour. Results show that our approach can be efficiently used to solve the problem at hand, and that a sound basin of attraction analysis may help developers and practitioners to design the experimental analysis.**

## I. Introduction

PORTFOLIO Selection main formulation dates back from the fifties and is concerned with selecting, out of a given set of assets, which assets to invest in and by how much, in order to minimise a risk measure for a given minimum required target return. Many measures can be used for assessing the risk, but the variance of portfolio's return was used in the seminal work by Markowitz [24] and is still the most used.

Portfolio Selection Problem (PSP) can be viewed as an optimisation problem to be described by three objects: variables, objective, and constraints. Every object have to be instantiated by a choice, and the combination of these choices leads to a specific formulation (model) of the problem, hence to different optimisation results. For instance, as stated by di Tollo and Roli[8], two main choices are possible for variables: continuous[15], [29], [31], [28] and integer[30], [21]. Choosing continuous variables is quite 'natural' and its representation is independent of the actual budget, while integer values (ranging between zero and the maximum available budget, or equal to the number of 'rounds') allow us to add constraints taking into account actual budget, minimum lots and to tackle other objective functions to better explain the problem at hand. As for the different results, the integer formulation is more suitable to explain the behaviour of rational operators such small investors, whose activity is strongly influenced by integer constraint[22].

Furthermore, the same representation can be modelled by means of different formulations, e.g., by adding auxiliary variables[20], symmetry breaking[27] or redundant[32] constraints. Although these extensions have no effect on the certified optimal solution found, they may affect the optimisation procedure. For example, it has been shown that symmetry breaking constraints have negative effect on local search performances[27].

In this work we will investigate how the use of different formulations for the very same problem can lead to different behaviours of the algorithm used. We will study this aspect by solving the Portfolio Selection Problem by metaheuristics[4], [8], which are general problem-solving strategies conceived as high level strategies that coordinate the behaviour of lower level heuristics, and provide the user with a solution which cannot be certified to be optimum, still it represents a good compromise when the optimal solution is impossible to be found. Through the use of meta-heuristic, and using the paradigm of separation between model and algorithm[17], we will show that different formulations affect the algorithm's performances and study the motivation of this phenomenon.

The paper will start recalling Portfolio Theory in Section II, before introducing the concept of meta-heuristics in Section III. Then we will introduce a meta-heuristic approach for the Portfolio Selection Problem in Section IV, while Section V will introduce the principles Search Space Analysis is based upon. Search Space Analysis will be applied to our instances on Section VI, before concluding with Section VII.

## II. Portfolio Selection Basis

We associate to each asset belonging to a set $A$ of $n$ *assets* ($A = \{a_1, \dots, a_n\}$) a real-valued *expected*

978-83-60810-66-8/$25.00©2015, IEEE

*return* $r_i$, and the corresponding return variance $\sigma_i$. We furthermore associate, to each pair of assets $\langle a_i, a_j \rangle$, a real-valued return *covariance* $\sigma_{ij}$. We are furthermore given a value $r_e$ representing the minimum required return.

In this context, a portfolio is defined as the $n$-sized real vector $X = \{x_1, \ldots, x_n\}$ in which $x_i$ represents the relative amount invested in asset $a_i$. For each portfolio we can define its variance as $\sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} x_i x_j$ and its return as $\sum_{i=1}^{n} r_i x_i$. In the original formulation [25], PSP is formulated as the minimization of portfolio variance, imposing that the portfolio's return must be not smaller than $r_e$, leading to the following optimisation problem:

$$\min \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} x_i x_j, \tag{1}$$

$$s.t. \quad \sum_{i=1}^{n} r_i x_i \geq r_e, \tag{2}$$

$$\sum_{i=1}^{n} x_i = 1, \tag{3}$$

$$x_i \geq 0 \qquad (i = 1, \ldots, n). \tag{4}$$

The aforecited return constrained is introduced in constraint (2); constraint (3) is referred to as *budget constraint*, meaning that all the capital must be invested; constraint (4) imposes that variables have to be non-negative (i.e., short sales are not allowed).

If we define a finite set of values for $r_e$ and solve the problem for all defined $r_e$ values, we obtain the *Unconstrained Efficient Frontier* (UEF), in which the minimum risk value is associated to each $r_e$.

This formulation may be improved to grasp financial market features, by introducing a binary variable $Z$ for each asset ($z_i = 1$ if asset $i$ is on the portfolio, $0$ otherwise). Additional constraints which can be added to the basic formulation are:

- **Cardinality constraint**, used either to impose an upper bound $k$ to the cardinality of assets in the portfolio

$$\sum_{i=1}^{n} z_i \leq k, \tag{5}$$

  or to force the resulting portfolio to contain exactly $k$ assets:

$$\sum_{i=1}^{n} z_i = k_{max}. \tag{6}$$

  This constraint is important for practitioners in order to reduce the portfolio management costs.
- **Floor and ceiling constraints**, used to set, for each asset, the minimum ($\varepsilon_i$) and maximum ($\delta_i$) quantity allowed to be held in the portfolio

$$\varepsilon_i z_i \leq x_i \leq \delta_i z_i. \tag{7}$$

  Those constraints are used to ensure diversification and to avoid tiny portions of assets in the portfolios, which would make their management difficult and lead to unnecessary transaction costs.
- **Preassignments.** This constraint is used to express subjective preferences: we want certain specific assets to be held in the portfolio, by determining a $n$-sized binary vector $P$ (i.e., $p_i = 1$ if $a_i$ has to be held in the portfolio) and imposing the following:

$$z_i \geq p_i \qquad (i = 1, \ldots, n). \tag{8}$$

## III. Meta-heuristics

As stated in the Introduction, in this work we are solving the PSP by using meta-heuristics[4], which can be defined as high-level strategies that coordinate the action of low-level algorithms (heuristics) in order to find near-optimal solutions for combinatorial optimization problem. They are used when it is impossible to find the certified optimum solution in a reasonable amount of time, and their features can be outlined as follows:

- They are used to explore the search space and to determine principles to guide the action of subordinated heuristics.
- Their level of complexity ranges from a simple escape-mechanism to complex populations procedures.
- They are stochastic, hence escape and restart procedures have to be devised in the experimental phase.
- The concepts they are built upon allow an abstract descriptions, that is useful to design hybrid procedures.
- They are not problem-specific, but additional components may be used to exploit the structure of the problem or knowledge acquired during the search process.
- They may make use of problem-specific knowledge in the form of heuristics that are controlled by the upper level strategy.

The main paradigm meta-heuristics are build upon is the *intensification-diversification* paradigm, meaning that they should incorporate a mechanism to balance the exploration of promising regions of the search landscape (intensification) and the identification of new areas in the search landscape (diversification). The way of implementing this balance is different depending on the specific meta-heuristic used. A completed description is out of the scope of this paper, and we forward the interested reader to Hoos and Stuetzle[18].

## IV. Our Approach for Portfolio Choice

We are using the solver introduced by di Tollo et al.[7], [9] to tackle a constrained PSP, in which the Markowitz' variance minimisation in a continuous formulation is enhanced by adding constraints (4), (6) and (7), leading to the following formulation:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} x_i x_j, \tag{9}$$

subject to

$$\sum_{i=1}^{n} r_i x_i \geq r_e, \tag{10}$$

$$\sum_{i=1}^{n} x_i = 1, \tag{11}$$

$$x_i \geq 0 \qquad i = 1 \ldots n, \tag{12}$$

$$k_{min} \leq \sum_{i=1}^{n} z_i \leq k_{max}, \tag{13}$$

$$\varepsilon_i z_i \leq x_i \leq \delta_i z_i, \tag{14}$$

$$x_i \leq z_i \quad i = 1 \ldots n. \tag{15}$$

Where $k_{min}$ and $k_{max}$ are respectively lower and upper bounds on cardinality. This problem formulation contains two classes of decision variables: integer (i.e., $Z$) and continuous (i.e., $X$). Hence, it is possible to devise an hybrid procedure in which each variable class is tackled by a different component. Starting from this principle, we have devised a master-slave decomposition, in which a meta-heuristic procedure is used in order to determine, for each search step, assets contained in the portfolio ($Z$). Once the assets contained in the portfolio are decided, the corresponding continuous $X$ values can be determined with proof of optimality. Hence at each step, after having selected which assets to be taken into account, we are resorting to a the Goldfarb-Idnani algorithm for quadratic programming (QP) [16] to determine their optimum value. The stopping criterion and escape mechanism depend on the metaheuristic used, which will be detailed in what follows.

As explained in section VI, this master-slave decomposition has a dramatic impact on the meta-heuristic performance due to the different structure determined by this formulation, in which the basin of attraction are greater than the ones determined by a monolithic approach based on the same meta-heuristic approaches. In what follows we are outlining the components of our meta-heuristic approach.

- **Search space** Since the *master* meta-heuristic component takes into account the $Z$ variables only, the search space $S$ is composed of the $2^n$ portfolios that are feasible w.r.t cardinality and pre-assignment constraints, while other constraints are directly ensured by the *slave* QP procedure. If the QP procedure does not succeed in finding a feasible portfolio, a greedy procedure is used to find the portfolio with maximum return and minimum constraint violations.
- **Cost function** In our approach the cost function corresponds to the objective function of the problem $\sigma^2$, and is computed, at each step of the search process, by the *slave* QP procedure.
- **Neighborhood relations** As in di Tollo et al.[9], we are using three neighborhood relations in which the neighbor portfolio are generated by *adding*, *deleting* or *replacing* one asset: the neighbor is created by defining the asset pair $\langle i, j \rangle (i \neq j)$, inserting asset $i$, and deleting asset $j$. Addition is implemented by setting $j = 0$; deletion is implemented by $i = 0$.
- **Initial solution** The initial solution must be generated to create a configuration of $Z$. Since the we aim to generate an approximation of the unconstrained efficient frontier, we are devising three different procedures for generating the starting portfolio, which are used w.r.t. different $r_e$ values: MaxReturn (in which the starting portfolio corresponds to the maximum return portfolio, without constraints on the risk); RandomCard (in which cardinality and assets are randomly generated); and WarmRestart (in which the starting portfolio corresponds to the optimal solution found for the previous $r_e$ value). MaxReturn is used when setting the highest $r_e$ value (i.e., first computed value); for all other $r_e$ values both RandomCard and WarmRestart have been used.

### A. Solution techniques

As specific meta-heuristics for the *master* procedure, we have used Steepest Descent (SD), First Descent (FD) and Tabu Search (TS). SD and FD are considered as the most simple meta-heuristic strategies, since they accept the candidate solution only when its cost function is better than the current one, otherwise the search stops. They differ to each other in the neighborhood exploration, since in SD all neighbors are generated and the best one is compared to the current solution, while in FD the first better solution found is selected as current one. TS enhances this schema by selecting, as the new current solution, the best one amongst the neighborhood, and using an additional memory (Tabu list) in which forbidden states (i.e., former solutions) are stored, so that they cannot be generated as neighbors. In our implementation, we have used a dynamic-sized tabu list, in which solutions are put in the Tabu list for a randomly generated period of time. The length range of the Tabu list has been determined by using F-Race [3], and has been set to $[3, 10]$.

The three meta-heuristics components have been coded in C++ by Luca Di Gaspero and Andrea Schaerf and are available upon request.

As for the *slave* Quadratic programming procedure, we have used the Goldfarb and Idnani dual set method [16] to determine the optimal $X$ values corresponding to $Z$ values computed by the *master* meta-heuristic component. This method has been coded in C++ by Luca Di Gaspero: it is available upon request, and has achieved good performances when matrices at hand are dense.

To sum up, the *master* meta-heuristic component determines the actual configuration of $Z$ variables (i.e., point of the search space), the *slave* QP procedure computes the cost of the determined configuration, which is accepted (or not) depending on the mechanism embedded in FD, SD or TS.

### B. Benchmark instances

We have used instances from the repository `ORlib` (http://people.brunel.ac.uk/~mastjjb/jeb/info.html) and instances used in Crama and Schyns[6], which have been kindly provided to us by the authors. The UEF for the `ORlib` instances is provided in the aforementioned website; the UEF for instances from Crama and Schyns[6] has been generated by us by using our *slave* QP procedure. In both cases, the resulting UEF consists of 100 portfolios corresponding to 100 equally distributed $r_e$ values. Benchmarks' main features are highlighted in Table I.

By measuring the distance of the obtained frontier (CEF) from the UEF we obtain the *average percentage loss*, which is an indicator of the solution quality and which is defined as:

$$\mathsf{apl} = \frac{100}{p} \sum_{l=1}^{p} (V(r_e) - V_U(r_e))/V_U(r_e) \qquad (16)$$

in which $r_e$ is the minimum required return, $p$ is the frontier cardinality, $V(r_e)$ and $V_U(r_e)$ are the values of the function $F$ returned by the solver and the risk on the UEF.

### C. Experimental analysis

Our experiments have been run on a computer equipped with a Pentium 4 (3.2 GHz), and in what follows we are showing results obtained on both instance classes. In order to assess the quality of our approach, in the following tables we also report results obtained by other works tackling the same instances. Table II reports results over `ORlib`instances, showing that our approach outperforms the meta-heuristic approach by Schaerf[29], and compares favourably with Moral-Escudero et al.[26]

Table III compares our results with the one by Crama and Schyns[6]: solutions found by our hybrid approach have better quality than the ones found by SA [6], but running times are higher, due to our QP procedure and to our complete neighbourhood exploration, which are not implemented by Crama and Schyns.

We have also compared our approach with Mixed Integer Non-linear Programming (MINLP) solvers, by encoding the problem in AMPL [14] and solving it using CPLEX 11.0.1 and MOSEK 5. We have run the MINLP solvers over ORLib instances, and compared their results with SD+QP (10 runs), obtaining the same solutions in the three approaches, hence showing that our approach is able to find the optimal solution in a low computational time. Computational times for SD+QP and for the MINLP solvers are reported in Table IV and in Figure 1. We can notice that for big-sized instances exact solvers require higher computation time to generate points in which cardinality constraints are binding (i.e., left part of the frontier). Our approach instead scales very well w.r.t. size and provides results which are comparable.

We can conclude this section by observing that SD+QP provides as satisfactory results as the more complex TS+QP. Since Tabu Search is conceived to better explore the search space, this can be considered rather surprising. The next sections will enlighten us about this phenomenon.

TABLE I
OUR INSTANCES.

| | ORlib dataset | | | | Crama and Schyns dataset | | |
|---|---|---|---|---|---|---|---|
| ID | Country | Assets | AVG(UEF)risk | ID | Country | Assets | AVG(UEF)risk |
| 1 | Hong Kong (Hang Seng) | 31 | $1.55936 \cdot 10^{-3}$ | S1 | USA (DataStream) | 20 | 4.812528 |
| 2 | Germany (DAX 100) | 85 | $0.412213 \cdot 10^{-3}$ | S2 | USA (DataStream) | 30 | 8.892189 |
| 3 | UK (FTSE 100) | 89 | $0.454259 \cdot 10^{-3}$ | S3 | USA (DataStream) | 151 | 8.64933 |
| 4 | USA (S&P 100) | 98 | $0.502038 \cdot 10^{-3}$ | | | | |
| 5 | Japan (NIKKEI) | 225 | $0.458285 \cdot 10^{-3}$ | | | | |

TABLE II
RESULTS OVER ORLIB INSTANCES.

| | FD+QP | | SD+QP | | TS+QP | | TS[29] | | GA+QP[26] | |
|---|---|---|---|---|---|---|---|---|---|---|
| Inst. | min apl | time | min apl | time | min apl | time | min apl | time | min apl | time |
| 1 | 0.00366 | 1.5 | 0.00321 | 3.1 | 0.00321 | 29.1 | 0.00409 | 251 | 0.00321 | 415.1 |
| 2 | 2.66104 | 9.6 | 2.53139 | 14.1 | 2.53139 | 100.9 | 2.53617 | 531 | 2.53180 | 552.7 |
| 3 | 2.00146 | 10.1 | 1.92146 | 16.1 | 1.92133 | 114.4 | 1.92597 | 583 | 1.92150 | 886.3 |
| 4 | 4.77157 | 11.2 | 4.69371 | 18.8 | 4.69371 | 130.5 | 4.69816 | 713 | 4.69507 | 1163.7 |
| 5 | 0.24176 | 25.3 | 0.20219 | 45.9 | 0.20210 | 361.8 | 0.20258 | 1603 | 0.20198 | 1465.8 |

TABLE III
RESULTS OVER CRAMA AND SCHYNS INSTANCES.

| | FD+QP | | | SD+QP | | | TS+QP | | | SA[6] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Inst. | apl | | time | apl | | time | apl | | time | apl | | time |
| S1 | 0.72 | 0.094 | 0.3 | 0.35 | 0.0 | 1.4 | 0.35 | 0.0 | 4.6 | 1.13 | 0.13 | 3.2 |
| S2 | 1.79 | 0.22 | 0.5 | 1.48 | 0.0 | 3.1 | 1.48 | 0.0 | 8.5 | 3.46 | 0.17 | 5.4 |
| S3 | 10.50 | 0.51 | 10.2 | 8.87 | 0.003 | 53.3 | 8.87 | 0.0003 | 124.3 | 16.12 | 0.43 | 30.1 |

TABLE IV
COMPUTATIONAL TIMES OVER ORLIB INSTANCES 1–4, SD+QP AND MINLP.

| Instance | avg(SD + QP) | CPLEX 11 | MOSEK 5 |
|---|---|---|---|
| 1 | 3.1s | 2.1s | 15.8s |
| 2 | 14.7s | 397.1s | 5.0s |
| 3 | 18.0s | 890.7s | 1,903.3s |
| 4 | 20.9s | 169,461.0s | 239,178.4s |

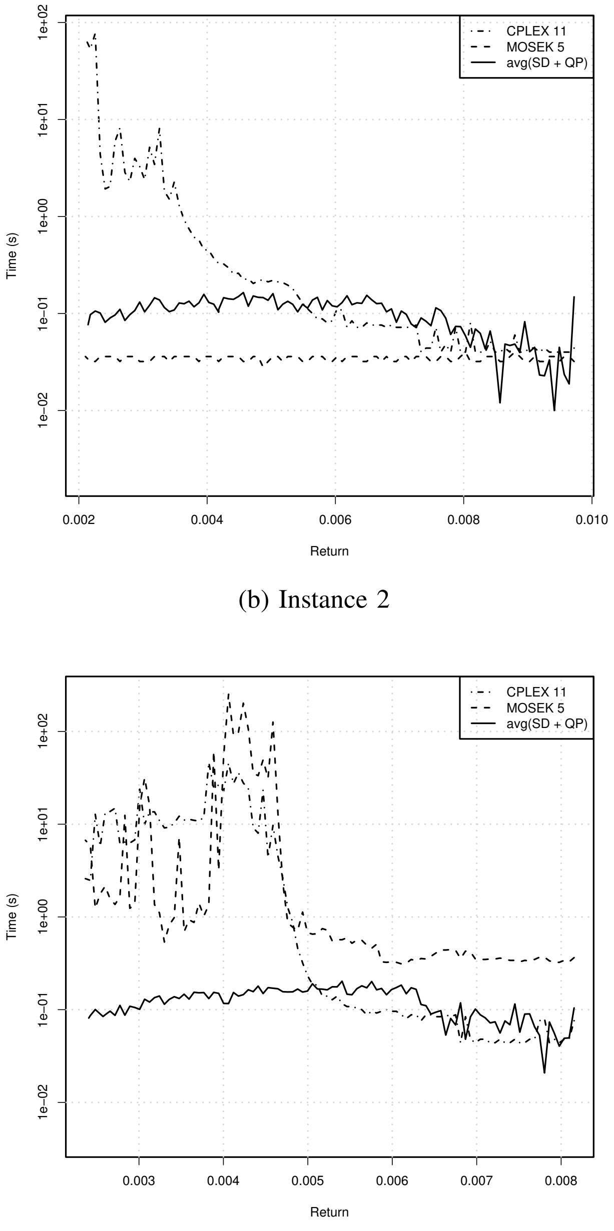
## V. SEARCH SPACE ANALYSIS

Search Space Analysis relies on the concept of *basin of attraction* (BOA), and is aimed to understand the features of the search space, when they are not deductible using exhaustive approaches.

In our meta-heuristic model, we are defining BOAs of search graph nodes. For this definition to be valid for any state of the search graph[2], we are relaxing the requirement that the goal state is an attractor. Therefore, the basin of attraction will also depend on the particular termination condition of the algorithm. In the following examples, we will suppose to end the execution as soon as a stagnation condition is detected, i.e., when no improvements are found after a maximum number of steps. In what follows we are following the definitions expressed by Roli[2], and we are applying our analysis to deterministic systems, before extending it to stochastic systems.

**Definition** Given a deterministic algorithm $\mathcal{A}$, the basin of attraction $\mathcal{B}(\mathcal{A}|s)$ of a point $s$, is defined as the set of states that, taken as initial states, give origin to trajectories that include point $s$.

Let $S^*$ be the set of global optima: for each $s \in S^*$ there exist a basin of attraction, and their union $I^* = \bigcup_{i \in S^*} \mathcal{B}(\mathcal{A}|i)$ contains the states that, taken as a starting solution, would have the search provide a certified global optimum. Hence, if we use a randomly chosen state as a starting solution,



(b) Instance 2

(c) Instance 3

Fig. 1. Computational time: comparison between SD+QP and MINLP approaches over ORLib Instances.

the ratio $|I^*|/|S|$ would measure the probability to find an optimal solution. As a generalization, we are defining a probabilistic basin of attraction as follows:

**Definition** Given a stochastic algorithm $\mathcal{A}$, the basin of attraction $\mathcal{B}(\mathcal{A}|s;p^*)$ of a point $s$, is defined as the set of states that, taken as initial states, give origin to trajectories that include point $s$ *with probability* $p \geq p^*$. Accordingly, the union of the BOA of global optima is defined as $I^*(p) = \bigcup_{i \in S^*} \mathcal{B}(\mathcal{A}|i;p)$. It is clear that that $\mathcal{B}(\mathcal{A}|s)$ is a special case for $\mathcal{B}(\mathcal{A}|s;p^*)$, hence in what follows we are using $\mathcal{B}(s;p^*)$ instead of $\mathcal{B}(\mathcal{A}|s;p^*)$, without loss of generalization. When $p^* = 1$ we want to find solutions belonging to trajectories that ends in $s$. Notice that $\mathcal{B}(s;p_1) \subseteq \mathcal{B}(s;p_2)$ when $p_1 > p_2$.

The effectiveness of a meta-heuristic $\mathcal{A}$ is dramatically influenced by the topology and structure of the search landscape, and since the aim is to reach an optimal solution, the need of an analysis of BOA features arises. Please notice that our definition of basins of attraction enables both a complete/analytical study —when probabilities can be deducted from the search strategy features— and a statistical/empirical analysis (e.g., by sampling).

## VI. Search space analysis for Portfolio Selection Problem

When solving an optimisation problem, a sound modelling and development phase should be based on the separation between the model and the algorithm: this stems from constraint programming, and several tools foster this approach (i.e., Comet[17]). In this way, it is possible to draw information about the structure of the optimisation problem, and this knowledge can be used, for instance, for the choice of the algorithm to be used. Up to the author's knowledge, literature about portfolio selection by meta-heuristics has hardly dealt with this aspect, though some attempts have been made to study the problem structure. For instance, Maringer and Winker [23] draw some conclusion about the objective function landscape by using a memetic algorithm which embeds, in turn, Simulated Annealing (SA)[19] and Threshold Acceptance (TA)[11]. They compare the use of SA and TA inside the memetic algorithm dealing with different objective functions: Value-at-Risk(*Var*) and Expected Shortfall (*ES*)[8]. Their results indicates that TA is suitable when using *VaR*, while SA performs best when using *ES*. An analysis of the search space is made to understand this phenomenon.

Other works compare different algorithms on the same instance to understand which algorithm perform best, and in what portion of the frontier. Amongst them, Crama and Schyns[6] introduce three different Simulated Annealing strategies, showing that there is no clear dominance among them. Armañanzas and Lozano [1] introduces Ant Colony Optimisation (ACO)[10], refining solutions with a greedy search, comparing results with Simulated Annealing and Iterative Improvement, and showing that ACO and SA performances greatly depends on the expected return (see Sec. II). A common way of tackling this analysis is to run the different algorithms, and then to pool the obtained solutions. After this phase, the dominated solutions are deleted and it is possible to understand which algorithm performs best w.r.t. a given part of the frontier [5], [13].

The main shortcoming of these approaches is that they identify which algorithm performs well in a given portion of the frontier, without explaining the motivation beneath this behaviour. Hence, an additional effort has to be made to understand the model and how it can affect the algorithm performance. In this section, we are aimed in comparing different formulations for the PSP and in understanding how the structure of the problem affects the algorithm's performances through Search Space Analysis.

When using a meta-heuristic, search space analysis represents an effective tool to assess the algorithm performances and the instance hardness. In what follows we are discussing results obtained over real instances and over hard-handmade instances in order to outline the connections between search space analysis and algorithm performances.

*Analysis for Real Instances:* We define five equally distributed $r_e$ values, referred to as $R_i$ $(i = 1 \ldots 5)$ and we analyse the search space corresponding to each $r_i$ over the five `ORlib` instances in order to assess the local minima distribution, that is an indicator of the search space ruggedness. This concept is important since it has been shown that there exists a negative correlation between ruggedness and meta-heuristic performances[18]. We have implemented and run a deterministic version of SD (referred to as $SD_{det}$) to estimate the number of minima of an instance of the problem discussed in Sec. IV, which combines continuous variables $x$ with integer variables $z$. As for the constraints, we have set both a minimum ($k_{min}$) or a maximum ($k_{max}$) bound on cardinality in order to understand the differences arising when using a maximum or strict cardinality constraint. As for determining the initial states, we have resorted either to complete enumeration (if the instance at hand is small) or to uniform sampling.

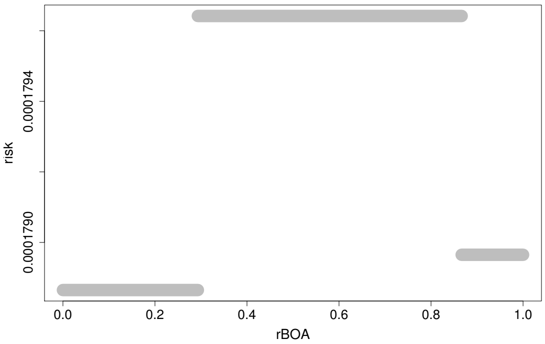


Fig. 2. Instance 4: BOA analysis. $k_{min} = 1$, $k_{max} = 10$, $R = 0.00375$.

Results are shown in table V, where we report the number of the different local minima found by 30 runs of $SD_{det}$. Dashed entries mean that no feasible solution exists.

Results indicate that instances at hand show a small number of local minima and only one global minimum. This clearly indicates a situation in which the search landscape is rather smooth, and explains why different strategies such TS and FD/SD lead to similar optimization results: since local optimum are few and far between, there is no need of using complex strategies or escape mechanisms, since the probability of meeting a trajectory leading to one of the optima are quite high. We recall that those values have been found by using a deterministic version of $SD$, and their inverse represents an upper bound on the probability to reach the certified optimum when using the stochastic SD and TS defined in section IV-A.

We conclude that when using our formulation, global minima have a quite large BOA. This can be seen in Fig. 2, in which segments length corresponds to *rBOA* (i.e., ratio between size of *BOA(s)* and search space size) and their y-value corresponds to the minimum found: global minima *rBOA* ranges from 30% to 60%.

In the next paragraph we will show that the same problem, modeled in a different way, leads to different basin of attractions.

*Monolithic Search Basin of Attraction:* In the previous paragraph we have shown that, when using our problem formulation, the BOAs of local optima are quite big, making the search landscape smooth and the problem easy to be tackled by our hybrid solver. BOAs depend on the search strategy used and on the problem

TABLE V
INSTANCE 4, NUMBER OF MINIMA FOUND.

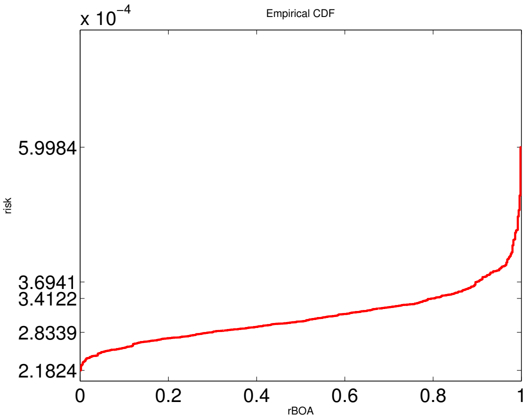
| $k_{min}, k_{max}$ | $R_1 = 0.00912$ | $R_2 = 0.00738$ | $R_3 = 0.00556$ | $R_4 = 0.00375$ | $R_5 = 0.00193$ |
|---|---|---|---|---|---|
| 1,3 | 1 | 1 | 1 | 1 | 1 |
| 1,6 | 1 | 1 | 1 | 5 | 1 |
| 1,10 | 1 | 1 | 1 | 1 | 3 |
| 3,3 | 1 | 1 | 3 | 5 | 3 |
| 6,6 | – | 1 | 1 | 2 | 1 |
| 10,10 | – | 1 | 1 | 3 | 2 |

formulation, and this can be shown by running a different strategy, i.e., a monolithic one, on the same problem instances. We have used a SD based on a variant of Threshold Accepting [12], in which only a variable class is considered, i.e., $w$ variables corresponding to actual asset weights. The desired outcome of this problem is the same as the previously introduced one, but they are represented in a different way. In the following we explain the main features of this meta-heuristic approach:

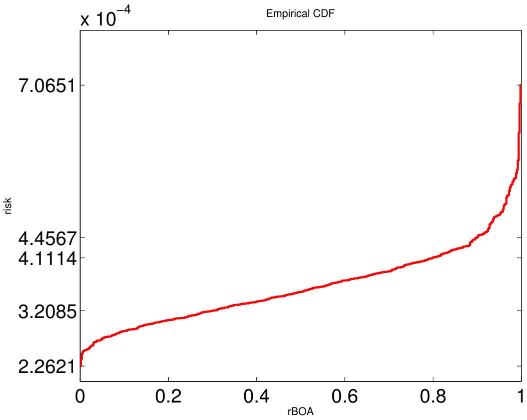
- **Search Space** The *master-slave* decomposition is not used anymore, and a state is represented by a sequence $W = w_1 \dots w_n$ such that $w_b$ corresponds to the relative amount invested in asset $b$. Furthermore, the portfolio has to be feasible w.r.t. cardinality, budget, floor and ceiling constraints.
- **Neighborhood relations** A given amount (*step*) is transferred from asset $a$ to another $b$, no matter if $b$ is already in the portfolio or not. If this leads one asset value to be smaller than $\epsilon_i$, its value is set to $\epsilon_i$. If the move consists in decreasing the value of an asset being set to $\epsilon_i$, its value is set to $0$.
- **Initial solution** The initial solution has to be feasible w.r.t. cardinality, budget, floor and ceiling constraints and is always created from scratch.
- **Cost Function** As for the cost function we are using a penalty approach, hence it is given by adding the degree of constraints violations to the portfolio risk.
- **Local Search Strategies** SD that explores the space of $w$ variables.

Results about BOAs analysis for this approach are shown in figure 3. Even from visual inspection only, it turns out that the number of local minima is dramatically higher than the one corresponding to the *master-slave* approach; furthermore basin of attraction are tiny, and the certified optimum has not been found.

*Analysis for artificial instances:* In the previous paragraph we have shown that, for the PSP we are solving, instances at hand are easy to solve, since our *master-slave* decomposition leads to search spaces with a small number of local optima with huge BOAs. Hence, there is no need for complex approaches and escape mechanisms, and this explains why simple meta-heuristics performances are comparable with more sophisticated one such TS. Furthermore, preliminary analysis have



(a) Instance 4: $k_{min} = 1$, $k_{max} = 10$, $R = 0.00375$



(b) Instance 4: $k_{min} = 1$, $k_{max} = 6$, $R = 0.00193$

Fig. 3. Two ORlib instances: Monolithic BOA analysis with different constraints.

$$\sigma = \left(\begin{array}{ccccccccccc}
1 & -1 & 0 & 0 & 0 & & & \cdots & & & 0\\
-1 & 1 & 0 & 0 & 0 & & & \cdots & & & 0\\
0 & 0 & 1 & -0.9 & 0 & 0 & 0 & \cdots & & & \vdots\\
0 & 0 & -0.9 & 1 & 0 & 0 & 0 & \cdots & & & \vdots\\
\vdots & & 0 & 0 & 1 & -0.9 & 0 & \cdots & & & \vdots\\
\vdots & & 0 & 0 & -0.9 & 1 & 0 & \cdots & & & \vdots\\
\vdots & & \ddots & & 0 & 0 & \ddots & & & & \vdots\\
\vdots & & & \ddots & \vdots & \vdots & & \ddots & & & \vdots\\
\vdots & & & & \cdots & & & & 1 & 0 & 0\\
\vdots & & & & \cdots & & & & 0 & 1 & -0.9\\
0 & & & & \cdots & & & & 0 & -0.9 & 1
\end{array}\right) \qquad (6)$$

suggested us that this is a common feature in financial market related instances: this could be considered as a good point for practitioners, but makes impossible to test the robustness of our approach, in which we have developed TS+QP in order to tackle more difficult instances. Hence, we have designed an artificial hand-made instance featuring a huge number of minima with tiny BOAs, containing an even number $n$ of assets $i$, in which $r_i = 1 \forall i$ and whose covariance matrix is depicted here above.

It is easy to see that for every $r_e$ the best portfolio contains the first two assets only, but also that portfolios consisting of assets $i$ (odd) and $i + 1$ only are local optima, since all their neighbors feature higher risk.

It can be shown show that it is necessary to visit a portfolio $s$ having $z_1 = 1$ or $z_2 = 1$ to reach the global optimum $s^*$. Furthermore, portfolios containing an odd asset $i$ ($i > 1$) whose $z_i = 1$ and $z_{i+1} = 1$ will never entry in a trajectory in which this couple would be removed. Hence, $\mathcal{B}(s^*)$ contains all portfolios featuring $z_1 = 1$ or $z_2 = 1$, and in which there is no $i$ odd and $> 1$ such that $z_i = 1$ and $z_{i+1} = 1$. In this case, *rBOA*$(s^*)$ is inversely proportional to $n$.

By running our master-slave approach over this instance ($\epsilon_i = 0.01$ and $\delta_i = 1$ for $i = 1 \dots n$) we have remarked that TS+QP easily find a solution comparable to that provided by CPLEX, while SD and FD performances are greatly affected by the starting solution (and anyhow much poorer than TS+QP).

It has to be noticed that such an instance could be hardly found over real markets, even its presence is not forbidden by structural properties, but when tackling it the need of larger neighborhoods arises. Anyhow, no matter the neighborhood size, it is always possible to devise artificial instances whose minima are composed by subsets that have to be moved jointly.

From the Search Space Analysis conducted in this section, we may conclude that different formulations (hybrid vs continuous only) lead to different Basin of Attraction analysis on the instances at hand. This turns into different algorithm behaviours. The formulation that leads to a smooth search landscape (hybrid) can be tackled by algorithms with weak diversification capabilities (i.e., SD in the proposed hybrid formulation), whilst these algorithms are to be replaced by more sophisticated ones when the search landscape becomes rugged (see the behaviour of SD in the monolithic version). The artificial instance places itself in the middle of these phenomena, as it provides room for the use of more complex strategies (i.e., TS) in the hybrid case, due to the neighbor moves used which make the search to get stuck in the first local optimum found, but when embedded in the continuous only formulation doesn't provide different performances from the real instances.

## VII. Conclusion

In this work we have used a meta-heuristic approach to study the impact of different formulations on the Portfolio Selection algorithm's behaviour, and we have devised a methodology to understand the root of the different behaviours (search space analysis through BOA analysis). To this aim we have compared an approach based on a master-slave decomposition with a monolithic approach. Results have shown that the search space defined by the monolithic approach is quite rugged and need an algorithm featuring an escape mechanism to be solved efficiently, whilst the hybrid approach leads to a smoother search landscape to be explored efficiently also by simpler algorithms such SD.

## References

[1] R. Armañanzas and J.A. Lozano. A multiobjective approach to the portfolio optimization problem. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, volume 2, pages 1388–1395, 2005.

[2] A.Roli. A note on a model of local search. Technical Report TR/IRIDIA/2004/23.01, IRIDIA, ULB, Belgium, 2004.

[3] M. Birattari, T. Stützle, L. Paquete, and K. Varrentrapp. A racing algorithm for configuring metaheuristics. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2002)*, pages 11–18. Morgan Kaufmann Publishers, 2002.

[4] C. Blum and A. Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys*, 35(3):268–308, 2003.

[5] T.J. Chang, N. Meade, J.E. Beasley, and Y.M. Sharaiha. Heuristics for cardinality constrained portfolio optimisation. *Computers & Operations Research*, 27(13):1271–1302, 2000.

[6] Y. Crama and M. Schyns. Simulated annealing for complex portfolio selection problems. *European Journal of Operational Research*, 150:546–571, 2003.

[7] L. Di Gaspero, G. di Tollo, A. Roli, and A. Schaerf. Hybrid local search for constrained financial portfolio selection problems. In *Proceedings of Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, pages 44–58, 2007.

[8] G. di Tollo and A. Roli. Metaheuristics for the portfolio selection problem. *International Journal of Operations Research*, 5(1):443–458, 2008.

[9] G. di Tollo, T. Stützle, and M. Birattari. A metaheuristic multi-criteria optimisation approach to portfolio selection. *Journal of Applied Operational Research*, 6(4):222–242, 2014.

[10] M. Dorigo, L. M. Gambardella, M. Middendorf, and T. Stützle, editors. *Special Section on "Ant Colony Optimization"*. *IEEE Transactions on Evolutionary Computation*, 6(4), 317–365, 2002.

[11] G. Dueck and T. Scheuer. Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. *Journal of Computational Physics*, 90(1):161–175, 1990.

[12] G. Dueck and P. Winker. New concepts and algorithms for portfolio choice. *Applied Stochastic Models and Data Analysis*, 8:159–178, 1992.

[13] A. Fernandez and S. Gomez. Portfolio selection using neural networks. *Computers & Operations Research*, 34:1177–1191, 2007.

[14] R. Fourer, D.M. Gay, and B.W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming*. Duxbury Press/Brooks/Cole Publishing Company, 2002.

[15] L. Di Gaspero, G. di Tollo, A. Roli, and A. Schaerf. Hybrid metaheuristics for constrained portfolio selection problems. *Quantitative Finance*, 11(10):1473–1487, 2011.

[16] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27:1–33, 1983.

[17] P. Van Hentenryck and L. Michel. *Constraint-Based Local Search*. The MIT Press, 2005.

[18] H. Hoos and T. Stützle. *Stochastic Local Search Foundations and Applications*. Morgan Kaufmann Publishers, 2005.

[19] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[20] R. Mansini, W. Ogryczak, and M.G. Speranza. LP solvable models for portfolio optimization: a classification and computational comparison. *IMA Journal of Management Mathematics*, 14(3):187–220, 2003.

[21] R. Mansini and M.G. Speranza. Heuristic algorithms for the portfolio selection problem with minimum transaction lots. *European Journal of Operational Research*, 114(2):219–233, 1999.

[22] D. Maringer. *Portfolio Management with heuristic optimization*. Springer, 2005.

[23] D. Maringer and P. Winker. Portfolio optimization under different risk constraints with modified memetic algorithms. Technical Report 2003-005E, University of Erfurt, Faculty of Economics, Law and Social Sciences, 2003.

[24] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[25] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[26] R. Moral-Escudero, R. Ruiz-Torrubiano, and A. Suárez. Selection of optimal investment with cardinality constraints. In *Proceedings of the IEEE World Congress on Evolutionary Computation*, pages 2382–2388, 2006.

[27] S. Prestwich and A. Roli. Symmetry breaking and local search spaces. In *Proceedings of the 2nd International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, pages 273–287, 2005.

[28] E. Rolland. A tabu search method for constrained real number search:applications to portfolio selection. Technical report, Department of Accounting and Management Information Systems, Ohio State University, Columbus. U.S.A., 1997.

[29] A. Schaerf. Local search techniques for constrained portfolio selection problems. *Computational Economics*, 20(3):177–190, 2002.

[30] M.G. Speranza. A heuristic algorithm for a portfolio optimization model applied to the Milan stock market. *Computers & Operations Research*, 23(5):433–441, 1996.

[31] F. Streichert, H. Ulmer, and A. Zell. Comparing discrete and continuous genotypes on the constrained portfolio selection problem. In *Proceedings of Genetic and Evolutionary Computation Conference*, volume 3103 of *LNCS*, pages 1239–1250, 2004.

[32] M. Yokoo. Why adding more constraints makes a problem easier for hill-climbing algorithms: Analyzing landscapes of CSPs. In *Proceedings of the Third Conference on Principles and Practice of Constraint Programming*, pages 356–370, 1997.

DOI: 10.15439/2015F366
ACSIS, Vol. 5

# Fast Solvers for Nonsmooth Optimization Problems in Phase Separation

Pawan Kumar
Department of Mathematics and Computer Science
Freie Universität Berlin
Arnimallee 6, 14195 Berlin

***Abstract*—The phase separation processes are typically modeled by well known Cahn-Hilliard equation with obstacle potential. Solving these equations correspond to a nonsmooth and nonlinear optimization problem. Recently a globally convergent Newton Schur method was proposed for the non-linear Schur complement corresponding to this $2 \times 2$ non-linear system. The proposed method is similar to an inexact active set method in the sense that the active sets are first identified by solving a quadratic obstacle problem corresponding to the $(1, 1)$ block of the $2 \times 2$ system, and later solving a reduced linear system by annihilating the rows and columns corresponding to identified active sets. For solving the quadratic obstacle problem, various optimal multigrid like methods have been proposed. However solving the reduced system remains a major bottleneck. In this paper, we explore an effective preconditioner for the reduced linear system that allows solving large scale optimization problem corresponding to Cahn-Hilliard and to possibly similar models.**

## I. Introduction

THE Cahn-Hilliard equation was first proposed in 1958 by Cahn and Hilliard [1] to study the phase separation process in a binary alloy. Here the term phase stands for the concentration of different components in the alloy. It has been empirically observed that the concentration changes from a possibly mixed state to a distinct spatially separated two phase state when the alloy under preparation is subjected to a rapid cooling at a certain critical temperature. This rapid reduction in the temperature the so-called *deep quench limit* has been found to be modeled efficiently by obstacle potential proposed by Oono and Puri [2] in 1987; also see Blowey and Elliot [3, p. 237, (1.14)]. The phase separation has been noted to be highly non-linear (point nonlinearity to be precise), and the obstacle potential emulates the nonlinearity and non-smoothness that is empirically observed. Dealing with the non-smoothness and designing robust iterative procedure has been the subject of much active research in last decades. Assuming semi-implicit time discretizations [4] to alleviate the time step restrictions, most of the proposed methods essentially differ in the way the nonlinearity and non-smoothness are handled. Two of the main approaches for such problems are: regularization around the non-smooth region [5] or an active set approach [6] i.e., identify the active sets and solve a reduced problem which is linear, [6] unlike [5] also ensures global convergence of the Newton method by using proper damping parameter. The nonlinear problem corresponding to Cahn-Hilliard problem with obstacle potential could be written as a non-linear system in block $2 \times 2$ matrix form as follows:

$$\begin{pmatrix} F & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} u^* \\ w^* \end{pmatrix} \ni \begin{pmatrix} f \\ g \end{pmatrix}, \quad u^*, w^* \in \mathbb{R}^n \tag{1}$$

where $u^*, w^*$ are unknowns, $F = A + \partial I_{\mathcal{K}}$, where $I_{\mathcal{K}}$ denotes the indicator functional of the admissible set $\mathcal{K}$. The matrices $A, C$ are essentially Laplacian with $A$ augmented by a non-local term (a rank one term) reflecting mass conservation. By nonlinear Gaussian elimination of the $u^*$ variables, the system above could be reduced to a nonlinear Schur complement system in $w^*$ variables [6], where the nonlinear Schur complement is given by $-(C + BF^{-1}B^T)$. In [6], a globally convergent Newton method is proposed for this nonlinear Schur complement system which is later interpreted as a preconditioned Uzawa iteration. Note that $F(x)$ is a set valued mapping due to the presence of set-valued operator $\partial I_{\mathcal{K}}$; to solve the inclusion $F(x) \ni y$, or equivalently, $x \in F^{-1}y$ corresponding to the quadratic obstacle problem, many methods have been proposed: projected block Gauss-Seidel [7], monotone multigrid method [8], [9], [10], truncated monotone multigrid [11], and recently introduced truncated Newton multigrid [11]. See the excellent review article [11] that compares all these methods. Solving the quadratic obstacle problem corresponds to identifying the active sets. By annihilating the corresponding rows and columns that belong to the identified active sets, we obtain a reduced linear system as follows:

$$\begin{pmatrix} \hat{A} & \hat{B}^T \\ \hat{B} & -C \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{w} \end{pmatrix} = \begin{pmatrix} \hat{f} \\ \hat{g} \end{pmatrix}, \quad \hat{u}, \hat{w} \in R^n \tag{2}$$

The overall nonlinear iteration is performed in the sense of inexact Uzawa, and the preconditioners are updated with next available active sets.

In this paper our goal is to design effective preconditioner for (2). In particular we consider the Schur complement preconditioner proposed in [5] and adapt it to our linear system. Our linear system differs from [5]; we have to deal with nontrivial kernels at (1,1) and (2,2) block. We study the effectiveness of the preconditioner for various active set configurations that allow solving large scale nonsmooth optimization problem corresponding to similar model problems when similar nonsmooth Newton Schur method is used.

The rest of this paper is organized as follows. In Section III, we describe the Cahn-Hilliard model with obstacle potential,

Funded by EC math and Matheon

we discuss the time and space discretizations and variational formulations. In Section IV, we discuss briefly the solver for Cahn-Hilliard with obstacle problem. In particular, we briefly discuss the recent Nonsmooth Newton Schur method, and the truncated monotone multigrid for the obstacle problem; we describe how the linear system appears. The preconditioners for the reduced linear systems are discussed in Section IV-E. Finally in Section V, we show numerical experiments with the proposed preconditioner.

## II. Notation

Let SPD and SPSD denote symmetric positive definite and symmetric positive semi-definite respectively. Let $|x|$ denote the absolute value of $x$, whereas, for a set $\mathcal{K}$, $|\mathcal{K}|$ denotes the number of elements in $\mathcal{K}$. Let $Id \in \mathbb{R}^{n\times n}$ denote the identity matrix. For a vector $u$ let $u(i)$ denote the $i^{th}$ entry of vector $u$. Similarly for a matrix we use the notation $K(i,j)$ to denote the $(i,j)^{th}$ entry of $K$. For any matrix $K$, $K^+$ shall denote a pseudoinverse of $K$.

## III. Cahn-Hilliard Problem with Obstacle Potential

### A. The Model

The Cahn-Hilliard equation in PDE form with inequality constraints reads:

$$\partial_t u = \Delta w, \tag{3}$$

$$w = -\epsilon\Delta u + \psi_0^{'}(u) + \mu, \tag{4}$$

$$\mu \in \partial\beta_{[-1,1]}(u), \tag{5}$$

$$|u| \le 1, \tag{6}$$

$$\frac{\partial u}{\partial n} = \frac{\partial w}{\partial n} = 0 \text{ on } \partial\Omega, \tag{7}$$

where $\partial\beta_{[-1,1]}(u)$ is the subdifferential of $\beta_{[-1,1]}(u) := \int_\Omega I_{[-1,1]}(u)$. The obstacle potential $\psi$ is given as follows:

$$\psi(u) = \psi_0(u) + I_{[-1,1]}(u), \text{ where } \psi_0(u) = \frac{1}{2}(1-u^2). \tag{8}$$

Here the indicator function $I_{[-1,1]}(u)$ is defined as follows:

$$I_{[-1,1]} = \begin{cases} 0, \text{ if } u(i) \in [-1,1], \\ \infty, \text{ otherwise.} \end{cases} \tag{9}$$

The subscript $[-1,1]$ correspond to the fact that $u$ is allowed to take values only between $-1$ and $+1$, which we sometimes refer to as upper and lower obstacles respectively.

In (3)-(7) the unknowns $u$ and $w$ are called order parameter and chemical potential respectively. For a given $\epsilon > 0$, final time $T > 0$ and initial condition $u_0 \in \mathcal{K}$ where

$$\mathcal{K} = \{v \in H^1(\Omega) \,:\, |v| \le 1\}, \tag{10}$$

the equivalent initial value problem for Cahn-Hilliard equation with obstacle potential interpreted as variational inequality reads:

$$\left\langle \frac{du}{dt}, v \right\rangle + (\nabla w, \nabla v) = 0, \; \forall v \in H^1(\Omega), \tag{11}$$

$$\epsilon(\nabla u, \nabla(v-u)) - (u, v-u) \ge (w, v-u), \; \forall v \in \mathcal{K}, \tag{12}$$

where we use the notation $\langle\cdot,\cdot\rangle$ to denote the duality pairing of $H^1(\Omega)$ and $H^1(\Omega)^{'}$. Note that we used the fact that $\psi_0^{'}(u) = -u$ in the second term on the left of inequality (12) above. The existence and uniqueness of the solution of (11), (12) above has been established in Blowey and Elliot [3]. We next consider an appropriate discretization in time and space for the model.

### B. Time and Space Discretizations

We consider a fixed non-adaptive grid in time $(0,T)$ and in space $\Omega = (0,1)\times(0,1)$. The time step $\tau = T/N$ is kept uniform, $N$ being the number of time steps. We consider the semi-implicit Euler discretization in time and finite element discretization as in Barrett et. al. [7] with triangulation $\mathcal{T}_h$ with the following spaces:

$$\mathcal{S}_h = \{v \in C(\overline{\Omega}) \,:\, v|_T \text{ is linear} \forall T \in T_h\}, \tag{13}$$

$$\mathcal{P}_h = \{v \in L^2(\Omega) \,:\, v_T \text{ is constant} \forall T \in \mathcal{T} \in \mathcal{T}_h\}, \tag{14}$$

$$\mathcal{K}_h = \{v \in \mathcal{P}_h \,:\, |v_T| \le 1 \forall T \in \mathcal{T}_h\} = \mathcal{K}\cap\mathcal{S}_h \subset \mathcal{K}, \tag{15}$$

which leads to the following discrete Cahn-Hilliard problem with obstacle potential:
Find $u_h^k \in \mathcal{K}_h, w_h^k \in \mathcal{S}_h$ such that

$$\langle u_h^k, v_h\rangle + \tau(\nabla w_h^k, \nabla v_h) = \langle u_h^{k-1}, v_h\rangle, \; \forall v_h \in \mathcal{S}_h, \tag{16}$$

$$\epsilon(\nabla u_h^k, \nabla(v_h - u_h^k)) - \langle w_h^k, v_h - u_h^k\rangle \ge \langle u_h^{k-1}, v_h - u_h^k\rangle, \tag{17}$$

$$\forall v_h \in \mathcal{K}_h.$$

holds for each $k = 1,\ldots,N$. The initial solution $u_h^0 \in \mathcal{K}_h$ is taken to be the discrete $L^2$ projection $\langle u_h^0, v_h\rangle = (u_0, v_h), \forall v_h \in \mathcal{S}_h$. Existence and uniqueness of the discrete Cahn-Hilliard equations has been established in [4]. The discrete Cahn-Hilliard equation is equivalent to the set valued saddle point block $2\times 2$ nonlinear system (1) with $F = A + \partial I_{\mathcal{K}_h}$ and

$$A = \epsilon(\langle\lambda_p, 1\rangle\langle\lambda_p, 1\rangle + (\nabla\lambda_p, \nabla\lambda_q))_{p,q\in\mathcal{N}_h}, \tag{18}$$

$$B = -(\langle\lambda_p, \lambda_q\rangle)_{p,q\in\mathcal{N}_h}, \; C = \tau((\nabla\lambda_p, \nabla\lambda_q))_{p,q\in\mathcal{N}_h}. \tag{19}$$

where $\mathcal{N}_h$ stands for the set of vertices in $\mathcal{T}_h$, and $\lambda_p, p \in \mathcal{N}_h$ denote the standard nodal basis. We write the above in more compact notation as follows

$$A = \epsilon(K + mm^T), \quad B = -M, \quad C = \tau K, \tag{20}$$

where $K$ and $M$ are stiffness and mass matrices respectively.

## IV. Iterative solver for Cahn-Hilliard with Obstacle Potential

In [6], a nonsmooth Newton Schur method is proposed which is also interpreted as a preconditioned Uzawa iteration. For a given time step $k$, the Uzawa iteration reads:

$$u^{i,k} = F^{-1}(f^k - B^T w^{i,k}), \tag{21}$$

$$w^{i+1,k} = w^{i,k} + \rho^{i,k}\hat{S}_{i,k}^{-1}(Bu^{i,k} - Cw^{i,k} - g^k) \tag{22}$$

for the saddle point problem (1). Here $i$ denotes the $i^{th}$ Uzawa step and $k$ denotes the $k^{th}$ time step. Here $f^k$ and $g^k$ are defined in (16) and (17). The time loop starts with an initial value for $w^{0,0}$ which is taken arbitrary, and with the initial value $u^{0,0}$. The Uzawa iteration requires three main computations that we describe below.

### A. Computing $u^{i,k}$

The first step (21) corresponds to a quadratic obstacle problem:

$$u^{i,k} = \arg \min_{v \in \mathcal{K}} \left( \frac{1}{2} \langle Av, v \rangle - \langle f^k - B^T w^{i,k}, v \rangle \right). \tag{23}$$

As mentioned in the introduction, this problem has been extensively studied during last decades [7], [8], [9], [11].

### B. Computing $\hat{S}_{i,k}^{-1}(Bu^{i,k} - Cw^{i,k} - g^k)$

The descent direction $d^{i,k+1} = \hat{S}_{i,k}^{-1}(Bu^{i,k} - Cw^{i,k} - g)$ in (22) is obtained as a solution of the following reduced linear block $2 \times 2$ system:

$$\begin{pmatrix} \hat{A} & \hat{B}^T \\ \hat{B} & -C \end{pmatrix} \begin{pmatrix} \tilde{u}^{i,k} \\ d^{i,k} \end{pmatrix} = \begin{pmatrix} 0 \\ g + Cw^{i,k} - Bu^{i,k} \end{pmatrix}, \tag{24}$$

where

$$\hat{A} = TAT + \hat{T}, \quad \hat{B} = TB. \tag{25}$$

Here $T$ and $\hat{T}$ are defined as follows:

$$T = \text{diag} \begin{pmatrix} 0, \text{if } u^{i,k}(j) \in \{-1, 1\} \\ 1, \text{otherwise} \end{pmatrix}, \ j = 1, \ldots, |\mathcal{N}_h|, \tag{26}$$

$$\hat{T} = Id - T, \quad Id \in \mathbb{R}^{|\mathcal{N}_h| \times |\mathcal{N}_h|}, \tag{27}$$

where $u^{i,k}(j)$ is the $j^{th}$ component of $u^{i,k}$. In other words, $\hat{A}$ is the matrix obtained from $A$ by replacing the $i^{th}$ row and the $i^{th}$ column by the unit vector $e_i$ corresponding to the active sets identified by diagonal entries of $T$. Similarly, $\hat{B}$ is the matrix obtained from $B$ by annihilating columns, and $\hat{B}^T$ is the matrix obtained from $B^T$ by annihilating rows.

### C. Computing Step Length $\rho^{i,k}$

The step length $\rho^{i,k}$ is computed using a bisection method. We refer the reader to [12, p. 88].

### D. Algebraic Monotone Multigrid for Obstacle Problem

To solve the quadratic obstacle problem (21), we use the truncated monotone multigrid method proposed in [8]. However, here we use algebraic coarsening [13] that we describe briefly.

*1) Aggregation Based Coarsening:* We first discuss the coarsening for two-grid, the multilevel interpolations are applied recursively. In classical two-grid, a set of coarse grid unknowns is selected and the matrix entries are used to build interpolation rules that define the prolongation matrix P, and the coarse grid matrix $A_c$ is computed from the following Galerkin formula

$$A_c = P^T A P. \tag{28}$$

In contrast to the classical two-grid approach, in aggregation based multigrid, first a set of aggregates $G_i$ is defined. Let $|\mathcal{N}_{h,c}|$ be the total number of such aggregates, then the interpolation matrix $P$ is defined as follows

$$P_{ij} = \begin{cases} 1, & \text{if } i \in G_j, \\ 0, & \text{otherwise}, \end{cases}$$

Here, $1 \le i \le |\mathcal{N}_h|$, $1 \le j \le |\mathcal{N}_{h,c}|$. Further, we assume that the aggregates $G_i$ are such that

$$G_i \bigcap G_j = \phi, \text{ for } i \neq j \text{ and } \bigcup_i G_i = \{i \in \mathbb{N} : 1 \le i \le |\mathcal{N}_h|\}. \tag{29}$$

The matrix $P$ defined above is a $|\mathcal{N}_h| \times |\mathcal{N}_{c,h}|$ matrix, but since it has only one non-zero entry (which are "one") per row, the matrix is compactly represented by a single array of length $\mathcal{N}_{h,c}$ storing the location of the non-zero entry on each row. The coarse grid matrix $A_c$ may be computed as follows

$$(A_c)(i, j) = \sum_{k \in G_i} \sum_{l \in G_j} A(k, l),$$

where $1 \le i, \ j \le |\mathcal{N}_{h,c}|$, and $A(k, l)$ is the $(k, l)^{th}$ entry of $A$.

Numerous aggregation schemes have been proposed in the literature, but in this paper we consider the standard aggregation based on strength of connection [14, Appendix A, p. 413] where one first defines a set of nodes $\mathcal{S}_i$ to which $i$ is strongly negatively coupled, using the Strong/Weak coupling threshold $\beta$:

$$\mathcal{S}_i = \{ \, j \neq i \mid A(i, j) < -\beta \ \text{max}|A(i, k)| \, \}.$$

Then an unmarked node $i$ is chosen such that priority is given to a node with minimal $M_i$, here $M_i$ being the number of unmarked nodes that are strongly negatively coupled to $i$. For a complete algorithm of aggregation, the reader is referred to Notay [13], [15].

### E. Preconditioner for Reduced Linear System

In Bosch et. al. [5], a preconditioner is proposed in the framework of semi-smooth Newton method combined with Moreau-Yosida regularization for the same problem. However, the preconditioner was constructed for a linear system which is different from the one we consider in (24). For convenience of notation, we rewrite the system matrix in (24) as follows

$$\mathcal{A} x = b, \tag{30}$$

where scripted $\mathcal{A}$ above is

$$\mathcal{A} = \begin{pmatrix} \hat{A} & \hat{B}^T \\ \hat{B} & -C \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \tag{31}$$

where

$$x_1 = \tilde{u}^{i,k}, \quad x_2 = d^{i,k}, \quad b_1 = 0, \quad b_2 = g + Cw^{i,k} - Bu^{i,k}. \tag{32}$$

The preconditioner proposed in [5] has the following block lower triangular form

$$\mathcal{B} = \begin{pmatrix} \hat{A} & 0 \\ \hat{B} & -S \end{pmatrix}, \tag{33}$$

where $S = C + \hat{B}\hat{A}^{-1}\hat{B}^T$ is the Schur complement. Note that such preconditioners are also called inexact or preconditioned Uzawa preconditioners for the *linear* saddle point problems. By block $2 \times 2$ inversion formula we have

$$\mathcal{B}^{-1} = \begin{pmatrix} \hat{A} & 0 \\ \hat{B} & -S \end{pmatrix}^{-1} = \begin{pmatrix} \hat{A}^{-1} & 0 \\ S^{-1}\hat{B}^T\hat{A}^{-1} & -S^{-1} \end{pmatrix}. \tag{34}$$

Let $\hat{S}$ be an approximation of Schur complement $S$ in $\mathcal{B}$, then the new preconditioner $\hat{\mathcal{B}}$ and the corresponding preconditioned operator $\hat{\mathcal{B}^{-1}}\mathcal{A}$ are given as follows

$$\hat{\mathcal{B}} = \begin{pmatrix} \hat{A} & 0 \\ \hat{B} & -\hat{S} \end{pmatrix}, \quad \hat{\mathcal{B}}^{-1}\mathcal{A} = \begin{pmatrix} I & \hat{A}^{-1}\hat{B}^T \\ 0 & \hat{S}^{-1}S \end{pmatrix}. \tag{35}$$

Using (35) above, we can note the following trivial result.

*Theorem 4.1:* Let $\mathcal{B}$ defined in (35) be a preconditioner for $\mathcal{A}$ defined in (31), then there are $|\mathcal{N}_h|$ eigenvalues of $\mathcal{B}^{-1}\mathcal{A}$ equal to one, and the rest are the eigenvalues of the preconditioned Schur complement $\hat{S}^{-1}S$.

*Remark 4.1:* When using GMRES [16], right preconditioning is preferred. Similar result as for the left preconditioner above Theorem 4.1 holds.

The preconditioned system $\mathcal{B}^{-1}\mathcal{A}x = \mathcal{B}^{-1}b$ is given as follows

$$\begin{pmatrix} I & \hat{A}^{-1}\hat{B}^T \\ 0 & \hat{S}^{-1}S \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \hat{A}^{-1} & 0 \\ S^{-1}\hat{B}^T\hat{A}^{-1} & -S^{-1} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \tag{36}$$

from which we obtain the following set of equations

$$x_1 + \hat{A}^{-1}\hat{B}^T x_2 = \hat{A}^{-1}b_1, \tag{37}$$

$$\hat{S}^{-1}Sx_2 = S^{-1}(\hat{B}^T\hat{A}^{-1}b_1 - b_2). \tag{38}$$

```
Algorithm 4.1: Objective: Solve $\mathcal{B}^{-1}\mathcal{A}x = \mathcal{B}^{-1}b$
1) Solve for $x_2 : \hat{S}^{-1}Sx_2 = \hat{S}^{-1}(\hat{B}^T\hat{A}^{-1}b_1 - b_2)$
2) Set $x_1 = \hat{A}^{-1}(b_1 - \hat{B}^T x_2)$
```

Here if Krylov subspace method is used to solve for $x_2$, then matrix vector product with $S$ and a solve with $\hat{S}$ is needed. However, when the problem size i.e. $|\mathcal{N}_h|$ is large, it won't be feasible to do exact solve with $\hat{A}$, and we need to solve it inexactly, for example, using algebraic multigrid methods. In the later case, the decoupling of $x_1$ and $x_2$ as in Algorithm 4.1 is not possible, and we shall need matrix vector product with $\mathcal{A}$ (31) and a solve (forward sweep) with $\hat{\mathcal{B}}$. We discuss at the end of this subsection on how to take advantage of the special structure of $\hat{A}$ in both cases of exact and inexact solves.

As a preconditioner $\tilde{S}$ of $S$, we choose the preconditioner first proposed in [5]. The preconditioner is given as follows:

$$\tilde{S} = S_1\hat{A}^{-1}S_2 = -(\hat{B} - \tau^{1/2}K)\hat{A}^{-1}(\hat{B}^T - \tau^{1/2}\hat{A}), \tag{39}$$

where $K$ is the stiffness matrix from (19). We observe that the preconditioned Schur complement $\tilde{S}^{-1}S$ is not symmetric, in particular, not symmetric w.r.t. $\langle\cdot,\cdot\rangle_S$ or w.r.t. $\langle\cdot,\cdot\rangle_{\tilde{S}}$ which is a sufficient condition for the convergence of preconditioned conjugate gradient method [16, p. 262]. Hence we shall use GMRES in Saad [16, p. 269] that allows nonsymmetric preconditioners.

*1) Exact and Inexact Solve with $\hat{A}$:* In step 1 of Algorithm 4.1, we need to solve with $\hat{A}$ when constructing right hand side, and also in step 2. Let $P$ be a permutation matrix, then solving a system of the form $\hat{A}h = g$ is equivalent to solving $P^T\hat{A}h = P^Tg$ as $P^T$ is nonsingular. With a change of variable $Py := h$, we then solve for $y$ in

$$P^T\hat{A}Py = Pg, \tag{40}$$

and we set $h = Py$ to obtain the desired solution. By choosing $P$ that renumbers the nodes corresponding to the coincidence set, we obtain

$$P^T\hat{A}P = \begin{pmatrix} I & \\ & R^TP^T\hat{A}PR \end{pmatrix}, \tag{41}$$

where $R$ is the restriction operator defined as follows

$$R^TP^T\hat{A}PR = \hat{A}\restriction_{\mathcal{N}_h \setminus \mathcal{N}_h^\bullet}, \tag{42}$$

where

$$\mathcal{N}_h^\bullet = \{i \, : \, T(i,i) = 0\} \tag{43}$$

is nothing but set of active nodes. Here $R$ is explicitly given as follows:

$$R = \begin{pmatrix} \begin{pmatrix} 0 & 0 & \dots & 0 \\ \vdots & \dots & \dots & 0 \\ 0 & 0 & \dots & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \ddots & \dots & 0 \\ 0 & 0 & \dots & 1 \end{pmatrix} \end{pmatrix}, \; R \in \mathbb{R}^{|\mathcal{N}_h| \times |\mathcal{N}_h \setminus \mathcal{N}_h^\bullet|} \tag{44}$$

Let $\hat{K} = TKT$, we have

$$\begin{aligned} R^TP^TDPR &= R^TP^T\hat{A}PR \\ &= R^TP^T\epsilon(\hat{K} + \hat{m}\hat{m}^T)PR \\ &= \epsilon(R^TP^T\hat{K}PR + R^TP^T\hat{m}\hat{m}^TPR), \end{aligned}$$

where

$$R^TP^T\hat{K}PR = (P^TKP)|_{\mathcal{N}_h \setminus \mathcal{N}_h^\bullet}, \; \hat{K} = TKT, \; \hat{m} = Tm, \tag{45}$$

where $m$ is the rank-one term defined in (20). For convenience of notation, we write

$$R^TP^T\hat{A}PR = \epsilon(\widetilde{K} + \tilde{z}\tilde{z}^T), \tag{46}$$

where $\widetilde{K} = R^TP^T\hat{K}PR$ and $\tilde{z} = R^TP^T\hat{z}$. In the new notation, we have

$$P^T\hat{A}P = \begin{pmatrix} I & \\ & \epsilon(\widetilde{K} + \tilde{z}\tilde{z}^T) \end{pmatrix}. \tag{47}$$

Thus (40) now reads

$$\begin{pmatrix} I & \\ & \epsilon(\widetilde{K} + \tilde{z}\tilde{z}^T) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = Pg =: \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}, \tag{48}$$

which reduces to two-set of equations

$$y_1 = g_1, \tag{49}$$

$$\epsilon(\widetilde{K} + \tilde{z}\tilde{z}^T)y_2 = g_2. \tag{50}$$

To solve the latter, we use the Sherman-Woodbury formula

$$(\widetilde{K} + \tilde{m}\tilde{m}^T)^+ = \widetilde{K}^+ - \frac{\widetilde{K}^+\tilde{m}\tilde{m}^T\widetilde{K}^+}{1 + \tilde{m}^T\widetilde{K}^+\tilde{m}}. \tag{51}$$

The AMG discussed before is used to pseudo-invert $\widetilde{K}$, and we thus avoid constructing the dense matrix which would be the case when rank one term is explicitely added.

## V. Numerical Experiments

All the experiments were performed in double precision arithmetic in MATLAB. The Krylov solver used was GMRES with subspace dimension of 200, and maximum number of iterations allowed was 300. The iteration was stopped as soon as the relative residual was below the tolerance of $10^{-7}$.

We consider two samples of active set configurations that occur when a square region evolves as shown in figures 1 and 2. The region between the two squares and the circles is the interface between two bulk phases taking values +1 and -1; initially we chose random values between -0.3 and 0.5 in the interface region. The width of the interface is kept to be 10 times the chosen mesh size. The time step $\tau$ is chosen to be equal to $\epsilon$. We compare various mesh sizes leading to number of grid points upto just above 1 million, and compare various values of epsilon for each mesh sizes. We observe that the number of iterations remain independent of the mesh size, however it depends on $\epsilon$. But we observe that for a fixed epsilon, with finer mesh, the number of iterations actually decrease significantly. For example the number of iterations for $h = 2^{-7}, \epsilon = 10^{-6}$ is 84 but the number of iterations for $h = 2^{-10}, \epsilon = 10^{-6}$ is 38, a reduction of 46 iterations! It seems that finer mesh size makes the preconditioner more efficient. We also observe that the time to solve is proportional to number of iterations; the inexact solve for the (1,1) block remains optimal because the (1,1) block is essentially Laplacian for which AMG remains very efficient.

## VI. Conclusion

For the solution of large scale optimization problem corresponding to Cahn-Hilliard problem with obstacle problem, we proposed an efficient preconditioning strategy that requires two elliptic solves. In our initial experiments upto over million unknowns, the preconditioner remains mesh independent. Although, for coarser mesh, there seems to be strong dependence on epsilon, but as the mesh becomes finer, we observe a significant reduction in iteration count, thus making the preconditioner effective and useful on finer meshes. It is likely that the iteration count continues to decrease on finer meshes.

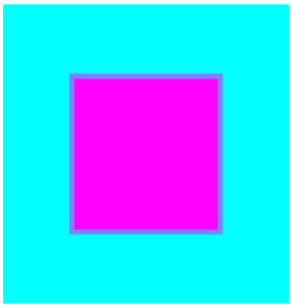


Fig. 1. Square

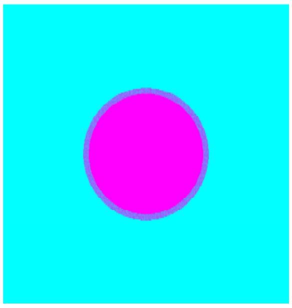


Fig. 2. Circle

TABLE I
COMPARE ITERATIONS COUNT FOR VARIOUS $\epsilon$ AND $h$

| | | square | | circle | |
|---|---|---|---|---|---|
| $h$ | $\epsilon$ | its | time | its | time |
| $2^{-7}$ | e-2 | 6 | 1.40 | 6 | 1.37 |
| | e-3 | 8 | 2.00 | 9 | 2.16 |
| | e-4 | 20 | 4.50 | 22 | 4.76 |
| | e-5 | 41 | 9.23 | 45 | 10.14 |
| | e-6 | 77 | 17.83 | 83 | 22.02 |
| $2^{-8}$ | e-2 | 6 | 6.22 | 6 | 4.61 |
| | e-3 | 5 | 4.88 | 5 | 5.12 |
| | e-4 | 13 | 10.56 | 15 | 11.91 |
| | e-5 | 31 | 24.39 | 35 | 27.31 |
| | e-6 | 60 | 50.44 | 67 | 54.09 |
| $2^{-9}$ | e-2 | 5 | 18.13 | 6 | 18.50 |
| | e-3 | 5 | 16.57 | 5 | 18.92 |
| | e-4 | 8 | 28.21 | 9 | 32.09 |
| | e-5 | 22 | 72.59 | 25 | 80.03 |
| | e-6 | 47 | 166.46 | 52 | 180.58 |
| $2^{-10}$ | e-2 | 5 | 81.86 | 6 | 89.16 |
| | e-3 | 5 | 81.69 | 5 | 85.23 |
| | e-4 | 5 | 98.62 | 5 | 97.76 |
| | e-5 | 14 | 218.52 | 15 | 254.48 |
| | e-6 | 34 | 527.49 | 38 | 612.41 |

## REFERENCES

[1] J. W. Cahn and J. E. Hilliard, "Free Energy of a Nonuniform System. I. Interfacial Free Energy," *The Journal of Chemical Physics*, vol. 28, no. 2, 1958. [Online]. Available: http://dx.doi.org/10.1063/1.1744102

[2] Y. Oono and S. Puri, "Study of phase-separation dynamics by use of cell dynamical systems. I. Modeling," *Physical Review A*, vol. 38, no. 1, 1987. [Online]. Available: http://dx.doi.org/10.1103/PhysRevA.38.434

[3] J. F. Blowey and C. M. Elliott, "The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy Part I: Numerical analysis," *European J. Appl. Math.*, no. 2, pp. 233–280, 1991. [Online]. Available: http://dx.doi.org/10.1017/S095679250000053X

[4] ——, "The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy Part II: Numerical analysis," *European J. Appl. Math.*, no. 3, 1992. [Online]. Available: http://dx.doi.org/10.1017/S0956792500000759

[5] J. Bosch, M. Stoll, and P. Benner, "Fast solution of Cahn-Hilliard variational inequalities using implicit time discretization and finite elements," *Journal of Computational Physics*, vol. 262, pp. 38–57, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jcp.2013.12.053

[6] C. Graeser and R. Kornhuber, "Nonsmooth newton methods for set-valued saddle point problems," *SIAM Journal on Numerical Analysis*, vol. 47, no. 2, pp. 1251–1273, 2009.

[7] J. Barrett, R. Nurnberg, and V. Styles, "Finite element approximation of a phase field model for void electromigration," *SIAM J. Numer. Anal.*, vol. 42, no. 2, pp. 738–772, 2004. [Online]. Available: http://dx.doi.org/10.1137/S0036142902413421

[8] R. Kornhuber, "Monotone multigrid methods for elliptic variational inequalities I," *Numerische Mathematik*, vol. 69, no. 2, pp. 167–184, 1994.

[9] ——, "Monotone multigrid methods for elliptic variational inequalities II," *Numerische Mathematik*, vol. 72, no. 4, pp. 481–499, 1996.

[10] J. Mandel, "A Multilevel lterative Method for Symmetric, Positive Definite Linear Complementarity Problems," *Applied Mathematics and Optimization*, vol. 11, pp. 77–95, 1984.

[11] C. Graser and R. Kornhuber, "Multigrid Methods for Obstacle Problems," *Journal of Computational Mathematics*, vol. 27, no. 1, pp. 1–44, 2009.

[12] C. Graser, "Convex Minimization and Phase Field Models," Ph.D. dissertation, FU Berlin, 2011.

[13] P. Kumar, "Aggregation based on graph matching and inexact coarse grid solve for algebraic two grid," *International Journal of Computer Mathematics*, vol. 91, no. 5, pp. 1061–1081, 2014. [Online]. Available: http://dx.doi.org/10.1080/00207160.2013.821115

[14] U. Trottenberg, C. Oosterlee, and A. Schuller, *Multigrid*. Academic Press, 2001. [Online]. Available: http://www.academicpress.com

[15] Y. Notay, "An aggregation-based algebraic multigrid method," *Electronic Transactions on Numerical Analysis*, vol. 37, pp. 123–146, 2010. [Online]. Available: http://dx.doi.org/10.1109/ISQED.2007.31

[16] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed. Philadelphia: SIAM, 2003. [Online]. Available: http://dx.doi.org/10.1137/1.9780898718003

# Computer Science & Systems

CSS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to more technical aspects of computer science and related disciplines. The CSNS area spans themes ranging from hardware issues close to the discipline of computer engineering via software issues tackled by the theory and applications of computer science and to communications issues of interest to distributed and network systems. Events that constitute CSNS are:

- BCPC'15—1st International Workshop on Biological, Chemical and Physical Computations
- CANA'15—8th Computer Aspects of Numerical Algorithms
- IWCPS'15—2nd International Workshop on Cyber-Physical Systems
- MMAP'15—8th International Symposium on Multimedia Applications and Processing
- WAPL'15—5th Workshop on Advances on Programming Languages

# 1st International Workshop on Biological, Chemical and Physical Computations

THE First International Workshop on Biological, Chemical and Physical Computations (BCPC'15) is a unique interdisciplinary workshop which brings together computer scientists and engineers with biologists, chemists, and physicists to initiate development of novel paradigms, architectures and implementations of computing devices adopting principles of information processing in physical, chemical and biological systems. The workshop aims to bring together world-leading scientists whose research focuses on non-traditional theoretical machines, experimental prototypes and genuine implementations of non-classical computing devices, who try to revisit existing approaches in unconventional computing, provide scientists and engineers with blueprints of realizable computing devices, and take a critical glance at the design of novel and emergent computing systems to point out failures and shortcomings of both theoretical and experimental approaches.

## Topics

The topics of interest include, but are not limited to:

- Physics of computation,
- Slime mould computing,
- Social insects computing,
- Chemical computing,
- Bio-molecular computing,
- Cellular automata as models of massively parallel computing,
- Logics of unconventional computing,
- Reaction-diffusion computing,
- Molecular machines incorporating information processing,
- Memristors,
- Organic electronics,
- Noise-based computing,
- Novel hardware systems,
- Mechanical computing,
- Physical limits to mechanical computation.

## Event Chairs

**Adamatzky, Andrew,** Professor, UWE, Bristol, UK, United Kingdom

**Pancerz, Krzysztof,** University of Management and Administration in Zamość, Poland

**Schumann, Andrew,** University of Information Technology and Management in Rzeszow, Poland

## Program Committee

**Akl, Selim,** Queen's School of Computing, Canada

**Armstrong, Rachel,** Architecture, Planning and Landscape, University of Newcastle, United Kingdom

**Asai, Tetsuya,** Laboratory of Advanced LSI Engineering, Hokkaido University, Japan

**Costa, Jose-Felix,** Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Portugal

**Costello, Ben De Lacy,** Centre for Research in Analytical, Material and Sensor Sciences, University of the West of England, United Kingdom

**Di Ventra, Max,** Department of Physics, University of California, United States

**Dittrich, Peter,** Bio Systems Analysis Group, Friedrich-Schiller-University Jena, Germany

**Erokhin, Victor,** Departament of Phisics, University of Parma, Italy

**Gorecki, Jerzy,** Institute of Physical Chemistry, PAN, Poland

**Hanczyc, Martin,** Institute of Physics Chemistry and Pharmacy, University of Southern Denmark, Denmark

**Miranda, Eduardo,** University of Plymouth, United Kingdom

**Petre, Ion,** Computer Science Programme at Åbo Akademi University, Finland

**Shirakawa, Tomohiro,** National Defense Academy of Japan, Japan

**Sirakoulis, Georgios,** Department of Electrical & Computer Engineering, Democritus University of Thrace, Greece

**Stannett, Mike,** University of Sheffield, United Kingdom

**Stepney, Susan,** University of York, United Kingdom

**Tsuda, Soichiro,** School of Chemisty, University of Glasgow, United Kingdom

**Tucker, John,** Department of Computer Science, Swansea University, United Kingdom

**Zauner, Klaus-Peter,** Faculty of Physical Sciences and Engineering, University of Southampton, United Kingdom

**Zelinka, Ivan,** Department of Computer Science, VSB Technical University in Ostrava, Czech Republic

DOI: 10.15439/2015F237
ACSIS, Vol. 5

# Basic Transitions of *Physarum Polycephalum*

Alice Dimonte
IMEM-CNR
Parco Area delle Scienze 37/A.
Parma, 43124, Italy
Email: alice.dimonte@imem.cnr.it

Tatiana Berzina
IMEM-CNR
Parco Area delle Scienze 37/A.
Parma, 43124, Italy
Email: tatiana.berzina@fis.unipr.it

Victor Erokhin
IMEM-CNR
Parco Area delle Scienze 37/A.
Parma, 43124, Italy
Email: victor.erokhin@fis.unipr.it

***Abstract*—The main charter of this work is the organism *Physarum polycephalum*, in particular plasmodium, *Physarum*'s vegetative phase. During this latter form, the organism is more active and moves searching for food. Plasmodium behaves like a giant amoeba, and more interestingly, its way of foraging can be interpreted as a computation. By comparing the reaction of this organism with attractors and repellents, knowing its capability of solving computational problems with natural parallelism, we dedicated the present work to study the behavior of *Physarum polycephalum* slime mold under different conditions.**

## I. Introduction

*PHYSARUM polycephalum's* behavior [1] has been the object of different studies belonging to different research fields, from biology to unconventional computing and from experimental to theoretical science. Thus, considering the sentence just mentioned, the question coming out is "Why does *Physarum* attract so many different scientists?"

It belongs to one of the predominant groups in the world of eukaryotes: Amoebozoa [2] a family hundreds millions of years old. Although they are probably one of the first forms of life on Earth, few are data about them because of the evolutionary variation happened across the years to this group. In this field, most of the interest of scientists is focused on slime molds [3]. *Physarum polycephalum*, in particular, can be considered as a reaction-diffusion medium enwrapped in a growing membrane. Moreover, the plasmodium of *Physarum* is simple enough to be modeled and implemented in unconventional computing algorithms as a non-linear media. However, on the other hand, it is a robust organism, reach of challenging features allowing many different computational studies [4].

In [5] *Physarum* has been compared to a labelled transition system whose behavior has been modeled by a rough set models, implemented in a specific language and properly developed to describe *Physarum polycephalum's* response to spatial configuration of stimuli [6]. Additionally, its foraging behavior has been related to computation [4], in fact, attracting/repelling sources are able to induce in it excitation waves, affecting its motion. Thus, because of the parallel processing of inputs and outputs, it can be seen also as a massive-parallel amorphous computer, taking data from food sources in the space around and giving, as a result, the protoplasmic network created by its own body [7]. A recent example of *Physarum's* capability of designing optimized networks, allows the retracing of Canadian Highway Networks [8]. Additionally it has been developed an algorithm [9] able to find the shortest path, as slime mold naturally does. Among all these feature, its behavior can be resumed as: capability of optimizing shortest path [10], calculation of proximity and planar graphs [11], calculation of basic logical operation [12], motion control [13], self-adaptation [14], and self-reparation [15].

This paper shows experiments done to study the behavior of this organism under different conditions and stimuli. In particular, we analyzed not only spatial distribution of attractants/repellents, but also the effect of external elements inserted and transported by *Physarum polycephalum*. Therefore, considering the obtained results, we can say that it should be possible to program this transition system [5] to realize deterministic adaptive networks and spatial distribution of nanoscale and microscale materials. Choosing, among transported materials, those able to change electrical, optical and, magnetic properties of the system it should be possible to enable information sensing and processing. The work is organized as follows: the first section analyzes *Physarum*'s behavior in presence of rose and geranium, while the second one is dedicated to chemical compounds. Then, we distributed food around the experimental area in a particular way, like the hours of a clock, in order to see how would appear *Physarum*'s realized network. The fourth study is devoted to light, a strong repellent for the mold, able also to induce sclerotization, its dormient phase, assumed in case of not proper environment conditions [16]. The fifth section presents the reaction of *Physarum* to music and sounds of different frequencies. As was already presented in [17] there is a correlation between

This work has been financially supported by: European project FP7-ICT-2011-8 "PhyChip-Physarum Chip: Growing Computers from Slime Mould" Grant Agreement n.316366.

motion of plasmodium and the music or sound frequency. Last section describes the experiments performed using *Physarum* as a bio-robot, able to transport microparticles whithin its own body, as in [18]. In particular, we worked with polystyrene and $MnCO_3$ microparticles.

The results here reported can be useful for unconventional computing and bio-computing application devices. In such field, *Physarum polycephalum's* networks become the scaffolds of hybrid circuits at the micro and nano-scale. For the development of self-growing computational systems, we must consider and study the slime mold as a medium capable of integrating and redistributing foreign particles. Therefore, *Physarum polycephalum* acts as a programmable transport medium as showed in [19].

.

## II. Experiments

*Physarum polycephalum*'s colony has been grown starting from sclerotized samples. The organism was fed with oat-flakes and kept in a dark and humid chamber in 9 cm diameter Petri dishes with 1,5% Agar non nutrient gel. Moreover, in order to allow a proper grown and safety conditions for the colony itself, the organism was periodically replanted to fresh Agar Petri.

### A. Biological elements

In wild nature *Physarum polycephalum* can be found in the underwood, near leaves, where the sun light is partially shielded by the trees. Therefore, it is reasonable to think that the mold could meet, during its motion, elements like flowers, leaves, grass, etc.

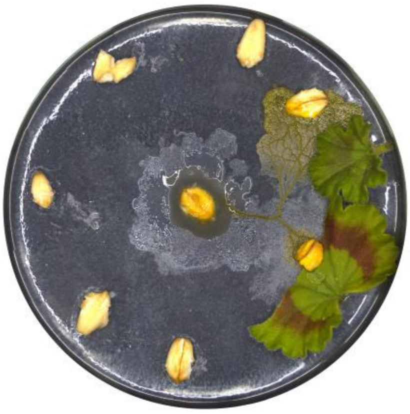


Fig. 1 Optical microscope photograph of a Petri dish with *Physarum polycephalum* growing networks firstly in the direction of the geranium.

In this section, we used as biological elements, geranium's leaves and rose-hip petals. Photographs in Fig.1 and Fig.2 show two similar experimental set-up. *Physarum polycephalum* was in the middle of a Petri dish, with Agar non nutrient gel, and starting from an oat-flake. Then, we distributed oat flakes along the perimeter, but near two of them, we added respectively: geranium's leaves in the first case and rose-hips petals in the second. The experiments have been developed keeping the set-up in the dark and humid chamber exploited for maintaining the colony and taken out of just for checking. After 10 hours, *Physarum* reached the oat-flakes near the leaves or petals, as shown respectively in Fig. 1 and Fig. 2. Repeating each case 10 times, we verified that mold's behavior was always the same. Therefore, we can conclude that both geranium and rose-hips work as attractors for *Physarum*.

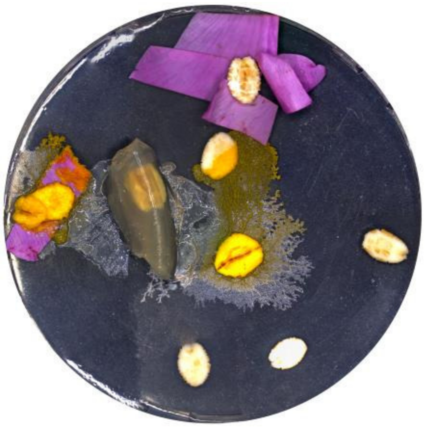


Fig. 2 Optical microscope photograph of a Petri dish with *Physarum polycephalum* growing networks firstly in the direction of rose-hip petals.

### B. Chemicals

In this experimental section, we studied slime mold's behavior in presence of two polysaccharides compounds, able to create a gel: pectin and chitosan. Firstly, we prepared the samples by dipping a circular glass, 10 mm in diameter, in the polysaccharide solution. The latter when drying forms a homogeneous gel layer on which we posed a small blob of *Physarum.* Then, the so created sample was positioned into a Petri dish with Agar non-nutrient gel, oat-flakes were spread around the glass to push *Physarum* growing its networks on the substrate and reaching them. As in the case of biological elements, the experimental Petri dishes were kept in the dark and humidity. We developed the same set-up for chitosan and for pectin; in both cases and for all the experiments, we got positive results, i.e. the mold survived creating networks towards the food. Fig. 3 shows one of the chitosan experiments performed, after one night. The mold reached the food surviving and forming protoplasmic tubes across the substrate, thus, proving the biocompatibility of chitosan for *Physarum*.

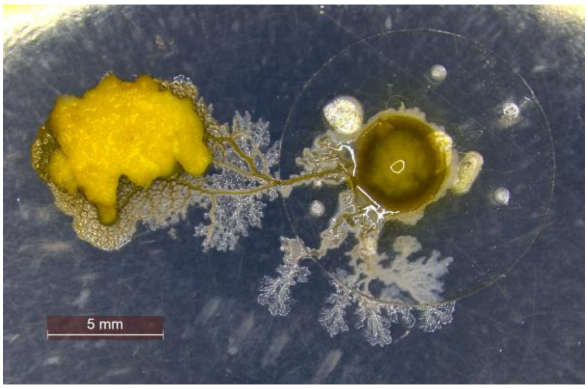


Fig. 3 Optical microscope photograph of a glass with chitosan from which the mould started its motion

Fig. 4 shows an experiment performed with pectin. The sample has been photographed after 14 hours; *Physarum* created networks and veins of protoplasmic tubes going out of the glass and reaching the oat-flakes. Therefore, also pectin can be considered a biocompatible element for slime mold.

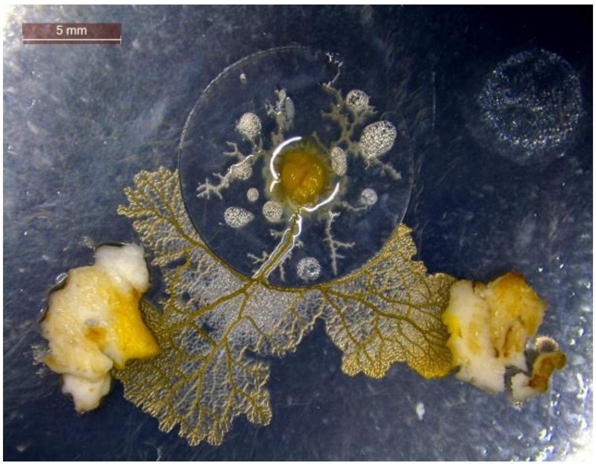


Fig. 4 Optical microscope photograph of a glass with pectin from which the mold started its motion

### *C. Clock*

We developed this series of experiment in order to understand if *Physarum* has a preferential direction, and how its networks will appear, considering a so particular food distribution.

These are three sequential optical microscope photographs, see Fig. 5, of the same sample defined as *clock*. Pictures refer to three times interval, after 24, 48 and 72 hours respectively.

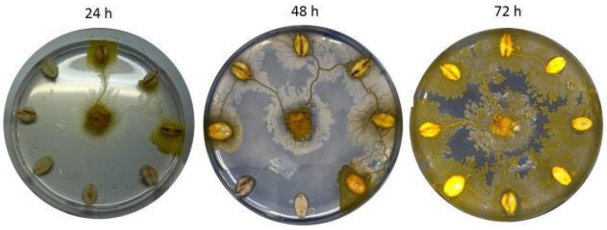


Fig. 5 Photograph of the same sample after 24h-48 and 72 hours of *Physarum* growing on agar non nutrient gel with oat-flakes distributed like hours in a clock

The mold started from the middle of the Petri, placed on an oat-flake; then, after disposing all other oats along the perimeter, like clock's hours. The sample was kept in a dark and humid environment for the whole time, and just photographed at the time-break abovementioned. As it is visible from the first Petri (i.e. after 24 hours), *Physarum polycephalum* created 2 protoplasmic tubes in the direction of two neighbors oats and, than it grew at the same time in the two opposite directions. Fig. 6 is another photograph of *clock* experiment after 12 and 24 h.

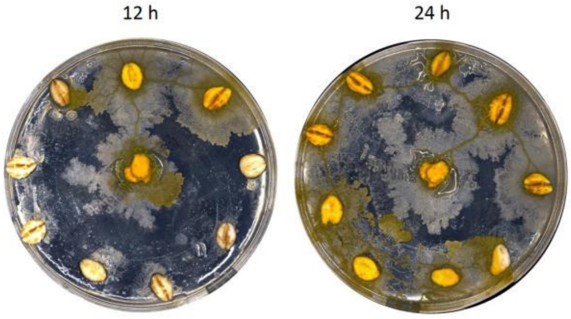


Fig. 6 Photograph of a sample in the clock series after 12 and 24 hours

Fig. 7 is another experiment of *clock*, but it is different from previous cases. Here oat-flakes have been placed closer to the starting point (in the middle of the Petri) and, as it is clearly visible, *Physarum* started on three oat flakes.

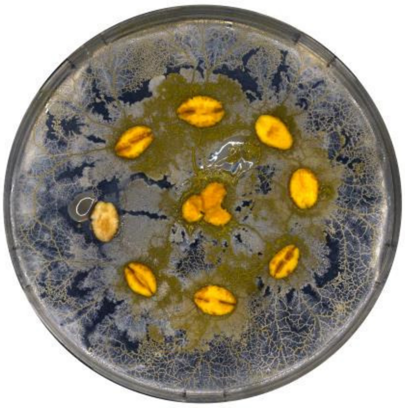


Fig. 7 Photograph of a sample in the *clock* series, in this case mold started from three oat-flakes in the middle of the Petri. Other flakes have been placed closer than in other cases.

What we got after all these experiments was that Physarum seems to have a preferential direction, on right, as many animals and insects [20]. However, there is still not statistical significance about these right decisions.

### *D. Light*

In Fig. 8 we worked with light and *Physarum*, studying the repellent effect that the first has to the latter. Starting from the generally used Petri with Agar gel inside, we masked with a black tape the outer part of the Petri dish, leaving transparent just an area with an *S* shape. Therefore, we placed a cold light source underneath the Petri and developed the experiment. What stands out from Fig. 8 is that *Physarum*, from its starting point moved in the food direction avoiding the illuminated area of the Petri.

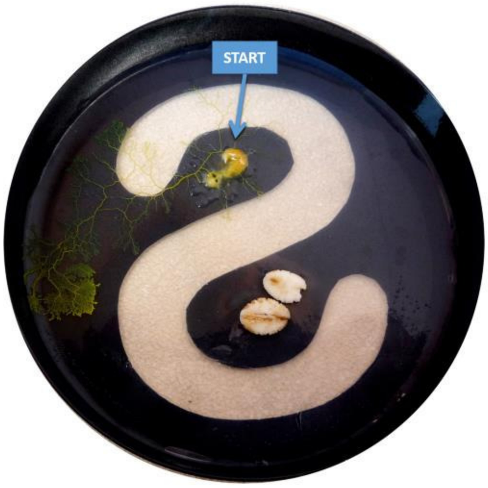


Fig. 8 Photograph of a sample in which has been tested the repellent effect of light on *Physarum polycephalum*

### *E. Music*

The Photograph of Fig. 9 presents a series of three Petri Dishes with the same set-up: agar non-nutrient gel, the same quantity of *Physarum* starting from the middle and 4 oats placed at 45 degree one with respect to the other.

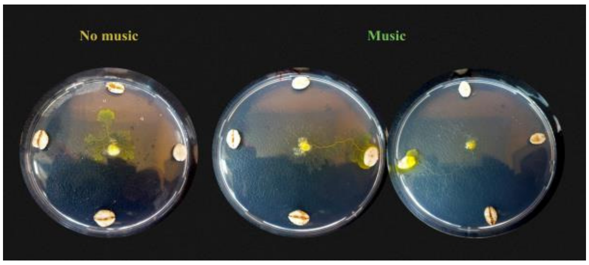


Fig. 9 Photograph of three sample in which the effect of Mozart's music on the mold's speed has been tested

Therefore, two of these samples were placed near a sound source, in particular Mozart symphony; the third one was in a silent place. What we observed, not only in these cases, but in all the so developed experiments, was that the frequencies of Mozart's music have a positive effect on *Physarum polycephalum* increasing its speed. Moreover, we performed other experiments changing kind of music, in particular hard rock songs (ACDC and Iron Maiden) and heavy metal (Carribean Corps and Marilyn Manson). It was interesting to verify that, in case of hard rock, there was not appreciable difference in the speed of *Physarum* samples under music with respect to those in a silence environment. However, when the mold was exposed to heavy metal, the sound resulted in the fact that the organism reduced substantially its speed, and in a 10% of cases, the whole motion was interrupted and it starts sclerotizing. For this study, we repeated each experiment ten times placing in parallel a sample under music and another under silence, thus keeping the same surrounding conditions.

### *F. Microparticles*

We loaded *Physarum polycephalum* slime mold with two kind of microparticles:

1_ Polystyrene microparticles (bought by Sigma-Aldrich), an aqueous dispersion (% 0.5 p/p) of particles 3 μm in diameter.

2_ $MnCO_3$ microparticles (bought by Plasma Chem., Berlin) with a diameter of 3 μm in aqueous dispersion (% 0.5 p/p).

3_ $BaFe_{12}O_{19}$ nanoparticles (obtained by a sol-gel method [14]) were solved in a homogeneous suspension in pure water with a final concentration equal to 2 mg/L. The solution was sonicated before its deal.

The performed experiments had, as an objective to verify the compatibility and the ability of the mold to transport micro and nan-objects, of different nature. Microparticles were chosen because of their nature, in particular organics, polystyrene, and non organics, $MnCO_3$. The latter, nanoparticles of Barium hexaferrite were selected because they are magnetic, a necessary feature to develope the experiments we were planning to carry out.

Additionally, we studied two kinds of transport: first case we loaded *Physarum* with particles, by directly mixing the two elements together; in the second one, we observed what happened if *Physarum* had to cross a strip of particles. The latter study, however, was performed only with polystyrene microparticles. Moreover, it has to be distinguished two main transport mechanisms: in a first case, particles are transported on *Physarum* veins, in a second one, they are taken inside the body of the mold. The first case is typical of those substances that *Physarum* almost ignore; they are not dangerous, nor interesting for it. Thus, they are transported just because, during its motion, they sticked to its body, but they are not incorporated. Therefore, it is possible to find them only on the surface of mold's body, and not inside it. In the second case, particles are mistaken as food, so they are engulfed inside *Physarum* that picks them up and tries to absorb.

We proceeded in parallel with both micro-particles (polystyrene and $MnCO_3$) solutions: 20–30 µL of particles were directly added to the plasmodium and mechanically mixed. The so obtained solution, of *Physarum*-particles, was placed on a silicon substrate. To induce the organism in moving and, to understand if it is able to transport these object, and in which way, we placed oat-flakes on a second silicon substrate, posed next to the one the mold was starting from. Then, we placed samples in a dark and humid environment and, after 12 hours, loaded *Physarum* not only survived, but, creating networks of protoplasmic tubes, reached also the food on the second substrate. The latter was used to perform scanning electron microscopy (SEM) analysis, after sclerotizing the network by light illumination. In this way, we are sure that the particles, we will found, have to be transported by *Physarum.*

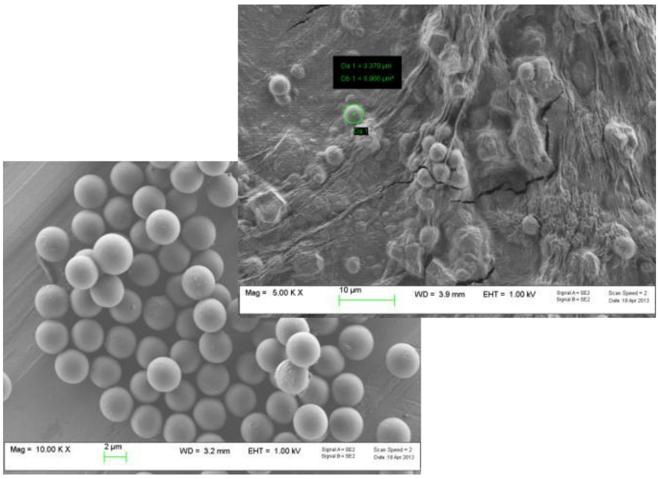


Fig. 10 Two SEM pictures: the one on left depicts polystyrene microparticles, while, on right, there is a sclerotic *Physarum* vein loaded with polystyrene microparticles by mechanically mixing method.

The right picture of Fig. 10 shows polystyrene particles within *Physarum*'s body. We found particles inside and on the surface of the mold, thus, *Physarum* incorporates and integrates polystyrene micro-objects. Moreover, we proved also that *Physarum* transports these materials, bringing them actively from the starting point, where they have been mixed, to the point the mold moved to in order to reach the food.

Figure 11 shows SEM measurements of a sclerotized vein of *Physarum* and $MnCO_3$ microparticles. The latter are clearly visible within the crack and near the veins of the mold.

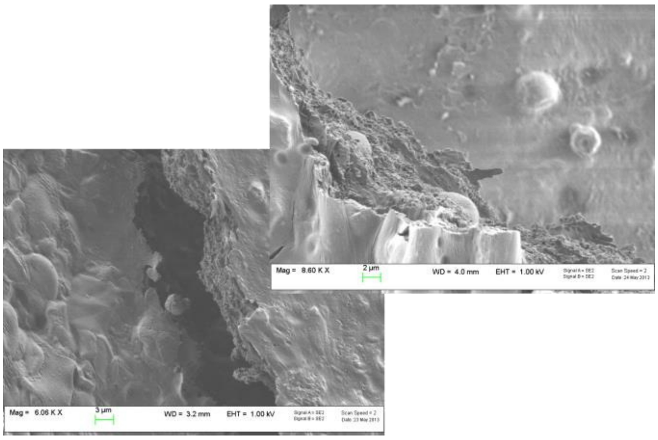


Fig. 11 SEM measurements performed on the sclerotized *Physarum* veins charged with MnCO3 particles. Also in this case the analyzed samples are those where *Physarum* moved to after loading. In both pictures particles are distinguishable within veins and cracks of the mold, in addition there are also two particles near the cut vein (right picture).

We performed a second type of study, just considering the polystyrene microparticles. Creating a strip of particles and an "ad hoc" experimental set-up, we forced *Physarum* in crossing particles' strip. Moreover, starting from a rounded glass, placed into a Petri dish with Agar non-nutrient gel, we designed a strip, one mm width, with the help of kapton, in a way that finally the strip divided the glass in two half. At this point, we placed a 2 µl of *Physarum* blob on one-half of the glass. However, in order to force *Physarum* in crossing the strip and not reaching the food, beyond the particles, by alternative paths, we created a semicircle with repellent agar, all around the mold. Therefore, the only possible way for *Physarum* to reach the food was through the particles strip. Fig. 12 shows that *Physarum polycephalum,* from the starting point, created protoplasmic tubes in the direction of food and, during its motion, crossed the strip made by particles. As in the previous cases, the mold transported particles that were found inside its slime after its crossing.

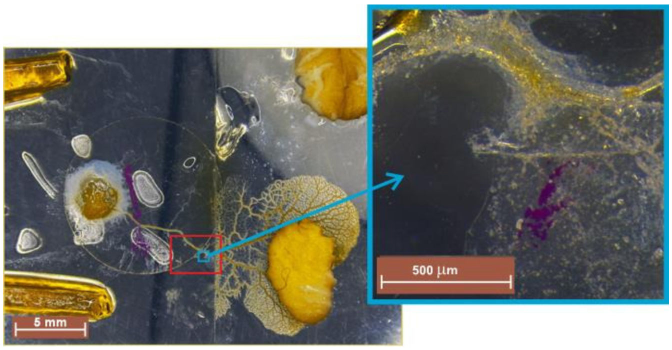


Fig. 12 Optical microscopy photograph of one of the developed experiment with slime mold and red polystyrene microparticles.

Last series of experiments have been performed with magnetic nanoparticles. In order to be able to recognize them once inside the mold we analyzed firstly the nanoparticles solution deposited on silicon substrate as in Fig.13

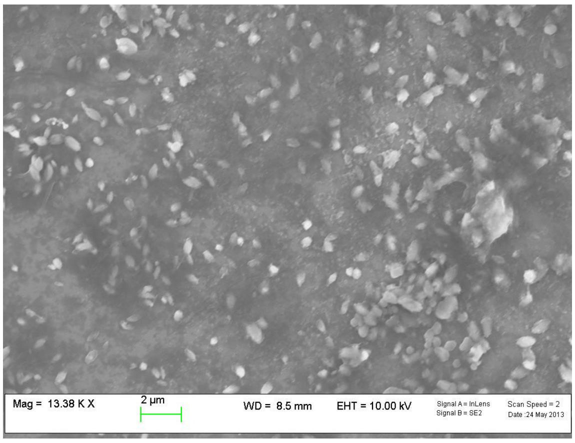


Fig. 13 SEM image of $BaFe_{12}O_{19}$ nanoparticles on silicon substrate

Slime mold was loaded with barium hexaferrite nanoparticles by mixing 100-200 µL of *Physarum* with 20-30 µL of the magnetic particles suspension.

Magnetic nanoparticles typically form big aggregates, due to the magnetic interactions, being ferromagnetic at that size. As in case of microparticles, we studied the efficiency of the slime mold loaded with magnetic nanoparticles. Firstly, we verified that the mixing procedure of the slime mold, with these particles, did not result in the organism death. As we saw, the slime mold kept its activity for more than a month, we focused the experiment on *Physarum*'s capability of transporting nanoparticles during its growth, proceeding exactly as in the case of polystyrene and $MnCO_3$ microparticles. However, even if *Physarum* moved to the second substrate, as we expected, a direct SEM imaging (Fig. 14) did not allow visualizing the presence of particles, due to their dimension that can be mistaken with nuclei of *Physarum*. For this reason, we fixed particles on the substrate surface and removed the slime mold

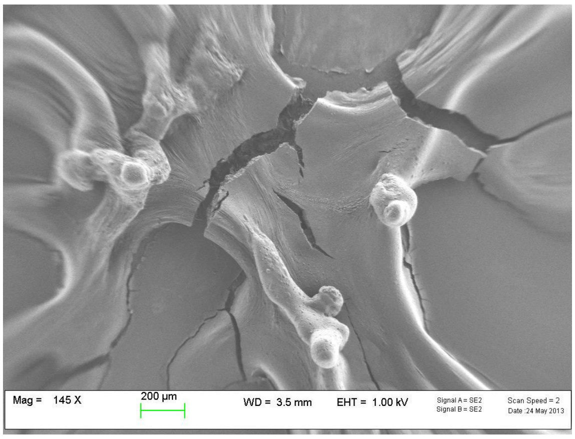


Fig. 14 SEM measurement of a sclerotized vein of Physarum polycephalum loaded with magnetic nanoparticles $BaFe_{12}O_{19}$. What stands out from this picture, as in the other done with such kind of particles inside the mold, are particular and curious 3D structures. The latter have been observed only in the samples when *Physarum* was loaded with magnetic nanoparticles of $BaFe_{12}O_{19}$

Considering that particles are magnetic, we decided to fix them by means of an external magnet placed under the sample where *Physarum* moved to by networks formation. Therefore, particles would be attracted by the magnet downwards and kept attached to the sample surface during the removal of the slime mold, by washing. Finally, the sample, always keeping the magnet attached, was dried and analyzed by SEM measurements. In order to show that the elements are really connected to the particles, element mapping was performed. The appropriate SEM image is shown in Fig. 15.

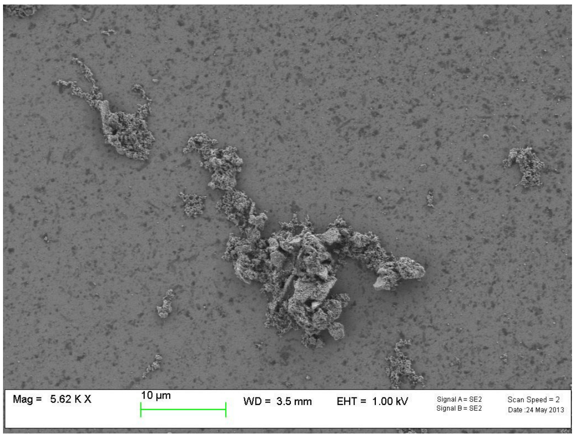


Fig. 15 SEM image of tubes formed by *Physarum* during its growth towards the food and subsequently washed, as described in the text.

## III. Conclusions

In this work we studied the behavior of *Physarum polycephalum* as an organism able to realize natural transition systems [4], *Physarum* is an interesting starting point for sensing, computing, novel biological substrate formation and circuit design. The possibility of targeting the growth with external disposition of attractants and repellents provides new capabilities for the realization of bio-inspired computational or robotic systems, based on *Physarum polycephalum*. In fact, such action can accelerate and/or suppress the propagation in certain directions. In order to illustrate this statement, let us consider a slime mold growth on the support, supplied with position sensors (in the simplest case – just electrodes, varying the capacity when the slime mold arrives to them). Moreover, analyzing *Physarum's* position, it will be possible to dispose attractants and/or repellent in such a way that the sequence of the food sources, where the slime mold will arrive, can be predetermined. By complementing with colored food [19], it will be possible to make any desirable mixture of colors. Of course, instead of the color, the food can include substances that, transported in special zones, evolve in some desired reactions. The final product will depend on the sequence of the reagents mixing. Thus, the *Physarum* will act as a bio-robotic system to delivery and control chemical reactions. The attractive and repellent parameters, such as the presence of food, light, humidity, chemistry of the surfaces, can be considered as inputs of this biocomputing/actuating system. The final distribution of the grown tubular pattern shows the system's output. The increase of the number of the control stimuli can improve the computational capabilities [11] of slime mold, especially if there will be a possibility of targeting the growing active zones of slime mold with robust external stimuli. The possibility, of targeting the growth with external actions, provided by author [21] serving *Physarum* also as a bio-living-robot, opens new capabilities for the realization of bio-inspired computational or robotic systems based on the slime mold. Future works involve the feeding of *Physarum* slime mold with nanoparticles that would be adsorbed in different way, depending on the biocompatibility; another field, that could be considered related to the previous one, concern nanoengineered polymeric capsules with shell made by mold-attractants, but filled with some solutions not so appealing for the bio-robot. These ideas open a huge amount of experiments, studies and analysis.

## Acknowledgment

We acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programmefor Research of the European Commission, under the Collaborative project PhyChip, grant agreement number 316366. We also acknowledge Yury Gunaza for the preparation of the artwork for the paper.

## References

[1] T. Kuroiwa, S. Kawano, and M. Hizume, "Studies on mitochondrial structure and function in physarum polycephalum. v. behaviour of mitochondrial nucleoids throughout mitochondrial division cycle," *J. cell Bio.*, vol. 72, no. 3, pp. 687–694, 1977.

[2] R. F. Watkins and M. W. Gray, "Sampling gene diversity across the supergroup amoebozoa: largest data sets from acanth amoeba castellanii, hartmannella vermiformis, physarum polycephalum, hyperamoeba dachnaya and hyperamoeba sp.," *Protist*, vol. 159, no. 2, pp. 269– 281, 2008.

[3] A. Fiore Donno, C. Berney, J. Pawlowski, and S. L. Baldauf, "Higher-order phylogeny of plasmodial slime molds (myxogastria) based on elongation factor 1-a and small subunit rrna gene sequences," *J. Eukaryotic Microbiol.*, vol. 52, no. 3, pp. 201–210, 2005B.

[4] A. Adamatzky, "Physarum machine: implementation of a kolmogorov-uspensky machine on a biological substrate," *Parallel Processing Letters*, vol. 17, no. 04, pp. 455–467, 2007.

[5] K. Pancerz, and A. Schumann, "Rough Set Models of Physarum Machines," *Int. J. General Syst.*, 2014.

[6] A. Schumann, and K. Pancerz, "Towards an Object-Oriented Programming Language for Physarum Polycephalum Computing." In Proceedings of the Workshop on Concurrency, Specification and Programming (CS&P'2013), edited by M. Szczuka, L. Czaja, and M. Kacprzak. 389–397. Warsaw, Poland.

[7] A. Tero, S. Takagi, T. Saigusa, K. Ito, D. P. Bebber, M. D. Fricker, K. Yumiki, R. Kobayashi, and T. Nakagaki, "Rules for biologically inspired adaptive network design," *Science*, vol. 327, no. 5964, pp. 439-442, 2010.

[8] A. Adamatzky, and S. G. Akl, "Trans-Canada slimeways: slime mould imitates the Canadian transport network," *ArXiv:1105.5084v1* [nlin.PS], pp. 1-18, 2011.

[9] X. Zhang, Q. Wang, A. Adamatzky, F. T. S. Chan, S. Mahadevan, and Y. Deng, "An improved Physarum polycephalum algorithm for the shortest path problem," *Scient. World J.*, vol. 2014, pp. 1-9, 2014.

[10] T. Nakagaki, H. Yamada, and M. Hara, "Smart network solutions in an amoeboid organism," *Biophys. Chem.*, vol. 107, pp. 1–5, January 2004.

[11] A. Adamatzky, "Developing proximity graphs by physarum polycephalum: does the plasmodium follow the Toussaint hierarchy?," *Parallel Process. Lett.*, vol. 19, no. 01, pp.105–127, 2009.

[12] M. Aono, S.-J. Kim, M. Hara, and T. Munakata, “Amoeba-inspired tug-of-war algorithms for exploration exploitation dilemma in extended bandit problem,” *Biosystems,* vol. 117, no. 0, pp. 1–9, 2014.

[13] A. Tero, R. Kobayashi, and T. Nakagaki, “Physarum solver: a biologically inspired method of road-networks navigation,” Phys. A: Statistical Mechanics and its Applications, vol. 363, no. 1, pp. 115–119, 2000.

[14] T. Saigusa, A. Tero, T. Nakagaki, and Y. Kuramoto, “Amoebae anticipate periodic events,” *Physical Review Letters,* vol. 100, p. 018101, 2008.

[15] W. Marwan, “Amoeba-inpsired network design,” *Science,* vol. 327, no. 5964, pp. 419-20, 2010.

[16] A. Takamatsu, E. Takaba and G. Takizawa “Environment-dependent morphology in plasmodium of true slime mold Physarum polycephalum and a network growth model,” *J. Theor.Biol.,* vol. 256, no. 1, pp. 29-44, 2009.

[17] E. Braund, and E. Miranda, “Music with unconventional computing: a system for Physarum polycephalum sound synthesis,” *Lect. Notes in Comp. Sci.,* vol. 8905, pp 175-189, 2014.

[18] A. Cifarelli, A. Dimonte, T. Berzina, and V. Erokhin, “On the loading of Physarum polycephalum with microparticles for unconventional computing application,” *BioNanoSci.,* vol. 4, pp. 92-96, 2014.

[19] A. Adamatzky, “Manipulating substances with Physarum polycephalum,” *Mat. Sci. and Eng. C,* vol. 30, pp. 1211–1220, 2010.

[20] E. R. Hunt, T. O’Shea-Wheller, G. F. Albery, T. H. Bridger, M. Gumn, and N. R. Franks, “Ants show a leftward turning bias when exploring unknown nest sites,” *Biol. Lett.,* vol. 10, pp. 1-4, 2015.

[21] A. Dimonte, A. Cifarelli, T. Berzina, V. Chiesi, P. Ferro, T. Besagni, F. Albertini, A. Adamatzky, and V. Erokhin, “Magnetic nanoparticles-loaded Physarum Polycephalum: directed growth and particles distribution, ” *Interdiscip. Sci. Life Sci.,* vol. 6, pp. 1-9, 2014.

DOI: 10.15439/2015F199
ACSIS, Vol. 5

# PhysarumSoft - a Software Tool for Programming Physarum Machines and Simulating Physarum Games

Andrew Schumann* and Krzysztof Pancerz* †
*University of Information Technology and Management in Rzeszów, Poland
Email: andrew.schumann@gmail.com
†University of Management and Administration in Zamość, Poland
Email: kpancerz@wszia.edu.pl

***Abstract*—In the paper, we describe selected functionality of the current version of a new software tool, called *PhysarumSoft*, developed for programming *Physarum* machines and simulating *Physarum* games. The tool was designed for the Java platform. A *Physarum* machine is a biological computing device implemented in the plasmodium of *Physarum polycephalum* or *Badhamia utricularis* that are one-cell organisms able to build programmable complex networks. The plasmodial stage of such organisms is a natural transition system that can be used as a medium for solving different computational tasks as well as creating bio-inspired strategy games.**

## I. Introduction

A *Physarum* machine is a programmable amorphous biological computing device, experimentally implemented in the plasmodium of *Physarum polycephalum*, also called true slime mould [1]. *Physarum polycephalum* is a single cell organism belonging to the species of order *Physarales*. In the considered case, the term of *Physarum* machine covers, in general, a hybrid device implemented in two plasmodia (cf. [2]), namely the plasmodium of *Physarum polycephalum* as well as the plasmodium of *Badhamia utricularis*. *Badhamia utricularis* is also the species of order *Physarales*. The plasmodium of *Physarum polycephalum* or *Badhamia utricularis*, spread by networks, can be programmable. In propagating and foraging behavior of the plasmodium, we can perform useful computational tasks. This ability was firstly discerned by T. Nakagaki et al. [3]. The *Physarum* machine comprises an amorphous yellowish mass with networks of protoplasmic veins, programmed by spatial configurations of attracting and repelling stimuli. When several attractants are scattered in the plasmodium range, the plasmodium looks for attractants, propagates protoplasmic veins towards them, feeds on them and goes on. Repellents play the role of elements blocking propagation of protoplasmic veins.

Solving computational tasks by means of *Physarum* machines is one of the main goals of the *Physarum* Chip Project: Growing Computers from Slime Mould [4] funded by the Seventh Framework Programme (FP7). In this project, we are going to construct an unconventional computer on programmable behaviour of *Physarum polycephalum*.

To program computational tasks for *Physarum* machines, we are developing a new object-oriented programming language [5], [6], [7], called a *Physarum* language. The *Physarum* language is a prototype-based language [8] consisting of inbuilt sets of prototypes corresponding to both the high-level models used for describing behaviour of *Physarum polycephalum* (e.g., ladder diagrams, transition systems, timed transition systems, Petri nets) and the low-level model (distribution of stimuli). More information is given in Section II.

Another task that can be performed in *Physarum* machine environments concerns bio-inspired strategy games. Fundamental topics of the research area related to bio-inspired games on *Physarum* machines were earlier considered, for example in [2] and [9]. Simulating *Physarum* games is considered in Section III.

To support reserach on programming *Physarum* machines and simulating *Physarum* games, we are developing a specialized software tool, called the *Physarum* software system, shortly *PhysarumSoft*. The tool was designed for the Java platform. In the paper, we describe selected functionality of the current version of *PhysarumSoft*. A general structure of this system is shown in Figure 1. We can distinguish three main parts of *PhysarumSoft*:

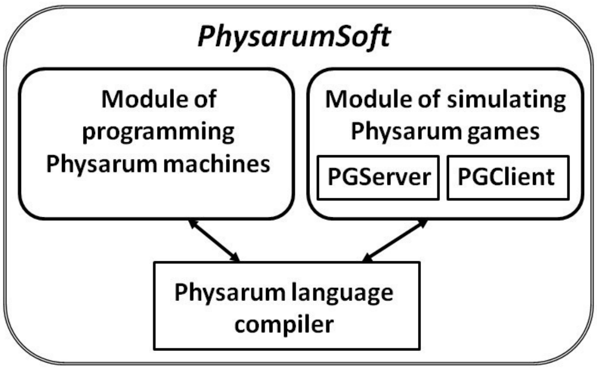

PhysarumSoft
Module of programming Physarum machines
Module of simulating Physarum games
PGServer
PGClient
Physarum language compiler

Fig. 1. A general structure of *PhysarumSoft*

- *Physarum* language compiler. The *Physarum* language is an object-oriented high-level programming language.

978-83-60810-66-8/$25.00©2015, IEEE

For generating the compiler of the language, the Java Compiler Compiler (JavaCC) tool [10] was used. JavaCC is the most popular parser generator for use with Java applications. For the programming purpose, a compiler embodied in our tool translates the high-level code describing a model of the *Physarum* machine into the spatial distribution (configuration) of stimuli (attractants, repellents) controlling propagation of protoplasmic veins of the plasmodium. A grammar of our language was described in [6].

- Module of programming *Physarum* machines described in Section II.
- Module of simulating *Physarum* games described in Section III.

The main features of *PhysarumSoft* are the following:

- Portability. Thanks to the Java technology, the created tool can be run on various software and hardware platforms. In the future, the tool will be adapted for platforms available in mobile devices and as a service in the cloud.
- User-friendly interface (see some screenshots shown in Sections II and III).
- Modularity. The project of *PhysarumSoft* and its implementation covers modularity. It makes the tool extend in the future easily.

## II. Programming Physarum Machines

To program *Physarum* machines (i.e., to set the spatial distribution (configuration) of stimuli (attractants, repellents) controlling propagation of protoplasmic veins of the plasmodium), we are developing a new object-oriented programming language [5], [6], [7], called the *Physarum* language. Our language is based on the prototype-based approach (cf. [8]) that is less common than the class-based one, although, it has a great deal to offer. This approach is also called classless or instance-based programming because prototype-based languages are based upon the idea that objects, representing individuals, can be created without reference to class-defining. In this approach, the objects, that are manipulated at runtime, are prototypes. In our language, there are inbuilt sets of prototypes corresponding to both the high-level models used for describing behaviour of *Physarum polycephalum* (e.g., ladder diagrams, transition systems, timed transition systems, Petri nets) and the low-level model (distribution of stimuli). According to the prototype-based approach, objects are created by means of a copy operation, called cloning, which is applied to a given prototype. Objects can be instantiated (cloned) via the keyword *new* using defined constructors. Different methods are used to manipulate features of the objects and create relationships between objects.

In our approach, the starting point, in programming the behaviour of the *Physarum* machine, is a high-level model describing propagation of protoplasmic veins of *Physarum*. We have proposed several high-level models used in programming *Physarum* machines, i.e.:

- ladder diagrams (see [11]),
- transition systems ([5]) and timed transition systems (see [12]),
- Petri nets (see [13]).

In the remaining part of this section, we recall basic definitions concerning transition systems, timed transition systems and Petri nets. They are general purpose tools that can be used to model dynamic systems with distinguished states and transitions between states. Application of ladder diagrams is restricted to modelling digital circuits. The recalled definitions are illustrated with some examples.

Transition systems are a simple and powerful tool for explaining the operational behaviour of models of concurrency. Formally, a transition system is a quadruple $TS = (S, E, T, I)$, cf. [14], where:

- $S$ is the non-empty set of states,
- $E$ is the set of events,
- $T \subseteq S \times E \times S$ is the transition relation,
- $I$ is the set of initial states.

Usually transition systems are based on actions which may be viewed as labelled events. If $(s, e, s') \in T$, then the idea is that $TS$ can go from $s$ to $s'$ as a result of the event $e$ occurring at $s$. Any transition system $TS = (S, E, T, I)$ can be presented in the form of a labelled graph with nodes corresponding to states from $S$, edges representing the transition relation $T$, and labels of edges corresponding to events from $E$.

The behaviour of *Physarum* machines is often dynamically changed in time. It is assumed, in the transition systems mentioned earlier, that all events happen instantaneously. Therefore, in [12], we proposed to use another high-level model, based on timed transition systems [15]. In the timed transition systems, timing constraints restrict the times at which events may occur. The timing constraints are classified into two categories: lower-bound and upper-bound requirements.

Let $N$ be a set of nonnegative integers. Formally, a timed transition system $TTS = (S, E, T, I, l, u)$ consists of:

- an underlying transition system $TS = (S, E, T, I)$,
- a minimal delay function (a lower bound) $l : E \rightarrow N$ assigning a nonnegative integer to each event,
- a maximal delay function (an upper bound) $u : E \rightarrow N \cup \infty$ assigning a nonnegative integer or infinity to each event.

In *Physarum* machines, timing constraints can be implemented through activation and deactivation of stimuli (attractants and/or repellents). Each state corresponds to either an original point of the plasmodium or an attractant. Especially, initial states of transition systems can be presented by original points, where protoplasmic veins originate from. Edges represent plasmodium transitions between attractants as well as the original points of the plasmodium.

In case of transition system and timed transition system models, the main prototypes defined in the *Physarum* language and their selected methods are collected in Table I.

Let us consider an exemplary timed transition system shown in Figure 2. Formally, we have $TTS = (S, E, T, I, l, u)$, where:

- $S = \{s_1, s_2, s_3, s_4\}$,

TABLE I
MAIN PROTOTYPES, CORRESPONDING TO TRANSITION SYSTEM AND TIMED TRANSITION SYSTEM MODELS, DEFINED IN THE PHYSARUM LANGUAGE, AND THEIR SELECTED METHODS

| Prototype | Selected methods |
|---|---|
| TS.State | *setDescription*, *setAsInitial* |
| TS.Event | *setDescription*, *setTimingConstraints* |
| TS.Transition | |

- $E = \{e_1, e_2, e_3\}$,
- $T = \{(s_1, e_1, s_2), (s_2, e_2, s_3), (s_3, e_3, s_4)\}$,
- $I = \{s_1, s_2\}$,
- $l(e_1) = l(e_2) = l(e_3) = 0$,
- and $u(e_1) = u(e_2) = \infty$, $u(e_3) = 3$.

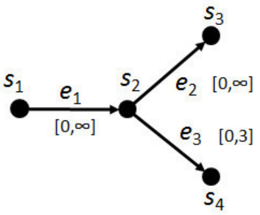
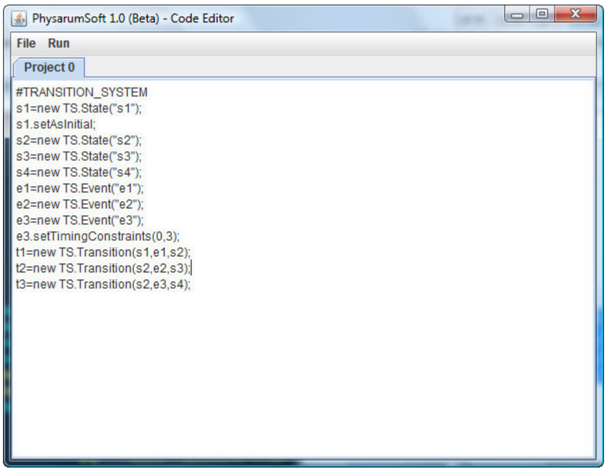


Fig. 2. An exemplary timed transition system $TTS$



```
#TRANSITION_SYSTEM
s1=new TS.State("s1");
s1.setAsInitial;
s2=new TS.State("s2");
s3=new TS.State("s3");
s4=new TS.State("s4");
e1=new TS.Event("e1");
e2=new TS.Event("e2");
e3=new TS.Event("e3");
e3.setTimingConstraints(0,3);
t1=new TS.Transition(s1,e1,s2);
t2=new TS.Transition(s2,e2,s3);
t3=new TS.Transition(s2,e3,s4);
```

Fig. 3. The code for the model in the form of $TTS$

The code, for the model in the form of $TTS$, in the *Physarum* language written in the editor of the module of programming *Physarum* machines, is shown in Figure 3. The result of compilation, i.e., the spatial distribution of stimuli (attractants, repellents) controlling propagation of protoplasmic veins of the plasmodium, is shown in Figure 4.

One can see that:

- *Physarum* $Ph\_1$ represents initial state $s_1$, attractants $A\_1$, $A\_2$, $A\_3$ represent states $s_2$, $s_4$, $s_3$, respectively.
- Splitting the plasmodium at $A\_1$ is supported by repellent $R\_1$.
- Repellent $R\_2$ is placed next to attractant $A\_2$ because timing constraints are set for event $e_3$.
- For $t > 3$, $R\_2$ must be activated to annihilate the vein of the plasmodium between $A\_1$ and $A\_2$.

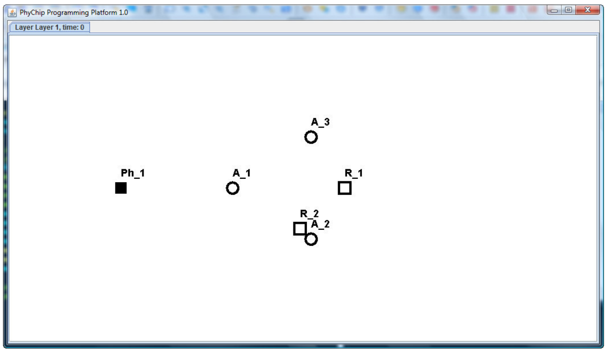


Fig. 4. The result of compilation

Petri nets introduced by C.A. Petri [16] are a formal tool used to model discrete event systems. In [13], we proposed to use Petri nets with inhibitor arcs (cf. [17]) to model behaviour of *Physarum polycephalum*. The inhibitor arcs test the absence of tokens in a place and they can be used to disable transitions. This fact can model repellents in *Physarum* machines. A transition can only fire if all its places connected through inhibitor arcs are empty (cf. [18]).

Formally, a marked Petri net with inhibitor arcs is a five-tuple

$$MPN = (Pl, Tr, Ar, w, m),$$

where:

- $Pl$ is the finite set of places (marked graphically with circles),
- $Tr$ is the finite set of transitions (marked graphically with rectangles),
- $Ar = Ar_O \cup Ar_I$ such that $Ar_O \subseteq (Pl \times Tr) \cup (Tr \times Pl)$ is the set of ordinary arcs (marked graphically with arrows) from places to transitions and from transitions to places whereas $Ar_I \subseteq Pl \times Tr$ is the set of inhibitor arcs (marked graphically with lines ended with small circles) from places to transitions,
- $w : Ar \to \{1, 2, 3, \dots\}$ is the weight function on the arcs,
- $m : Pl \to \{0, 1, 2, \dots\}$ is the initial marking function on the places.

In describing the Petri net behaviour, it is convenient to use for any $t \in Tr$:

- $I_O(t) = \{p \in Pl : (p, t) \in Ar_O\}$ - a set of input places connected through ordinary arcs to the transition $t$,
- $I_I(t) = \{p \in Pl : (p, t) \in Ar_I\}$ - a set of input places connected through inhibitor arcs to the transition $t$,
- $O(t) = \{p \in Pl : (t, p) \in Ar_O\}$ - a set of output places connected through ordinary arcs from the transition $t$.

In the proposed approach, we have additionally assumed the following limits for the Petri net:

TABLE II
THE MEANING OF TOKENS IN PLACES REPRESENTING CONTROL STIMULI

| Token | Meaning |
|---|---|
| Present | Stimulus activated |
| Absent | Stimulus deactivated |

TABLE III
THE MEANING OF TOKENS IN PLACES REPRESENTING OUTPUT STIMULI

| Token | Meaning |
|---|---|
| Present | Stimulus occupied by plasmodium of *Physarum polycephalum* |
| Absent | Stimulus not occupied by plasmodium of *Physarum polycephalum* |

TABLE IV
MAIN PROTOTYPES, CORRESPONDING TO PETRI NET MODELS, DEFINED IN THE *Physarum* LANGUAGE, AND THEIR SELECTED METHODS

| Prototype | Selected methods |
|---|---|
| PN.Place | *setDescription*, *setRole* |
| PN.Transition | *setDescription* |
| PN.Arc | *setAsInhibitor*, *setAsBidirectional* |

- $w(a) = 1$ for each $a \in Ar$,
- $m(p) \leq 1$ for each $p \in Pl$ (the capacity limit).

If $m(p) = 1$, then a token (i.e., a black dot) is drawn in the graphical representation of the place $p$. Assuming limits as the ones above, a transition $t \in Tr$ is said to be enabled if and only if $m(p) = 1$ for all $p \in I_O(t)$, i.e., the token is present in all input places $p$ connected with the transition $t$ through the ordinary arcs, and $m(p) = 0$ for all $p \in I_I(t)$, i.e., the token is absent in all input places $p$ connected with the transition $t$ through the inhibitor arcs, and $m(p) = 0$ for all $p \in O(t)$, i.e., the token is absent in all output places $p$ of the transition $t$. If the transition $t$ is enabled, we say that it can fire. A new marking function $m' : Pl \to \{0, 1, 2, \dots\}$ defines the next state of the Petri net after firing the transition $t$:

$$m'(p) = \begin{cases} m(p) - 1 & \text{if } p \in I_O(t) \text{ and } p \notin O(t), \\ m(p) + 1 & \text{if } p \in O(t) \text{ and } p \notin I_O(t), \\ m(p) & \text{otherwise.} \end{cases}$$

It is worth noting that in all figures including Petri net models, to simplify them, we have used bidirectional arcs between input places and transitions instead of arcs from input places to transitions and from transitions to input places. A bidirectional arc causes that the token is not consumed (removed) from the input place after firing a transition. This fact has a natural justification, i.e., firing a transition does not cause deactivation of the attractants and disappearance of the plasmodium from the original point. The plasmodium grows to build a dendritic network of veins.

In the proposed Petri net models of *Physarum* machines, we can distinguish three kinds of places:

- Places representing *Physarum polycephalum*.
- Places representing control stimuli (repellents).
- Places representing output stimuli (attractants).

In the *Physarum* language, the kind of a place is determined by the role played by it.

For each kind of places, we adopt different meaning (interpretation) of tokens. The meaning of tokens in places representing *Physarum polycephalum* is natural, i.e., the token in a given place corresponds to the presence of the plasmodium of *Physarum polycephalum* in an original point, where it starts to grow. The meaning of tokens in places representing control stimuli is shown in Table II, whereas the meaning of tokens in places representing output stimuli is shown in Table III. In case of control stimuli, we are interested in whether a given stimulus is activated or not. In case of output stimuli (attractants), we are interested in whether a given attractant is occupied by the plasmodium of *Physarum polycephalum*. Transitions in Petri net models represent the flow (propagation) of the plasmodium from the original points to attractants as well as between attractants.

In case of Petri net models, the main prototypes defined in the *Physarum* language and their selected methods are collected in Table IV.

Let us consider an exemplary Petri net shown in Figure 5. Formally, we have $MPN = (Pl, Tr, Ar, w, m)$, where:

- $Pl = \{P_1, P_2, P_3, P_4\}$,
- $Tr = \{T_1, T_2\}$,
- $Ar = Ar_O \cup Ar_I$, such that
  - $Ar_O = \{(P_1, T_1), (T_1, P_2), (P_2, T_2), (T_2, P_3), (T_2, P_4)\}$,
  - and $Ar_I = \emptyset$,
- $w(a) = 1$ for all $a \in Ar$,
- and $m(p) = 0$ for each $p \in Pl$,

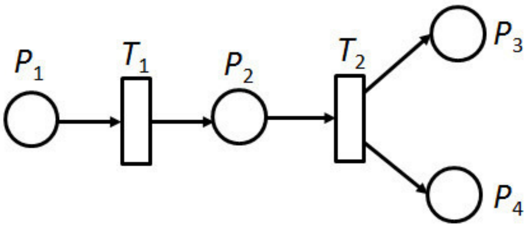
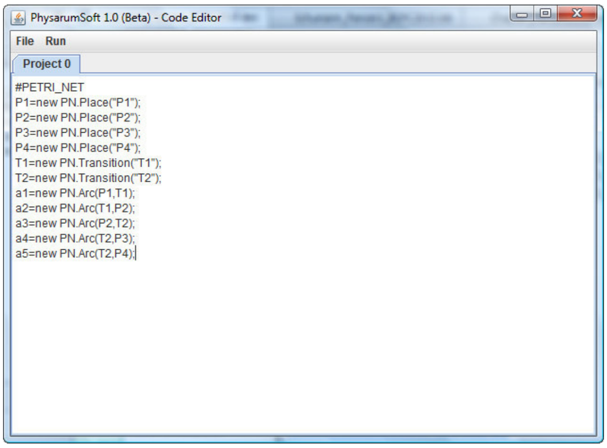


Fig. 5. An exemplary Petri net $MPN$

The code, for the model in the form of $MPN$, in the *Physarum* language written in the editor of the module of programming *Physarum* machines, is shown in Figure 6.

The result of compilation is shown in Figure 7. One can see that *Physarum* $Ph\_1$ represents place $P_1$, attractants $A\_1$, $A\_2$, $A\_3$ represent places $P_2$, $P_4$, $P_3$, respectively. Splitting the plasmodium at $A\_2$ is supported by repellent $R\_1$.

In [6], we showed how to use high-level models (transition systems and Petri nets) to describe four basic forms of *Physarum* motions:



```
#PETRI_NET
P1=new PN.Place("P1");
P2=new PN.Place("P2");
P3=new PN.Place("P3");
P4=new PN.Place("P4");
T1=new PN.Transition("T1");
T2=new PN.Transition("T2");
a1=new PN.Arc(P1,T1);
a2=new PN.Arc(T1,P2);
a3=new PN.Arc(P2,T2);
a4=new PN.Arc(T2,P3);
a5=new PN.Arc(T2,P4);
```

Fig. 6. The code for the model in the form of $MPN$

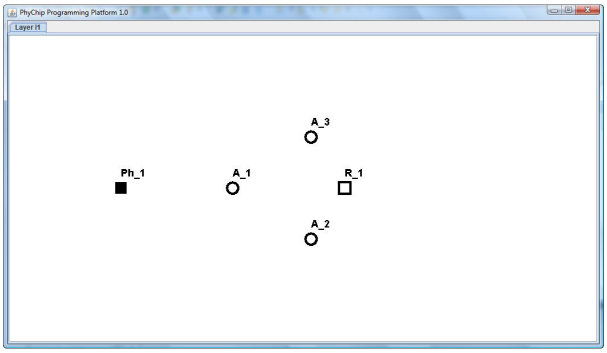


Fig. 7. The code for the model in the form of $MPN$

- *direct* - direction, i.e., a movement from one point, where the plasmodium is located, towards another point, where there is a neighbouring attractant,
- *fuse* - fusion of two plasmodia at the point, where they meet the same attractant,
- *split* - splitting the plasmodium from one active point into two active points, where two neighbouring attractants with a similar power of intensity are located,
- *repel* - repelling of the plasmodium or inaction.

It is worth noting that four basic forms are fundamental components used to build or describe more complex systems.

## III. Simulating Physarum Games

In [2], we showed that the slime mould (*Physarum polycephalum*) is a natural transition system which can be considered a biological model for strategic games. General assumptions for such games were presented both in [2] and [9]. They are based on the experiments performed by A. Adamatsky and M. Grube. If there are only two agents of the plasmodium game, where the first agent is presented by a usual *Physarum polycephalum* plasmodium and the second agent by its modification, a *Badhamia utricularis* plasmodium, then both start to compete with each other.

To simulate games on *Physarum* machines, we are developing a special module of *PhysarumSoft* called the *Physarum* game simulator. This module works under the client-server paradigm. A general structure of the *Physarum* game simulator is shown in Figure 8.

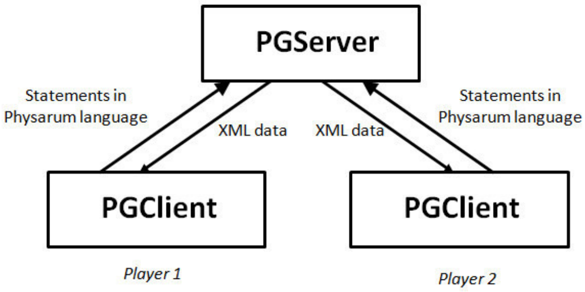
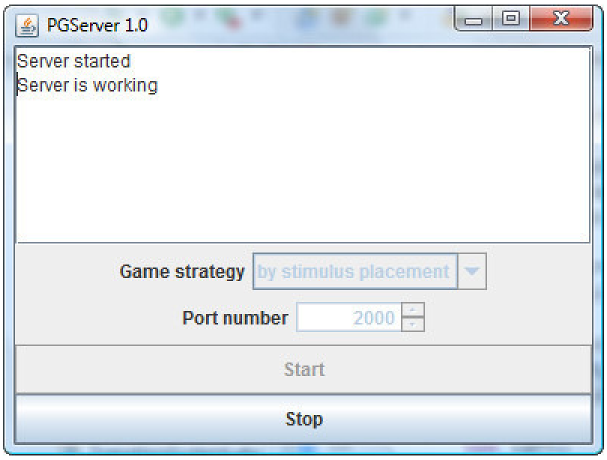

PGServer
Statements in Physarum language
XML data
XML data
Statements in Physarum language
PGClient
PGClient
Player 1
Player 2

Fig. 8. A general structure of the *Physarum* game simulator

The server-side application of the *Physarum* game simulator is called *PGServer*. The main window of *PGServer* is shown in Figure 9. In this window, the user can:

- select the port number on which the server listens for connections,
- start and stop the server,
- set the game strategy:
  - strategy by stimulus placement,
  - strategy by stimulus activation,
- shadow information about actions undertaken.



```
Server started
Server is working
```

Game strategy: by stimulus placement
Port number: 2000

Fig. 9. The main window of *PGServer*

The client-side application of the *Physarum* game simulator is called *PGClient*. The main window of *PGClient* is shown in Figures 10 and 12. In this window, the user can:

- set the server IP address and its port number,
- start the participation in the game,
- manipulate stimuli (place or activate them) during the game,
- monitor the current result.

In the *Physarum* game simulator, we have two players:

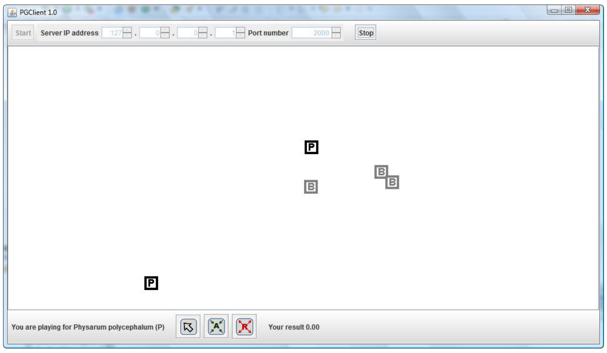
- the first one plays for the *Physarum polycephalum* plasmodia,

- the second one plays for the *Badhamia utricularis* plasmodia.

Locations of the original points of both plasmodia are randomly generated. The players can control motions of plasmodia via attracting or repelling stimuli. There are two strategies which can be defined for the game:

1) Locations of attractants and repellents are *a priori* generated in a random way. During the game, each player can activate one stimulus (attractant or repellent) at each step.
2) Locations of attractants and repellents are determined by the players during the game. At each step, each player can put one stimulus (attractant or repellent) at any location and this stimulus becomes automatically activated.

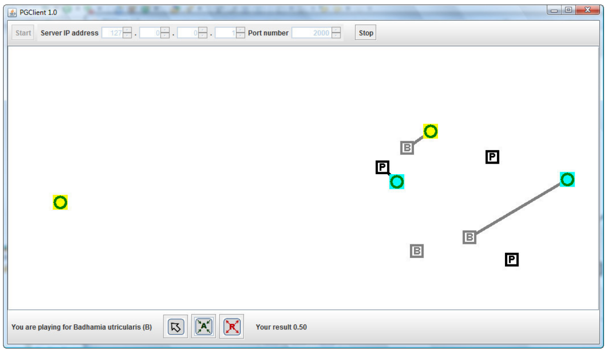
The client-side main window for the first strategy (locations of attractants and repellents are *a priori* generated in a random way) is shown in Figure 10. At the beginning, the original points of *Physarum polycephalum* and *Badhamia utricularis*, as well as stimuli, are scattered randomly on the plane. The window after several player's movements is shown in Figure 11. A box labelled by $P$ represents an original point of *Physarum polycephalum*. A box labelled by $B$ represents an original point of *Badhamia utricularis*. A single circle denotes an attractant whereas a double circle - repellent. Different background colors of stimuli differentiate between players.



You are playing for Physarum polycephalum (P)

Your result 0.00

Fig. 10. The main window of *PGClient* for the first strategy

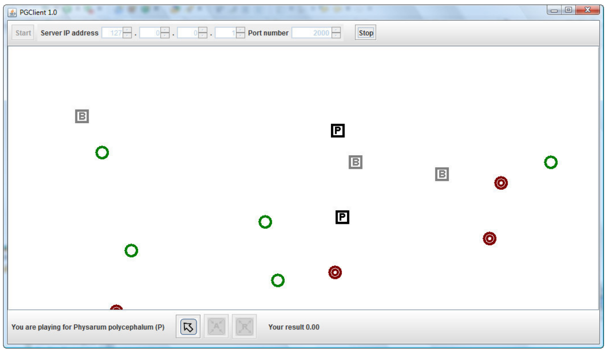


You are playing for Badhamia utricularis (B)

Your result 0.50

Fig. 11. The main window of *PGClient* for the first strategy after several player's movements

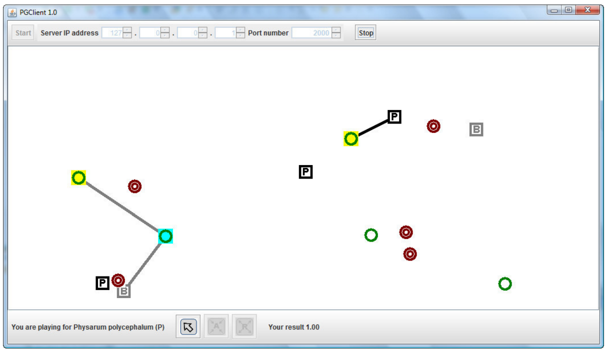
The client-side main window for the second strategy (locations of attractants and repellents are determined by the players during the game) is shown in Figure 12. At the beginning, the original points of *Physarum polycephalum* and *Badhamia utricularis* are scattered randomly on the plane. During the game, players can place stimuli. New veins of plasmodia are created. The window after several player's movements is shown in Figure 13.



You are playing for Physarum polycephalum (P)

Your result 0.00

Fig. 12. The main window of *PGClient* for the second strategy



You are playing for Physarum polycephalum (P)

Your result 1.00

Fig. 13. The main window of *PGClient* for the second strategy after several player's movements

Communication between clients and the server is realized through text messages containing statements of the *Physarum* language. The exemplary code responsible for creation of stimuli has the form:

```
p1_a1=new Attractant(195,224,1);
p1_a2=new Attractant(541,310,1);
p1_a1=new Attractant(580,92,2);
p2_r1=new Repellent(452,130,2);
p2_r1=new Repellent(659,327,1);
```

The first two parameters of stimulus constructors determine the location whereas the last parameter is the player's ID.

The server sends to clients information about the current configuration of the *Physarum* machine (localization of the original points of *Physarum polycephalum* and *Badhamia utricularis*, localization of stimuli, as well as a list of edges, corresponding to veins of plasmodia, between active points) through the XML file. The exemplary XML file has the form:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<network>
<elements>
```

```
<element id="0" player="0" type="0" x="96" y="310"/>
<element id="1" player="0" type="0" x="766" y="178"/>
<element id="2" player="0" type="0" x="566" y="248"/>
<element id="3" player="0" type="3" x="550" y="53"/>
<element id="4" player="0" type="3" x="374" y="534"/>
<element id="5" player="0" type="3" x="746" y="217"/>
<element id="6" player="1" type="1" x="195" y="224"/>
<element id="7" player="1" type="1" x="541" y="310"/>
<element id="8" player="2" type="1" x="580" y="92"/>
<element id="9" player="2" type="2" x="452" y="130"/>
<element id="10" player="1" type="2" x="659" y="327"/>
</elements>
<veins>
<vein createdBy="0" firstNodeID="2" secondNodeID="7"/>
<vein createdBy="3" firstNodeID="3" secondNodeID="8"/>
</veins>
</network>
```

The attribute "player" equal to 0 means that elements are created by the system, in this case, the original points of plasmodia (*Physarum polycephalum* or *Badhamia utricularis*).

As payoffs for the created bio-inspired games on *Physarum* machines, we may define a variety of tasks, including simple ones like achieving as many attractants as possible, occupied by plasmodia of organisms for which we play or constructing the longest path consisting of attractants occupied by plasmodia. Determining different payoffs for *Physarum* games appears to be an interesting field of research due to a huge number of different methodologies and paradigms which can be applied.

The activated attractant $A^*$ causes that the plasmodia propagate protoplasmic veins towards it and feed on it. It means that new transitions are created between the current active points of plasmodia and a new one on the attractant $A^*$. Propagating protoplasmic veins is possible if the current active points are located in the region of influence (ROI) of $A^*$. It means that a proper neighborhood of $A^*$ is taken into consideration. From that moment, the activated attractant $A^*$ is occupied by plasmodia. It is worth noting that, as the experiments showed, the attractant occupied by the plasmodium of *Physarum polycephalum* cannot be simultaneously occupied by the plasmodium of *Badhamia utricularis* and vice versa. Moreover, the *Physarum polycephalum* plasmodium grows faster and could grow into branches of *Badhamia utricularis*, while the *Badhamia utricularis* plasmodium could grow over *Physarum polycephalum* veins.

The activated repellent $R^*$ can change the direction of plasmodium motions or can avoid propagating plasmodium protoplasmic veins towards activated attractants. Such influences are possible if plasmodia are in the region of influence (ROI) of $R^*$.

The control capabilities presented above enable the players to choose, at each step, one of the possible tactics:

1) The attractant or repellent activated by the player can help propagation of his/her plasmodia (of either *Physarum polycephalum* or *Badhamia utricularis*).
2) The attractant or repellent activated by the player can disturb propagation of the second player's plasmodia.

The second possibility is worth considering if we adopt the payoff approximations based on the rough set model described in [19]. It will be the main direction of further investigations. During the game, the players can switch between two possible tactics according to the current game configuration. At the end of the game, we determine who wins.

## IV. Conclusions and Further Work

In the paper, we have described selected functionality of the current version of a new software tool called *PhysarumSoft*. This tool is intended for programming *Physarum* machines and simulating *Physarum* games. For over the last two years, we have designed a new object-oriented programming language, called the *Physarum* language, that constitutes the basis for modelling behaviour of *Physarum* maachines. This language is used in *PhysarumSoft*. In the nearest future, we plan to extend *PhysarumSoft* to other high-level models, e.g., $\pi$-calculus and cellular automata. Moreover, we are developing a new model, based on rough sets [20], to approximate payoffs of strategy games created on *Physarum* machines.

## Acknowledgment

This research is being fulfilled by the support of FP7-ICT-2011-8.

## References

[1] A. Adamatzky, *Physarum Machines: Computers from Slime Mould.* World Scientific, 2010.

[2] A. Schumann, K. Pancerz, A. Adamatzky, and M. Grube, "Bio-inspired game theory: The case of Physarum polycephalum," in *Proceedings of the 8th International Conference on Bio-inspired Information and Communications Technologies (BICT'2014)*, Boston, Massachusetts, USA, 2014. doi: 10.4108/icst.bict.2014.257869

[3] T. Nakagaki, H. Yamada, and A. Toth, "Maze-solving by an amoeboid organism," *Nature*, vol. 407, pp. 470–470, 2000. doi: 10.1038/35035159

[4] A. Adamatzky, V. Erokhin, M. Grube, T. Schubert, and A. Schumann, "Physarum Chip Project: Growing computers from slime mould," *International Journal of Unconventional Computing*, vol. 8, no. 4, pp. 319–323, 2012.

[5] K. Pancerz and A. Schumann, "Principles of an object-oriented programming language for Physarum polycephalum computing," in *Proceedings of the 10th International Conference on Digital Technologies (DT'2014)*, Zilina, Slovak Republic, 2014. doi: 10.1109/DT.2014.6868725 pp. 273–280.

[6] ——, "Some issues on an object-oriented programming language for Physarum machines," in *Applications of Computational Intelligence in Biomedical Technology*, ser. Studies in Computational Intelligence, R. Bris, J. Majernik, K. Pancerz, and E. Zaitseva, Eds. Springer International Publishing, Switzerland, 2016, vol. 606, pp. 185–199.

[7] A. Schumann and K. Pancerz, "Towards an object-oriented programming language for Physarum polycephalum computing," in *Proceedings of the Workshop on Concurrency, Specification and Programming (CS&P'2013)*, M. Szczuka, L. Czaja, and M. Kacprzak, Eds., Warsaw, Poland, 2013, pp. 389–397.

[8] I. Craig, *Object-Oriented Programming Languages: Interpretation.* London: Springer-Verlag, 2007.

[9] A. Schumann and K. Pancerz, "Interfaces in a game-theoretic setting for controlling the plasmodium motions," in *Proceedings of the 8th International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS'2015)*, Lisbon, Portugal, 2015. doi: 10.5220/0005285203380343 pp. 338–343.

[10] JavaCC, http://java.net/projects/javacc/.

[11] A. Schumann, K. Pancerz, and J. Jones, "Towards logic circuits based on physarum polycephalum machines: The ladder diagram approach," in *Proceedings of the International Conference on Biomedical Electronics and Devices (BIODEVICES'2014)*, A. Cliquet Jr., G. Plantier, T. Schultz, A. Fred, and H. Gamboa, Eds., Angers, France, 2014. doi: 10.5220/0004839301650170 pp. 165–170.

[12] A. Schumann and K. Pancerz, “Timed transition system models for programming Physarum machines: Extended abstract,” in *Proceedings of the Workshop on Concurrency, Specification and Programming (CS&P'2014)*, L. Popova-Zeugmann, Ed., Chemnitz, Germany, 2014, pp. 180–183.

[13] ——, “Towards an object-oriented programming language for Physarum polycephalum computing: A Petri net model approach,” *Fundamenta Informaticae*, vol. 133, no. 2-3, pp. 271–285, 2014. doi: 10.3233/FI-2014-1076

[14] M. Nielsen, G. Rozenberg, and P. Thiagarajan, “Elementary transition systems,” *Theoretical Computer Science*, vol. 96, no. 1, pp. 3–33, 1992. doi: 10.1016/0304-3975(92)90180-N

[15] T. A. Henzinger, Z. Manna, and A. Pnueli, “Timed transition systems,” in *Real-Time: Theory in Practice*, ser. Lecture Notes in Computer Science, J. de Bakker, C. Huizing, W. de Roever, and G. Rozenberg, Eds. Berlin Heidelberg: Springer, 1992, vol. 600, pp. 226–251.

[16] C. A. Petri, “Kommunikation mit automaten,” Institut für Instrumentelle Mathematik, Bonn, Schriften des IIM Nr. 2, 1962.

[17] T. Agerwala and M. Flynn, “Comments on capabilities, limitations and 'correctness' of Petri nets,” in *Proceedings of the 1st Annual Symposium on Computer Architecture (ISCA'1973)*, Atlanta, USA, 1973, pp. 81–86.

[18] H. Verbeek, M. Wynn, W. van der Aalst, and A. ter Hofstede, “Reduction rules for reset/inhibitor nets,” *Journal of Computer and System Sciences*, vol. 76, no. 2, pp. 125–143, 2010. doi: 10.1016/j.jcss.2009.06.003

[19] K. Pancerz and A. Schumann, “Rough set models of Physarum machines,” *International Journal of General Systems*, vol. 44, no. 3, pp. 314–325, 2015. doi: 10.1080/03081079.2014.997529

[20] Z. Pawlak, *Rough Sets. Theoretical Aspects of Reasoning about Data*. Dordrecht: Kluwer Academic Publishers, 1991.

DOI: 10.15439/2015F236
ACSIS, Vol. 5

# Go Games on Plasmodia of *Physarum Polycephalum*

Andrew Schumann
University of Information
Technology and Management in
Rzeszow,
ul. Sucharskiego 2, 35-225
Rzeszow, Poland
Email:
andrew.schumann@gmail.com

*Abstract*—**We simulate the motions of *Physarum polycephalum* plasmodium by the game of Go, the board game originated in ancient China more than 2,500 years ago. Then we concentrate just on Go games, where locations of black and white stones simulate syllogistic reasoning, in particular reasoning of Aristotelian syllogistic and reasoning of performative syllogistic. For the first kind of reasoning we need a special form of coalition games. For the second kind of reasoning we appeal to usual antagonistic games.**

## I. Introduction

IN the *Physarum Chip Project: Growing Computers From Slime Mould* [1] we are working on designing a biological computer, where logic circuits are represented by programmable behaviors of *Physarum polycephalum* plasmodium, the one-cell organism that behaves according to different chemical stimuli called attractants and repellents and propagates networks connecting all reachable food attractants [2], [12]. The behavior of plasmodia is very sensitive and intelligent [4], [5], [6], [7], [8], [15], [17], [18], [19]. This behavior can be represented as a bio-inspired game theory on plasmodia [14], i.e. an experimental game theory, where, on the one hand, all basic definitions are verified in the experiments with *Physarum polycephalum* and *Badhamia utricularis* and, on the other hand, all basic algorithms are implemented in the object-oriented language for simulations of plasmodia [13].

We show that the slime mold can be a model for concurrent games and context-based games defined in [11]. In context-based games, players can move concurrently as well as in concurrent games, but the set of actions is ever infinite. In our experiments, we follow the following interpretations of basic entities: (1) attractants as payoffs; (2) attractants occupied by the plasmodium as states of the game; (3) active zones of plasmodium as players; (4) logic gates for behaviors as moves (available actions) for the players; (5) propagation of the plasmodium as the transition table which associates, with a given set of states and a given move of the players, the set of states resulting from that move.

In this game theory we can demonstrate creativity of primitive biological substrates of plasmodia. The point is that plasmodia do not strictly follow spatial algorithms like Kolmogorov-Uspensky machines, but perform many additional actions. So, the plasmodium behavior can be formalized within strong extensions of spatial algorithms, e.g. within concurrent games or context-based games [11].

In this paper we show how we can represent the plasmodium behavior as a *Go game*. It is board game with two players (called Black and White) who alternately place black and white stones, accordingly, on the vacant intersections (called points) of a board with a 19x19 grid of lines. Black moves first. Stones are placed until they reach a point where stones of another color are located. There are the following two basic rules of the game: (1) each stone must have at least one open point (called liberty) directly next to it (up, down, left, or right), or must be part of a connected group that has at least one such open point; stones which lose their last liberty are removed from the board; (2) the stones must never repeat a previous position of stones. The aim of the game is in surrounding more empty points by player’s stones. At the end of game, the number of empty points player's stones surround are counted, together with the number of stones the player captured. This number determines who the winner is. The Go game originated in ancient China more than 2,500 years ago and it is very popular till now.

We can consider the game of Go as a model of plasmodium motions. In this view the black stones are considered attractants occupied by the plasmodium and the white stones are regarded as repellents. By this interpretation, we have two players, also: Black (this player places attractants) and White (this player places repellents). The winner is determined by the number of empty points player's stones surround.

Notice that the number of possible Go games is too large, $10^{761}$. Therefore it is better to focus just on games, where locations of black and white stones simulate spatial reasoning. In this paper we propose two logics in the universe of possible Go games: (1) Aristotelian syllogistic [3]; (2) performative syllogistic [9], [10].

978-83-60810-66-8/$25.00©2015, IEEE

## II. Classical Game of Go on Plasmodia of *Physarum Polycephalum*

The plasmodium of *Physarum polycephalum* moves to attractants to connect them and in the meanwhile it avoids places, where repellents are located. The radius, where chemical signals from attractants (repellents) can be detected by the plasmodium to attract (repel) the latter, determines the structure of natural *Voronoi cells*, where each Voronoi cell is a place, where a chemical signal holds (see Fig. 1).

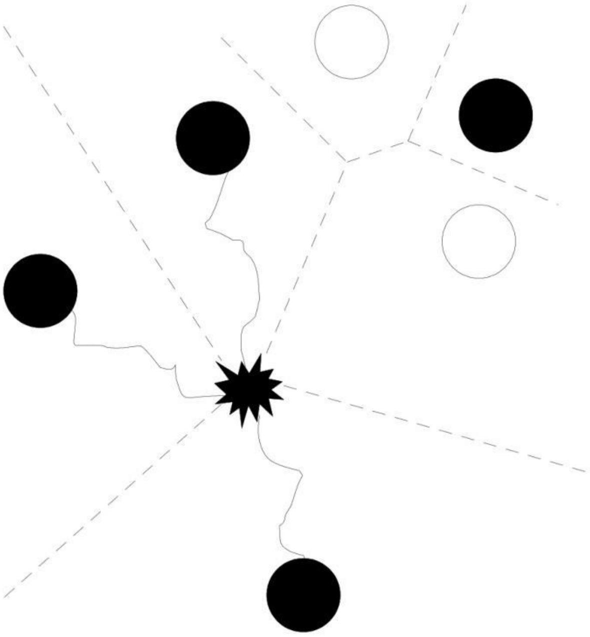


Fig. 1 The six Voronoi cells in accordance with the four attractants denoted by the black stones and with the two repellents denoted by the white stones. The plasmodium located in the center of the picture connects the three attractants by the three protoplasmic tubes. It cannot see the fourth attractant because of the two repellents

For the plasodium of Fig. 1 we have just the four neighbor cells. Notably that in Go games at each point we have only four neighbors everywhere. So, we can design the space for plasmodia in the way to have just four neighbors at each point. So, the cells of a Go game board are considered Voronoi cells with the same radius of intensity and power of attractants and repellents located in these cells. We associate black stones with attractants occupied by plasmodia and white stones with repellents. For the sake of convenience and more analogy with Voronoi cells, let us consider cells (not intersections of lines) as points for stone locations (see Fig. 2). Then we can use all rules of Go games to simulate plasmodium motions.

Thus, our Go game is represented in the universe of 18x18 Voronoi cells (see Fig. 2).

## III. Aristotelian Go Game on Plasmodia

Let us assume that the game of Go is a coalitional game with two players who choose only strategies to place black and white stones so that their locations can be interpreted as a spatial Aristotelian syllogistic reasoning.

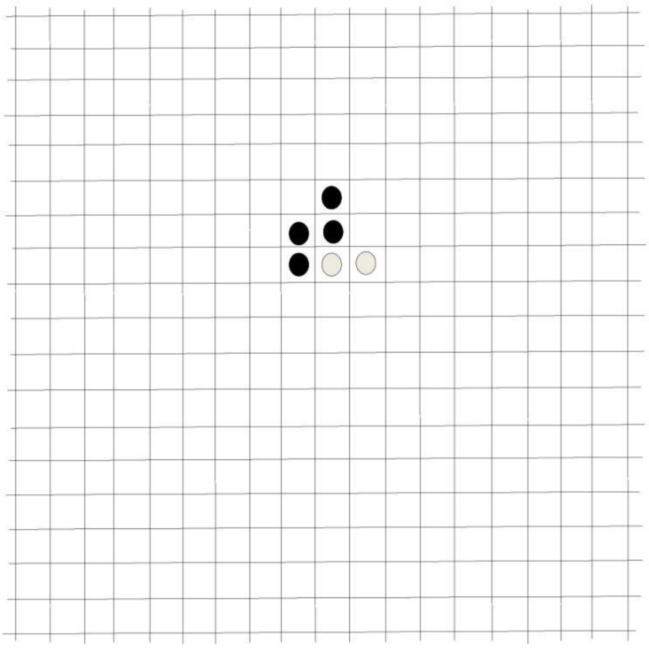


Fig. 2 The Go game board with two white stones designating repellents and four black stones designating attractants

Let us recall that the axiomatization of the Aristotelian syllogistic was laid first by Łukasiewicz [3]. In his axiomatization, the alphabet consists of the syllogistic letters *S*, *P*, *M*, …, the syllogistic connectives *a*, *e*, *i*, *o*, and the propositional connectives $\neg$, $\vee$, $\wedge$, $\Rightarrow$. Atomic propositions are defined as follows: *SxP*, where $x \in \{a, e, i, o\}$. All other propositions are defined in the following way: (i) each atomic proposition is a proposition, (ii) if *X*, *Y* are propositions, then ¬*X*, ¬*Y*, *X* å *Y*, where $å \in \{\vee, \wedge, \Rightarrow\}$, are propositions, also. The axioms proposed by Łukasiewicz are as follows:

$$SaP := (\exists A(AisS) \wedge \forall A(AisS \Rightarrow AisP)); \tag{1}$$

$$SiP := \exists A(AisS \wedge AisP); \tag{2}$$

$$SeP := \neg(SiP); \tag{3}$$

$$SoP := \neg(SaP); \tag{4}$$

$$SaS; \tag{5}$$

$$SiS; \tag{6}$$

$$(MaP \wedge SaM) \Rightarrow SaP; \tag{7}$$

$$(MaP \wedge MiS) \Rightarrow SiP. \tag{8}$$

In the Go implementation of Aristotelian syllogistic, the syllogistic letters *S*, *P*, *M*, … are interpreted as single cells of the board with the 18x18 Voronoi cells. The letter *S* is understood as empty if and only if the white stone is located on an appropriate cell denoted by *S*. Let us recall that in our plasmodium interpretation of Go games white stones mean ever repellents so that their location in a Voronoi cell means that this cell cannot be occupied by plasmodia. This letter *S* is treated as non-empty if and only if the black stone is located on an appropriate cell denoted by *S*. Recall that black stones mean ever attractants so that their location in a Voronoi cell means that this cell is occupied by plasmodia. If a cell does not contain any stone, this means that this cell is out of the game.

Hence, in the *Physarum* interpretation of this Go game, the non-empty syllogistic letters *S*, *P*, *M*, …, i.e. the cells denoted by *S*, *P*, *M*, … containing black stones are considered attractants and the empty syllogistic letters *S*, *P*, *M*, …, i.e. the cells denoted by *S*, *P*, *M*, … containing white stones are considered repellents. So, a data point *S* is regarded as non-empty if and only if an appropriate attractant located in *S* is occupied by plasmodium. This data point *S* is regarded as empty if and only if an appropriate repellent located in *S* repell plasmodium.

Thus, in the Aristotelian version of the Go game we have syllogistic strings of the form *SP* with the following interpretation: '*S* is *P*', and with the following meaning: *SP* is true if and only if *S* and *P* are neighbors and both *S* and *P* are not empty, otherwise *SP* is false. We can extend this meaning as follows: *SP* is true if and only if *S* and *P* are not empty and there is a line of non-empty cells between points *S* and *P*, otherwise *SP* is false. By the definition of true syllogistic strings, we can define atomic syllogistic propositions as follows:

*In the formal syllogistic*: there exists *A* such that *A* is *S* and for any *A*, if *A* is *S*, then *A* is *P*. *In the Go game model*: there is a cell *A* containing the black stone and for any *A*, if *AS* is true, then *AP* is true. *In the Physarum model*: there is a plasmodium in the cell *A* and for any *A*, if *AS* is true, then *AP* is true.

*In the formal syllogistic*: there exists *A* such that both *AS* is true and *AP* is true. *In the Go game model*: there exists a cell *A* containing the black stone such that *AS* is true and *AP* is true. *In the Physarum model*: there exists a plasmodium in the cell *A* such that *AS* is true and *AP* is true.

*In the formal syllogistic*: for all *A*, *AS* is false or *AP* is false. *In the Go game model*: for all cells *A* containing the black stones, *AS* is false or *AP* is false. *In the Physarum model*: for all plasmodia *A*, *AS* is false or *AP* is false.

*In the formal syllogistic*: for any *A*, *AS* is false or there exists *A* such that *AS* is true and *AP* is false. *In the Go game model*: for all cells *A* containing the black stones, *AS* is false or there exists *A* such that *AS* is true and *AP* is false. *In the Physarum model*: for any plasmodia *A*, *AS* is false or there exists *A* such that *AS* is true and *AP* is false.

Formally, this semantics is defined as follows. Let *M* be a set of attractants. Take a subset $|X| \subseteq M$ of cells containing the black stones (i.e. of cells containing attractants and occupied by the plasmodium) as a meaning for each syllogistic variable *X*. Next, define an ordering relation $\subseteq$ on subsets $|S|, |P| \subseteq M$ as: $|S| \subseteq |P|$ iff all attractants from $|P|$ are reachable for the plasmodium located at the attractants from $|S|$, i.e. iff for all cells of $|S|$ with black stones there are lines of black stones connecting them to cells of $|P|$ also containing black stones. Hence, $|S| \cap |P| \neq \varnothing$ means that some attractants from $|P|$ are reachable for the plasmodium located at the attractants from $|S|$ and $|S| \cap |P| = \varnothing$ means that no attractants from $|P|$ are reachable for the plasmodium located at the attractants from $|S|$. In the Go game model $|S| \cap |P| \neq \varnothing$ means that some cells from $|P|$ occupied by the black stones are connected by the lines of black stones with the cells from $|S|$ occupied by the black stones and $|S| \cap |P| = \varnothing$ means that there are no lines of black stones from the cells of $|P|$ to the cells of $|S|$.

This gives rise to models $\mathsf{M} = \langle M, |\cdot| \rangle$ such that

$$\mathsf{M} \models SaP \text{ iff } |S| \subseteq |P|;$$

$$\mathsf{M} \models SiP \text{ iff } |S| \cap |P| \neq \varnothing;$$

$$\mathsf{M} \models SeP \text{ iff } |S| \cap |P| = \varnothing;$$

$$\mathsf{M} \models p \wedge q \text{ iff } \mathsf{M} \models p \text{ and } \mathsf{M} \models q;$$

$$\mathsf{M} \models \quad p \vee q \text{ iff } \mathsf{M} \models p \text{ or } \mathsf{M} \models q;$$

$$\mathsf{M} \models \neg p \text{ iff it is false that } \mathsf{M} \models p.$$

*Proposition 1:* The Aristotelian syllogistic is sound and complete relatively to $\mathsf{M}$ if we understand $\subseteq$ as an inclusion relation (it is a well-known result [16]).

However, relatively to all possible Go games (plasmodium behaviors) the Aristotelian syllogistic is not complete. Indeed, the relation $\subseteq$ can have the following verification on the *Aristotelian Go game model on plasmodia* according to our definitions: $|S| \subseteq |P|$ and $|S| \subseteq |P'|$, where $|P| \cap |P'| = \varnothing$, i.e. all attractants from $|P|$ are reachable for the plasmodium located at the attractants from $|S|$ and all attractants from $|P'|$ are reachable for the plasmodium located at the attractants from $|S|$, but between $|P|$ and $|P'|$ there are no paths. In this case $\subseteq$ is not an inclusion relation and proposition 1 does not hold. Hence, we need repellents to make $\subseteq$ the inclusion relations in all cases. Therefore, to obtain the Aristotelian Go game model on plasmodia we shall deal with the coalitional game, where two players will cooperate to build spatial reasoning satisfying the rules of Aristotelian syllogistic. The first player places black stones to designate places of growing plasmodia. The second player places white stones to designate places of repelling plasmodia. So, both players follow coalitional strategies to simulate Aristotelian syllogistic reasoning.

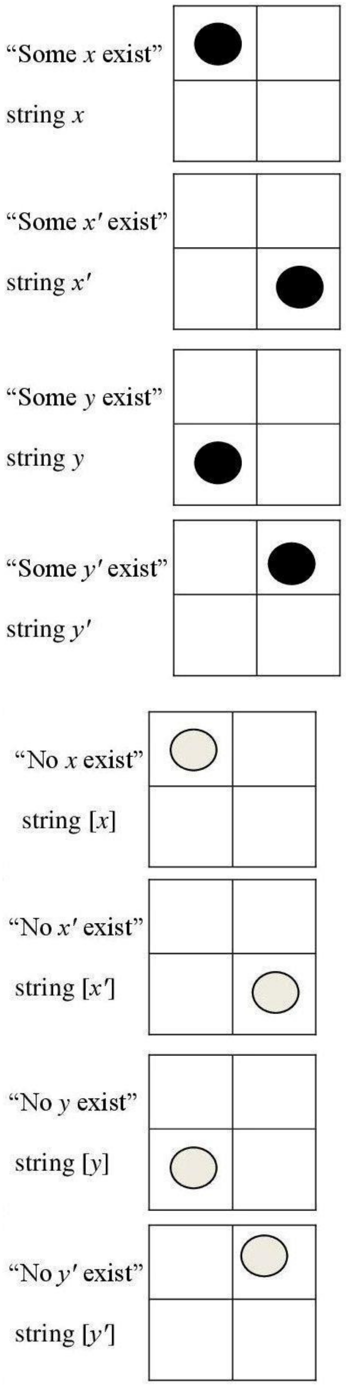


Fig. 3 The Aristotelian Go game diagrams for the basic existence strings

### A. *Coalition Go Game for Verifying Aristotelian Reasoning*

In the Aristotelian Go game model for verifying all the basic syllogistic propositions, we will use the following four cells: $x$, $y$, $x'$, $y'$ of the game board with the 19x19 grid of lines, where $x'$ means all cells which differ from $x$, but they are neighbors for $y$, and $y'$ means all cells which differ from $y$ and are neighbors for $x$. These cells express appropriate meanings of syllogistic letters. The corresponding universe of discourse will be denoted by means of the following diagram:

| | |
|---|---|
| $x$ | $y'$ |
| $y$ | $x'$ |

Assume that a black stone denotes an attractant and if it is placed within a cell $x$, this means that "this Voronoi cell contains an attractant $N_x$ activated and occupied by the plasmodium". It is a verification of the syllogistic letter $S_x$ at cell $x$ of the board. A white stone denotes a repellent and if it is placed within a cell $x$, this means that "this Voronoi cell contains a repellent $R_x$ activated and there is no plasmodium in it". It is a verification of a new syllogistic letter $[S_x]$. For the sake of convenience, we will denote $S_x$ by $x$ and $[S_x]$ by $[x]$. Using these stones, we can verify all the basic existence syllogistic propositions (see Fig. 3).

Aristotelian Go game strings of the form $xy$, $yx$ are interpreted as particular affirmative propositions "Some $x$ are $y$" and "Some $y$ are $x$" respectively, strings of the form $[xy]$, $[yx]$, $x[y]$, $y[x]$ are interpreted as universal negative propositions "No $x$ are $y$" and "No $y$ are $x$". A universal affirmative proposition "All $x$ are $y$" are presented by a complex string $xy \,\&\, x[y']$. The sign & means that we have strings $xy$ and $x[y']$ simultaneously and they are considered the one complex string. All these strings are verified on the basis of the diagrams of Fig. 4. So, we use only black stones for building particular propositions, only white stones for building universal negative propositions, and we combine black and white stones for building universal affirmative propositions. Consequently, we need a cooperation of two players to implement a spatial version of Aristotelian syllogistic within coalition Go games.

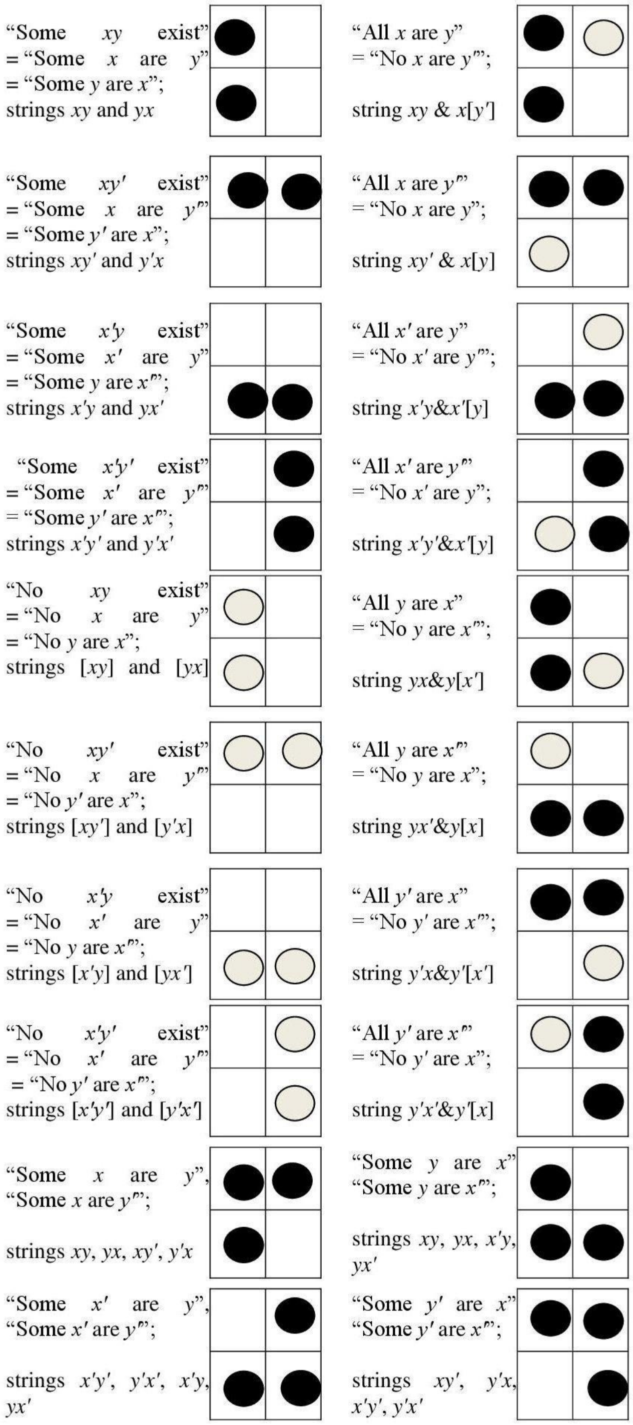
For verifying syllogisms we will use the following diagrams symbolizing some neighbor cells:

| | | | |
|---|---|---|---|
| | $m$ | $m'$ | |
| $m'$ | $x$ | $y'$ | $m$ |
| $m$ | $y$ | $x'$ | $m'$ |
| | $m'$ | $m$ | |

The motion of plasmodium starts from one of the central cells ($x$, $y$, $x'$, $y'$) and goes towards one of the four directions (northwest, southwest, northeast, southeast). The syllogism shows a connection between two not-neighbor cells on the basis of its joint neighbor and says if there was either multiplication or fusion of plasmodia (i.e. either splitting or fusion of the lines of black stones). As a syllogistic conclusion, we obtain another diagram:

| | |
|---|---|
| $x$ | $m'$ |
| $m$ | $x'$ |

Different syllogistic conclusions derived show directions of plasmodium's propagation. Some examples are provided in Fig. 5 – 7.


“Some $xy$ exist” = “Some $x$ are $y$” = “Some $y$ are $x$”; strings $xy$ and $yx$

“All $x$ are $y$” = “No $x$ are $y'$”; string $xy$ & $x[y']$

“Some $xy'$ exist” = “Some $x$ are $y'$” = “Some $y'$ are $x$”; strings $xy'$ and $y'x$

“All $x$ are $y'$” = “No $x$ are $y$”; string $xy'$ & $x[y]$

“Some $x'y$ exist” = “Some $x'$ are $y$” = “Some $y$ are $x'$”; strings $x'y$ and $yx'$

“All $x'$ are $y$” = “No $x'$ are $y'$”; string $x'y$&$x'[y]$

“Some $x'y'$ exist” = “Some $x'$ are $y'$” = “Some $y'$ are $x'$”; strings $x'y'$ and $y'x'$

“All $x'$ are $y'$” = “No $x'$ are $y$”; string $x'y'$&$x'[y]$

“No $xy$ exist” = “No $x$ are $y$” = “No $y$ are $x$”; strings $[xy]$ and $[yx]$

“All $y$ are $x$” = “No $y$ are $x'$”; string $yx$&$y[x']$

“No $xy'$ exist” = “No $x$ are $y'$” = “No $y'$ are $x$”; strings $[xy']$ and $[y'x]$

“All $y$ are $x'$” = “No $y$ are $x$”; string $yx'$&$y[x]$

“No $x'y$ exist” = “No $x'$ are $y$” = “No $y$ are $x'$”; strings $[x'y]$ and $[yx']$

“All $y'$ are $x$” = “No $y'$ are $x'$”; string $y'x$&$y'[x']$

“No $x'y'$ exist” = “No $x'$ are $y'$” = “No $y'$ are $x'$”; strings $[x'y']$ and $[y'x']$

“All $y'$ are $x'$” = “No $y'$ are $x$”; string $y'x'$&$y'[x]$

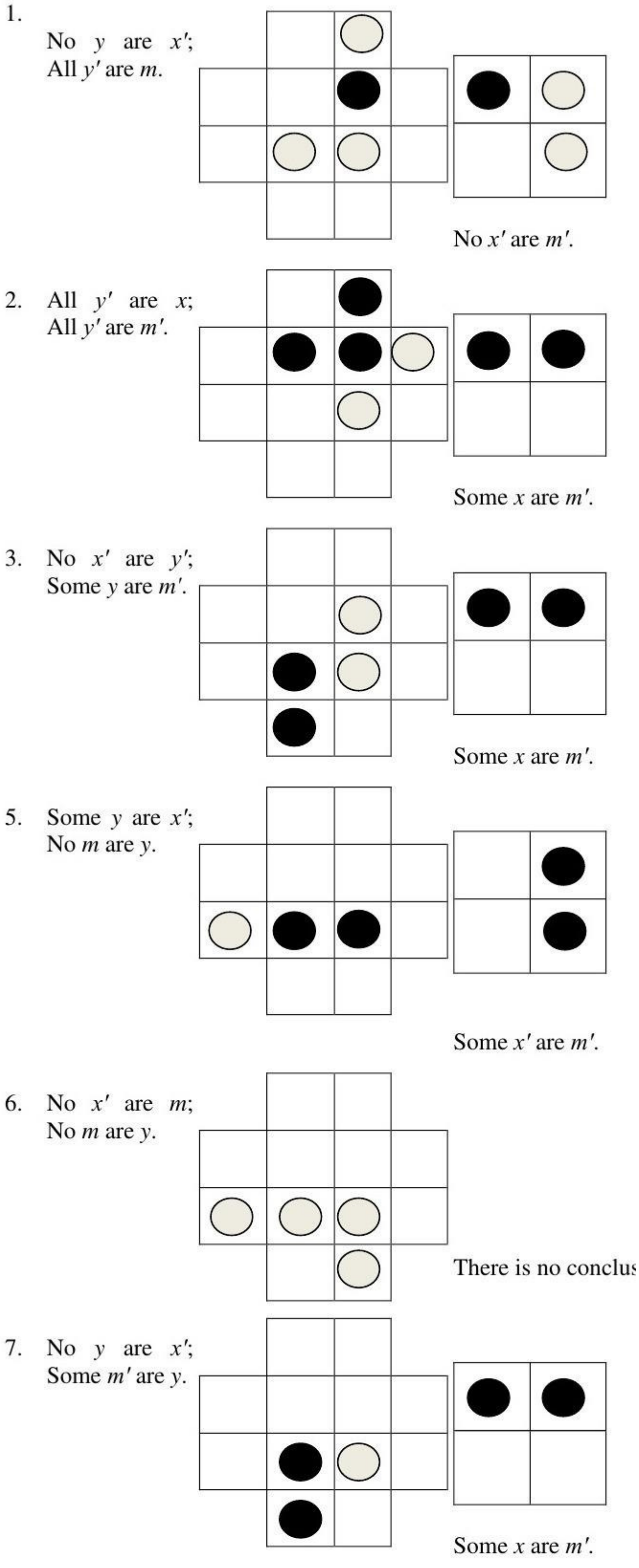
“Some $x$ are $y$”, “Some $x$ are $y'$”; strings $xy$, $yx$, $xy'$, $y'x$

“Some $y$ are $x$” “Some $y$ are $x'$”; strings $xy$, $yx$, $x'y$, $yx'$

“Some $x'$ are $y$”, “Some $x'$ are $y'$”; strings $x'y'$, $y'x'$, $x'y$, $yx'$

“Some $y'$ are $x$” “Some $y'$ are $x'$”; strings $xy'$, $y'x$, $x'y'$, $y'x'$

Fig. 4 The Aristotelian Go game diagrams for syllogistic propositions


1. No $y$ are $x'$; All $y'$ are $m$.
No $x'$ are $m'$.

2. All $y'$ are $x$; All $y'$ are $m'$.
Some $x$ are $m'$.

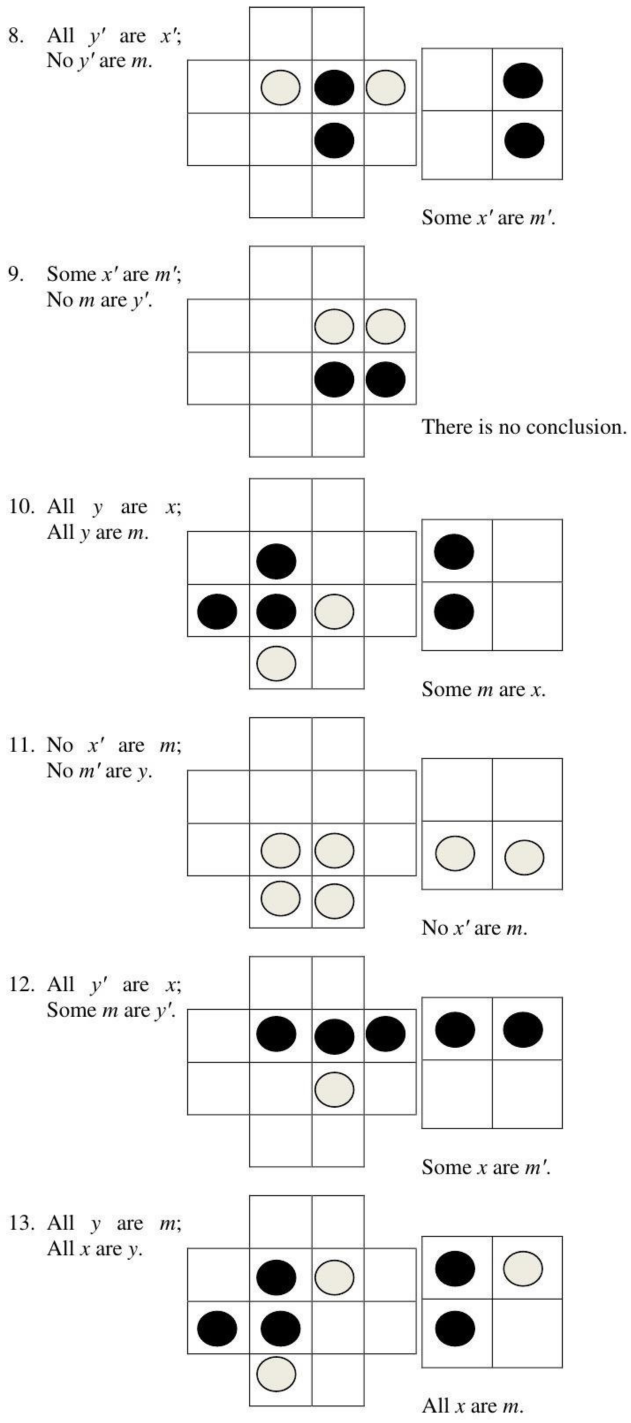
3. No $x'$ are $y'$; Some $y$ are $m'$.
Some $x$ are $m'$.

5. Some $y$ are $x'$; No $m$ are $y$.
Some $x'$ are $m'$.

6. No $x'$ are $m$; No $m$ are $y$.
There is no conclus

7. No $y$ are $x'$; Some $m'$ are $y$.
Some $x$ are $m'$.

Fig. 5 The Aristotelian Go game diagrams for syllogisms (part 1)



8. All $y'$ are $x'$;
No $y'$ are $m$.

Some $x'$ are $m'$.

9. Some $x'$ are $m'$;
No $m$ are $y'$.

There is no conclusion.

10. All $y$ are $x$;
All $y$ are $m$.

Some $m$ are $x$.

11. No $x'$ are $m$;
No $m'$ are $y$.

No $x'$ are $m$.

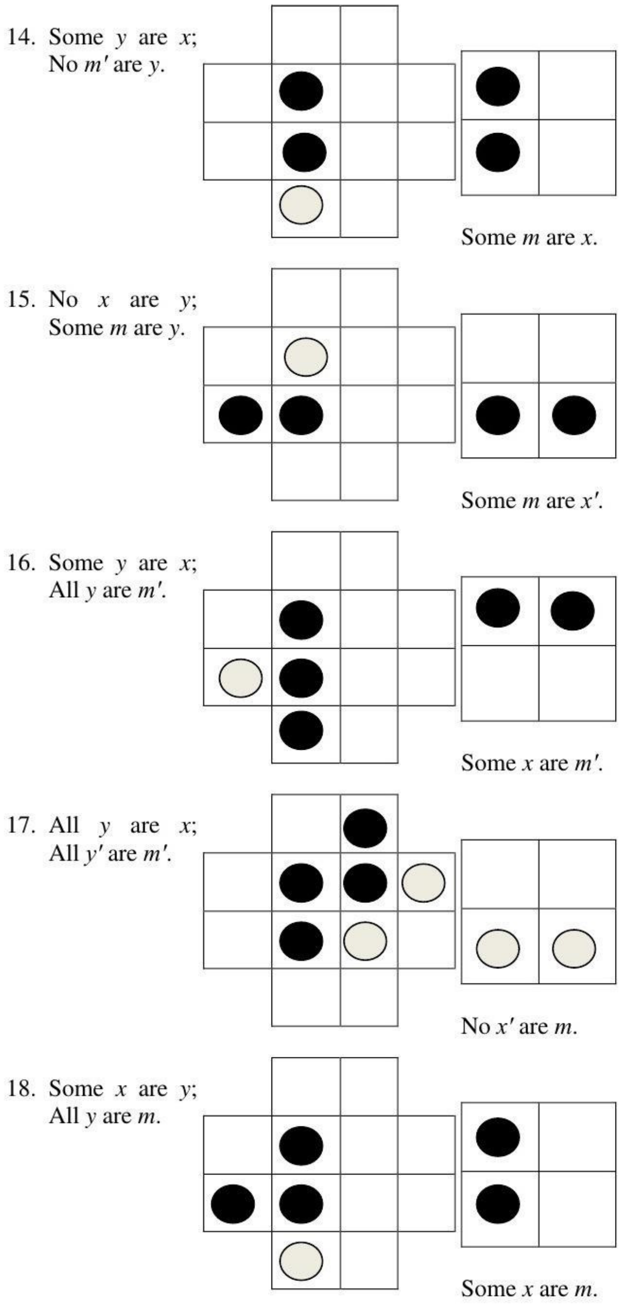
12. All $y'$ are $x$;
Some $m$ are $y'$.

Some $x$ are $m'$.

13. All $y$ are $m$;
All $x$ are $y$.

All $x$ are $m$.

Fig. 6 The Aristotelian Go game diagrams for syllogisms (part 2)



14. Some $y$ are $x$;
No $m'$ are $y$.

Some $m$ are $x$.

15. No $x$ are $y$;
Some $m$ are $y$.

Some $m$ are $x'$.

16. Some $y$ are $x$;
All $y$ are $m'$.

Some $x$ are $m'$.

17. All $y$ are $x$;
All $y'$ are $m'$.

No $x'$ are $m$.

18. Some $x$ are $y$;
All $y$ are $m$.

Some $x$ are $m$.

Fig. 7 The Aristotelian Go game diagrams for syllogisms (part 3)

Continuing in the same way, we can construct a syllogistic system, where conclusions are derived from three premises. The Aristotelian Go game (i.e. the suitable motion of plasmodium) starts from one of the central cells ($x$, $y$, $x'$, $y'$) and goes towards one of the four directions (northwest, southwest, northeast, southeast), then towards one of the eight directions (north-northwest, west-northwest, south-southwest, west-southwest, north-northeast, east-northeast, south-southeast, east-southeast), etc.

Hence, a Go game or spatial expansion of plasmodium is interpreted as a set of syllogistic propositions. The universal affirmative proposition $xy$ & $x[y']$ means that the plasmodium at the place $x$ goes only to $y$ and all other directions are excluded. The universal negative proposition $x[y]$ or $[xy]$ means that the plasmodium at the place $x$ cannot go to $y$ and we know nothing about other directions. The particular affirmative proposition $xy$ means that the plasmodium at the place $x$ goes to $y$ and we know nothing about other directions. Syllogistic conclusions allow us to mentally reduce the number of syllogistic propositions showing plasmodium's propagation.

For the implementation of Aristotelian syllogistic we appeal to repellents to delete some possibilities in the plasmodium propagation. So, model $\mathsf{M}$ defined above should be understood as follows:

$\mathsf{M} \models$ *All x are y* iff $xy$ & $x[y']$, i.e. the plasmodium is located at $x$ and can move only to $y$ and cannot move towards all other directions (the black stone is placed at $x$ and we can build the line of black stones only to $y$);

$\mathsf{M} \models$ *Some x are y* iff $xy$, i.e. the plasmodium is located at $x$ and can move to $y$ (the black stone is placed at $x$ and we can build the line of black stones to $y$);

$\mathsf{M} \models$ *No x are y* iff $x[y]$ or $[xy]$, i.e. the plasmodium cannot move to $y$ in any case (there is no line of black stones to $y$).

It is evident in this formulation that the Aristotelian syllogistic is so unnatural for plasmodia. Without repellents (the coalition game of two players), this syllogistic system cannot be verified in the medium of plasmodium propagations (Go game). In other words, we can prove the next proposition:

*Proposition 2:* The Aristotelian syllogistic is not sound and complete on the plasmodium without repellents. In other words, the Aristotelian syllogistic is not sound and complete in the Go game without a coalition of two players.

In other words, the Aristotelian syllogistic reasoning can be implemented as a Go game if and only if two players agree to play cooperatively to place black and white stones in accordance with spatial implementation of syllogisms.

### B. Examples of Aristotelian Go Game

Let us consider a game of Go at time step 10, i.e. when White and Black players have placed the 10 white stones and the 10 black stones respectively. Let this game be pictured in Figure 8. Each Voronoi cell is denoted from $S_{1,1}$ to $S_{18,18}$. So, in Figure 8 syllogistic letters $S_{6,4}$, $S_{7,5}$, $S_{7,6}$, $S_{8,7}$, $S_{8,8}$, $S_{7,9}$, $S_{8,10}$, $S_{6,11}$, $S_{4,9}$, $S_{4,10}$ are understood as non-empty and syllogistic letters $S_{4,6}$, $S_{5,6}$, $S_{6,8}$, $S_{6,9}$, $S_{6,10}$, $S_{4,11}$, $S_{7,10}$, $S_{7,12}$, $S_{11,8}$, $S_{12,8}$ as empty. As a result, we can build some true syllogistic propositions in this universe like that: 'Some $S_{7,5}$ are $S_{7,6}$', 'Some $S_{8,7}$ are $S_{8,8}$', 'Some $S_{4,9}$ are $S_{4,10}$', 'No $S_{4,6}$ are $S_{5,6}$', 'No $S_{6,8}$ are $S_{6,9}$', 'No $S_{6,9}$ are $S_{6,10}$', 'No $S_{6,10}$ are $S_{7,10}$', 'No $S_{11,8}$ are $S_{12,8}$', etc.

Let us notice that in the universe of Fig. 8 we do not have universal affirmative propositions. But we can draw some syllogistic conclusions such as 'If $S_{4,10}[S_{4,11}]$ and $S_{4,9}$ $S_{4,10}$, then $S_{4,9}[S_{4,11}]$' (i.e. 'If no $S_{4,10}$ are $S_{4,11}$ and some $S_{4,9}$ are $S_{4,10}$, then no $S_{4,9}$ are $S_{4,11}$').

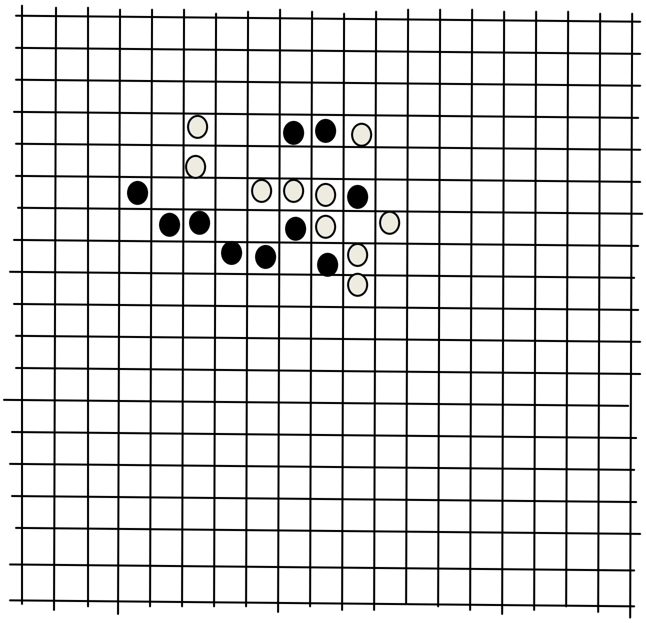


Fig. 8 The Aristotelian Go game 1 at time step 10

On the contrary, in the universe pictured in the Go game of Fig. 9 we have universal affirmative propositions such as 'All $S_{8,10}$ are $S_{8,11}$' and 'All $S_{4,9}$ are $S_{3,9}$'. Some possible conclusions: 'If no $S_{8,11}$ are $S_{8,12}$ and all $S_{8,10}$ are $S_{8,11}$, then no $S_{8,10}$ are $S_{8,12}$' and 'If no $S_{3,9}$ are $S_{3,10}$ and all $S_{4,9}$ are $S_{3,9}$, then no $S_{4,9}$ are $S_{3,10}$'.

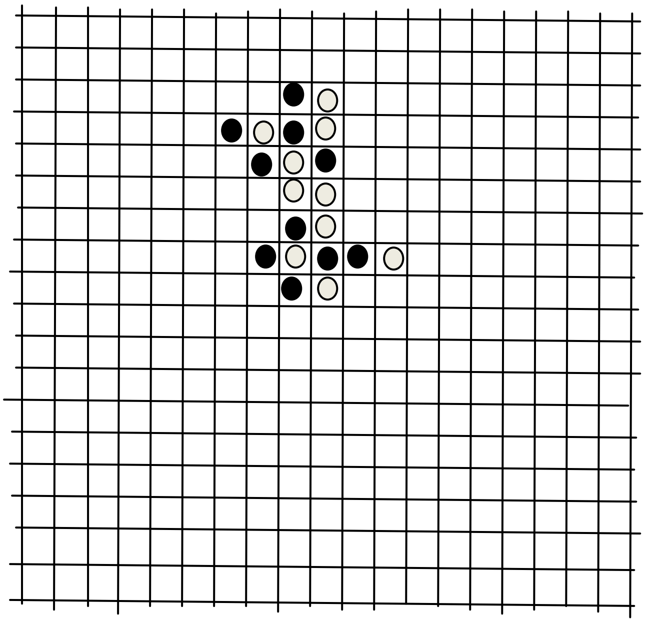


Fig. 9 The Aristotelian Go game 2 at time step 10

## IV. Non-Aristotelian Syllogistic Go Game on Plasmodia

While in Aristotelian syllogisms we are concentrating on one direction of many *Physarum* motions, and dealing with acyclic directed graphs with fusions of many protoplasmic tubes toward one data point, in most cases of *Physarum* behavior, not limited by repellents, we observe a spatial expansion of *Physarum* protoplasm in all directions with many cycles. Under these circumstances it is more natural to define all the basic syllogistic propositions $SaP$, $SiP$, $SeP$, $SoP$ in a way they satisfies the inverse relationship when all converses are valid: $SaP \Rightarrow PaS$, $SiP \Rightarrow PiS$, $SeP \Rightarrow PeS$, $SoP \Rightarrow PoS$. In other words, we can draw more natural conclusions for protoplasmic tubes which are decentralized and have some cycles. The formal syllogistic system over propositions with such properties is constructed in [9], [10]. This system is called the *performative syllogistic*. The alphabet of this system contains as descriptive signs the syllogistic letters $S$, $P$, $M$, …, as logical-semantic signs the syllogistic connectives $a$, $e$, $i$, $o$, and the propositional connectives $\neg$, $\vee$, $\wedge$, $\Rightarrow$. Atomic propositions are defined as follows: $SxP$, where $x \in \{a, e, i, o\}$. All other propositions are defined thus: (i) each atomic proposition is a proposition, (ii) if $X$, $Y$ are propositions, then $\neg X$, $\neg Y$, $X$å $Y$, where å $\in \{\vee, \wedge, \Rightarrow\}$, are propositions, too.

Let us consider Go games with the two different kinds of plasmodia: (i) plasmodia of *Physarum polycephalum* and (ii) plasmodia of *Badhamia utricularis* [14]. They try to occupy free attractants antagonistically. So, if an attractant is occupied by the plasmodium of *Physarum polycephalum*, it cannot be occupied by the plasmodium of *Badhamia utricularis* and if it is occupied by the plasmodium of *Badhamia utricularis*, it cannot be occupied by the plasmodium of *Physarum polycephalum*. In this way we observe a competition between two plasmodia.

In order to implement the performative syllogistic in the Go games with *Physarum polycephalum* and *Badhamia utricularis* plasmodia, we will interpret data points denoted by appropriate syllogistic letters as black stones (attractants) if we assume that appropriate cells are occupied by the plasmodium of *Physarum polycephalum* and we will interpret data points denoted by appropriate syllogistic letters as white stones (attractants) if we assume that appropriate cells are occupied by the plasmodium of *Badhamia utricularis*. A data point $S$ is considered empty for the Black player if and only if an appropriate attractant denoted by $S$ is occupied by the white stone (plasmodium of *Badhamia utricularis*). A data point $S$ is considered empty for the White player if and only if an appropriate attractant denoted by $S$ is occupied by the black stone (plasmodium of *Physarum polycephalum*). Let us define syllogistic strings of the form $SP$ with the following interpretation: (i) '$S$ is $P$': $SP$ is true for the Black player if and only if $S$ and $P$ are reachable for each other by the plasmodium of *Physarum polycephalum* and both $S$ and $P$ are not empty for the Black player, otherwise $SP$ is false; (ii) '$S$ is $P$': $SP$ is true for the White player if and only if $S$ and $P$ are reachable for each other by the plasmodium of *Badhamia utricularis* and both $S$ and $P$ are not empty for the White player, otherwise $SP$ is false. In other words, $SP$ is true for the Black player (respectively, for the White player) if and only if $S$ and $P$ are not empty for the Black player (respectively, for the White player) and there is a line of non-empty cells for the Black player (respectively, for the White player) between points $S$ and $P$, otherwise $SP$ is false. Using this definition of syllogistic strings, we can define atomic syllogistic propositions as follows:

*In the formal performative syllogistic*: there exists $A$ such that $A$ is $S$ and for any $A$, $AS$ is true and $AP$ is true. *In the Go game model:* there is a black (white) stone in $A$ connected by black (white) stones to $S$ and connected by black (white) stones to $P$. *In the Physarum model*: a plasmodium of *Physarum polycephalum* (a plasmodium of *Badhamia utricularis*) in $A$ occupies $S$ and for any plasmodia $A$ of *Physarum polycephalum* (for any plasmodia $A$ of *Badhamia utricularis*) which is a neighbor for $S$ and $P$, there are strings $AS$ and $AP$. This means that we have a massive-parallel occupation of the region by plasmodia of *Physarum polycephalum* (plasmodia of *Badhamia utricularis*) where the cells $S$ and $P$ are located.

*In the formal performative syllogistic*: for any $A$, both $AS$ is false and $AP$ is false. *In the Go game model*: for any cell $A$ there are no lines of black (white) stones connecting $A$ to $S$ and $A$ to $P$. *In the Physarum model*: for any plasmodium of *Physarum polycephalum* (of *Badhamia utricularis*) $A$ which is a neighbor for $S$ and $P$, there are no strings $AS$ and $AP$. This means that the plasmodium of *Physarum polycephalum* (of *Badhamia utricularis*) cannot reach $S$ from $P$ or $P$ from $S$ immediately.

*In the formal performative syllogistic*: there exists $A$ such that if $AS$ is false, then $AP$ is true. *In the Go game model*: there exists a cell $A$ with the black (white) stone which is a neighbor for cells $S$ and $P$ such that there is a string $AS$ or there is a string $AP$. *In the Physarum model*: there exists the plasmodium of *Physarum polycephalum* (of *Badhamia utricularis*) $A$ which is a neighbor for $S$ and $P$ such that there is a string $AS$ or there is a string $AP$. This means that the plasmodium of *Physarum polycephalum* (of *Badhamia utricularis*) occupies $S$ or $P$, but not the whole region where the cells $S$ and $P$ are located.

*In the formal performative syllogistic*: for any $A$, $AS$ is false or there exists $A$ such that $AS$ is false or $AP$ is false. *In the Go game model*: for any cell $A$ with the black (white) stone which is a neighbor for $S$ and $P$ there is no string $AS$ or there exists a black (white) stone in $A$ which is a neighbor for $S$ and $P$ such that there is no string $AS$ or there is no string $AP$. *In the Physarum model*: for any plasmodium of *Physarum polycephalum* (of *Badhamia utricularis*) $A$ which is a neighbor for $S$ and $P$ there is no string $AS$ or there exists $A$ which is a neighbor for $S$ and $P$ such that there is no string $AS$ or there is no string $AP$. This means that the plasmodium of *Physarum polycephalum* (of *Badhamia utricularis*) does not occupy $S$ or there is a neighboring cell which is not connected to $S$ or $P$ by a protoplasmic tube.

Notice that there are the following semantic correlations between propositions in the sense of the Black player and propositions in the sense of the White player:

- *SaP* is true for the Black player iff *SiP* is true for the White player with the same cells $S$ and $P$;
- *SoP* is true for the Black player iff *SeP* is true for the White player with the same cells $S$ and $P$;
- *SaP* is false for the Black player iff *SiP* is false for the White player with the same cells $S$ and $P$;
- *SoP* is false for the Black player iff *SeP* is false for the White player with the same cells $S$ and $P$.

Composite propositions are defined in the standard way.

In the performative syllogistic we have the following axioms:

$$SaP := (\exists A(A is S) \wedge (\forall A(A is S \wedge A is P))); \tag{9}$$

$$SiP := \forall A(\neg(A is S) \wedge \neg(A is P)); \tag{10}$$

$$\begin{aligned} &SoP := \neg(\exists A(A is S) \vee (\forall A(A is P \wedge A is S))), i.e. \\ &(\forall A \neg(A is S) \wedge \exists A(\neg(A is P) \vee \neg(A is S))); \end{aligned} \tag{11}$$

$$\begin{aligned} &SeP := \neg \forall A(\neg(A is S) \wedge \neg(A is P)), i.e. \\ &\exists A(A is S \vee A is P). \end{aligned} \tag{12}$$

$$SaP \Rightarrow SeP; \tag{13}$$

$$SaP \Rightarrow PaS; \tag{14}$$

$$SiP \Rightarrow PiS; \tag{15}$$

$$SaM \Rightarrow SeP; \tag{16}$$

$$MaP \Rightarrow SeP; \tag{17}$$

$$(MaP \wedge SaM) \Rightarrow SaP; \tag{18}$$

$$(MiP \wedge SiM) \Rightarrow SiP. \tag{19}$$

The formal properties of this axiomatic system are considered in [9], [10]. In the performative syllogistic we can analyze the collective dimension of behavior. Within this system we can study how the plasmodium of *Physarum polycephalum* and the plasmodium of *Badhamia utricularis* occupy all possible attractants in any direction if they can see them. So, this system shows logical properties of a massive-parallel behavior (i.e. the collective dimension of behavior). One of the most significant notions involved in this implementation of the performative syllogistic in plasmodia topology is a *neighborhood*. We can define a distance for the neighborhood differently, i.e. we can make it broader or narrower. So, from different neighborhoods it will follow that we deal with different 'universes of discourse'.

### A. Antagonistic Go Game for Verifying Performative Syllogistic Reasoning

In the Go game diagrams for the performative syllogistic, the 'universe of discourse' cover cells $x$, $y$, non-$x$ (which be denoted by $x'$), non-$y$ (which be denoted by $y'$):

| | |
|---|---|
| $x$ | $y'$ |
| $y$ | $x'$ |

where $x$, $y$, $x'$, $y'$ are neighbor cells containing black stones (interpreted as attractants for *Physarum polycephalum*) and white stones (interpreted as attractants for *Badhamia utricularis*), $x'$ are all neighbors for $y$ which differ from $x$, and $y'$ are all neighbors for $x$ which differ from $y$. Let us consider the Go game from the point of view just of the Black player. Suppose that we have black, white, and grey stones and (i) if a black stone is placed within a cell, this means that "this cell is occupied by the plasmodium of *Physarum polycephalum*" (i.e. "there is at least one thing in it for the Black player"), (ii) if a white stone is placed within a cell, this means that "this cell is not occupied the plasmodium of *Physarum polycephalum*" (i.e. "there is not thing in it for the Black player"), (ii) if a grey stone is placed within a cell, this means that "it is not known if this cell is occupied by the plasmodium of *Physarum polycephalum*" or "it is not known what color of stone placed within a cell is from the point of view of the Black player". All possible combinations of Go game diagrams for atomic propositions within our universe of discourse are pictured in Fig. 10.

The universe of discourse for simulating performative syllogisms by means of *Physarum* behaviors covers cells $x$, $y$, $m$, $x'$, $y'$, $m'$ in the following manner:

| | | | |
|---|---|---|---|
| $y'$ | $m$ | $m'$ | $x'$ |
| $m'$ | $x$ | $y'$ | $m$ |
| $m$ | $y$ | $x'$ | $m'$ |
| $x$ | $m'$ | $m$ | $y$ |

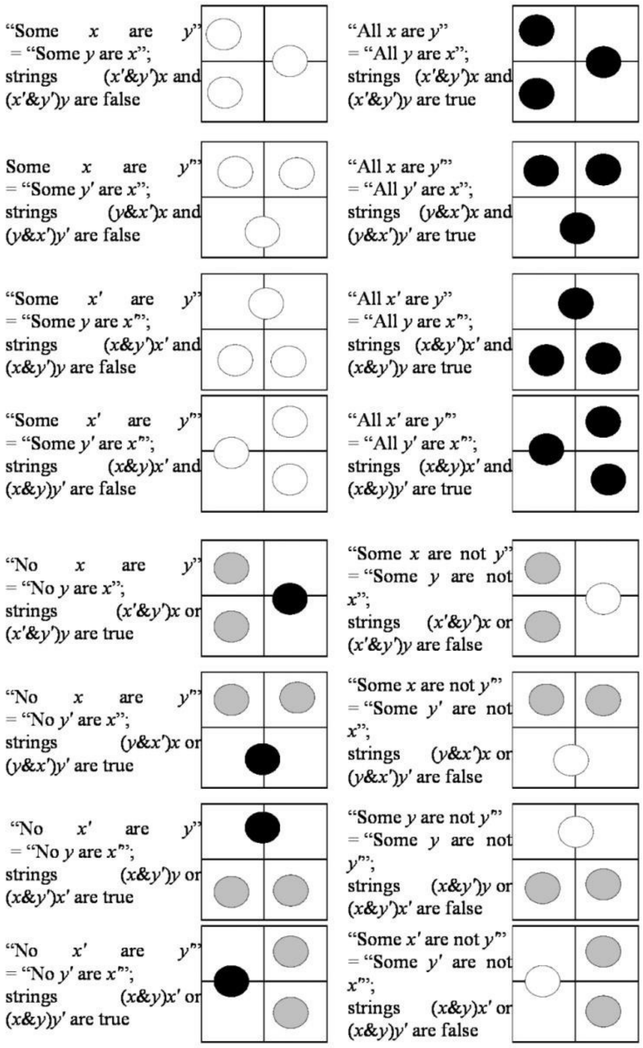
The motion of plasmodium starts from one of the central cells ($x$, $y$, $x'$, $y'$) and goes towards one of the four directions (northwest, southwest, northeast, southeast). The Go game diagram for syllogistic conclusions is as follows:

| | |
|---|---|
| $x$ | $m'$ |
| $m$ | $x'$ |

Some examples of performative syllogistic conclusions are regarded in Fig. 11.

Thus, the performative syllogistic allows us to study different zones containing attractants for *Physarum* *polycephalum* and *Badhamia utricularis* if they are connected by protoplasmic tubes homogenously.

A model $\mathsf{M}' = \langle M', |\cdot|_x \rangle$ for the performative syllogistic, where $M'$ is the set of attractants and $|X|_x \subseteq M'$ is a meaning of syllogistic letter $X$ which is understood as all attractants reachable for the plasmodium of *Physarum polycephalum* from the point $x$, is defined as follows:



“Some $x$ are $y$” = “Some $y$ are $x$”; strings $(x'\&y')x$ and $(x'\&y')y$ are false

“All $x$ are $y$” = “All $y$ are $x$”; strings $(x'\&y')x$ and $(x'\&y')y$ are true

Some $x$ are $y'$” = “Some $y'$ are $x$”; strings $(y\&x')x$ and $(y\&x')y'$ are false

“All $x$ are $y'$” = “All $y'$ are $x$”; strings $(y\&x')x$ and $(y\&x')y'$ are true

“Some $x'$ are $y$” = “Some $y$ are $x'$”; strings $(x\&y')x'$ and $(x\&y')y$ are false

“All $x'$ are $y$” = “All $y$ are $x'$”; strings $(x\&y')x'$ and $(x\&y')y$ are true

“Some $x'$ are $y'$” = “Some $y'$ are $x'$”; strings $(x\&y)x'$ and $(x\&y)y'$ are false

“All $x'$ are $y'$” = “All $y'$ are $x'$”; strings $(x\&y)x'$ and $(x\&y)y'$ are true

“No $x$ are $y$” = “No $y$ are $x$”; strings $(x'\&y')x$ or $(x'\&y')y$ are true

“Some $x$ are not $y$” = “Some $y$ are not $x$”; strings $(x'\&y')x$ or $(x'\&y')y$ are false

“No $x$ are $y'$” = “No $y'$ are $x$”; strings $(y\&x')x$ or $(y\&x')y'$ are true

“Some $x$ are not $y'$” = “Some $y'$ are not $x$”; strings $(y\&x')x$ or $(y\&x')y'$ are false

“No $x'$ are $y$” = “No $y$ are $x'$”; strings $(x\&y')y$ or $(x\&y')x'$ are true

“Some $y$ are not $y'$” = “Some $y$ are not $y'$”; strings $(x\&y')y$ or $(x\&y')x'$ are false

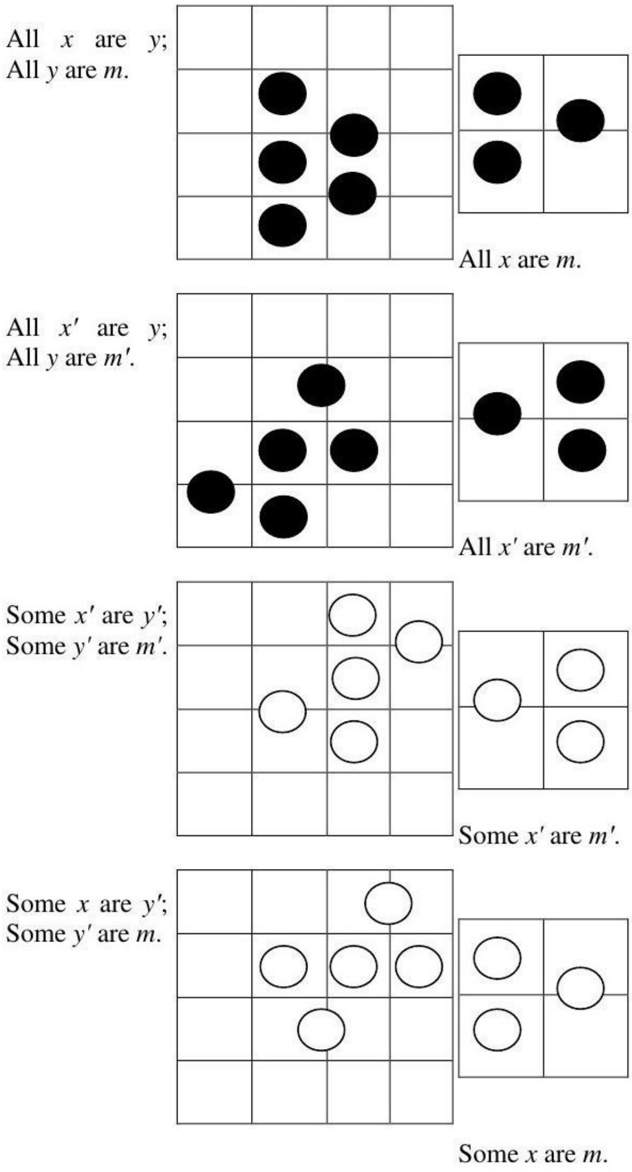
“No $x'$ are $y'$” = “No $y'$ are $x'$”; strings $(x\&y)x'$ or $(x\&y)y'$ are true

“Some $x'$ are not $y'$” = “Some $y'$ are not $x'$”; strings $(x\&y)x'$ or $(x\&y)y'$ are false

Fig. 10 The Go game diagrams for premises of performative syllogisms. Strings of the form $(x' \& y')x$ mean that in cells $x'$ and $y'$ there are neighbors $A$ for $x$ such that $Ax$, i.e. $(x' \& y')$ is a metavariable in $(x' \& y')x$ that is used to denote all attractants of $x'$ and $y'$ which are neighbors for the attractant of $x$

$\mathsf{M}' \models$ *All x are y* iff $|X|_x \neq \varnothing$, $|X|_y \neq \varnothing$, and $|X|_x \cap |X|_y \neq \varnothing$, more precisely both $(x' \& y')x$ and $(x' \& y')y$ hold in $\mathsf{M}'$, i.e. the plasmodium of *Physarum polycephalum* can move from neighbors of $y$ to $x$ and it can move from neighbors of $x$ to $y$ (we can place black stones in the line from neighbors of $y$ to $x$ and from neighbors of $x$ to $y$);

$\mathsf{M}' \models$ *Some x are y* iff $y \notin |X|_x$ and $x \notin |X|_y$, more precisely neither $(x' \& y')x$ nor $(x' \& y')y$ hold in $\mathsf{M}'$, i.e. the plasmodium of *Physarum polycephalum* cannot move from neighbors of $y$ to $x$ and it cannot move from neighbors of $x$ to $y$ (we can place white stones in the line from neighbors of $y$ to $x$ and from neighbors of $x$ to $y$);



All $x$ are $y$;
All $y$ are $m$.

All $x$ are $m$.

All $x'$ are $y$;
All $y$ are $m'$.

All $x'$ are $m'$.

Some $x'$ are $y'$;
Some $y'$ are $m'$.

Some $x'$ are $m'$.

Some $x$ are $y'$;
Some $y'$ are $m$.

Some $x$ are $m$.

Fig. 11 The Go game diagrams for performative syllogisms with true conclusions from the point of view of the Black player

$\mathsf{M}' \models$ *No x are y* iff $y \in |X|_x$ or $x \in |X|_y$, more precisely $(x' \& y')x$ or $(x' \& y')y$ hold in $\mathsf{M}'$, i.e. the plasmodium of *Physarum polycephalum* can move from neighbors of $y$ to $x$ or it can move from neighbors of $x$ to $y$ (we can place black stones in the line from neighbors of $y$ to $x$ or from neighbors of $x$ to $y$);

$$\mathrm{M}' \models \textit{Some } x \textit{ are not } y \text{ iff } y \notin |X|_x \text{ or } x \notin |X|_y,$$

more precisely ($x'$ & $y'$)$x$ or ($x'$ & $y'$) do not hold in $\mathrm{M}'$, i.e. the plasmodium of *Physarum polycephalum* cannot move from neighbors of $y$ to $x$ or it cannot move from neighbors of $x$ to $y$ (we can place white stones in the line from neighbors of $y$ to $x$ or from neighbors of $x$ to $y$);

$$\mathrm{M}' \models p \wedge q \text{ iff } \mathrm{M}' \models p \text{ and } \mathrm{M}' \models q;$$

$$\mathrm{M}' \models p \vee q \text{ iff } \mathrm{M}' \models p \text{ or } \mathrm{M}' \models q;$$

$$\mathrm{M}' \models \neg p \text{ iff it is false that } \mathrm{M}' \models p.$$

*Proposition 3:* The performative syllogistic is sound and complete in $\mathrm{M}'$.

For more details on formal properties of performative syllogistic, please see [9], [10]. This syllogistic describes the logic of plasmodium propagation in all possible directions. For the implementation of this syllogistic we do not need repellents. It is a natural system. The performative syllogistic as a Go game is an antagonistic game, where two players draw own conclusions without any coalition (see Fig. 10).

### B. Examples of Performative-Syllogistic Go Game

Let us also examine a game of Go at time step 10 to provide an example for performative syllogistic. Let this game be shown in Figure 12. As usual, each Voronoi cell is denoted from $S_{1,1}$ to $S_{18,18}$. In the universe of Fig. 12 there are no universal affirmative propositions and particular affirmative propositions. We face only universal negative propositions and particular negative propositions such as 'No $S'_{4,9}$ are $S'_{4,10}$', where $S'_{4,9}$ are neighbors for $S_{4,10}$ differing from $S_{4,9}$ and $S'_{4,10}$ are neighbors for $S_{4,9}$ differing from $S_{4,10}$, and 'Some $S'_{4,11}$ are not $S'_{5,11}$', where $S'_{4,11}$ are neighbors for $S_{5,11}$ differing from $S_{4,11}$ and $S'_{5,11}$ are neighbors for $S_{4,11}$ differing from $S_{5,11}$.

In the universe of Figure 13 there is a universal affirmative proposition: 'All $S_{6,5}$ are $S_{6,6}$', and a particular affirmative proposition: 'Some $S_{5,8}$ are $S_{5,9}$'.

## V. Conclusion

We have just shown that we can simulate the plasmodium motion as a Go game, where black stones are interpreted as attractants and white stones as repellents. We can consider configurations of stones as spatial reasoning. If we implement the Aristotelian syllogistic, we need a coalition of two players. If we implement the performative syllogistic [9], [10], we deal with an antagonistic game. In the latter case black stones are interpreted as attractants for *Physarum polycephalum* and white stones as attractants for *Badhamia utricularis.*

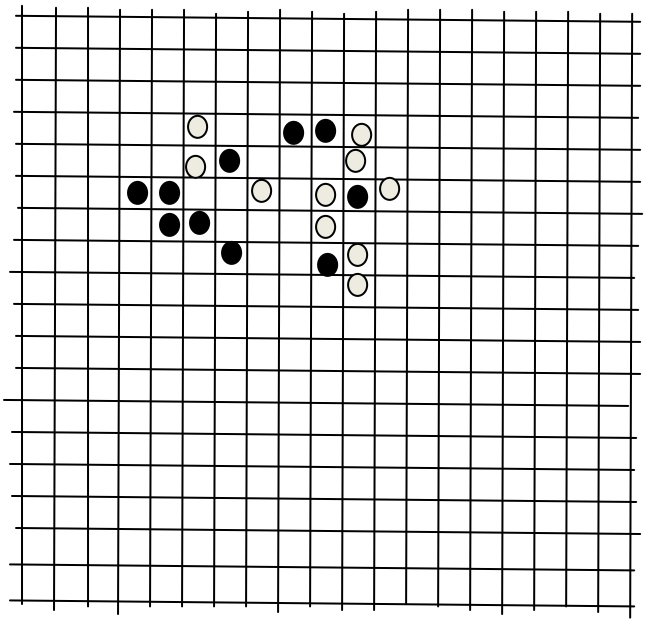


Fig. 12 The performative syllogistic Go game 1 at time step 10. The black stones are attractants occupied by the plasmodium of *Physarum polycephalum* and the white stones are attractants occupied by the plasmodium of *Badhamia utricularis*. So, we construct syllogisms from the point of view of the Black player

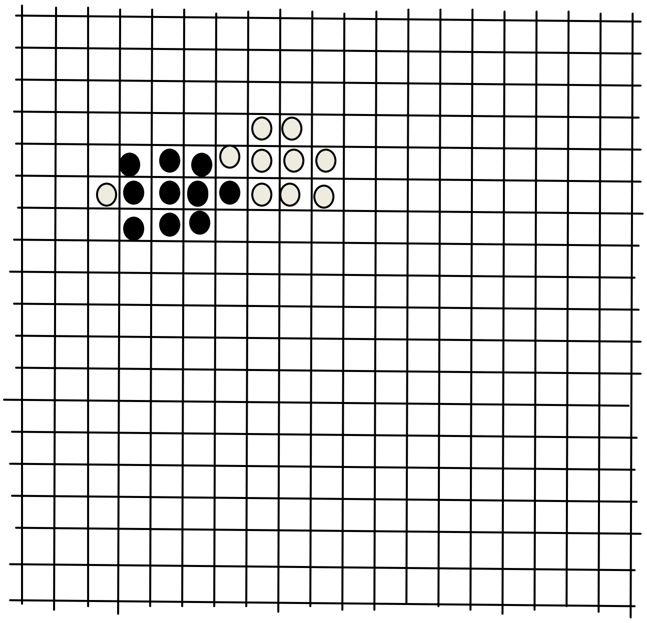


Fig. 13 The performative syllogistic Go game 2 at time step 10. The black stones are attractants occupied by the plasmodium of *Physarum polycephalum* and the white stones are attractants occupied by the plasmodium of *Badhamia utricularis*. We build syllogisms from the point of view of the Black player

## References

[1] A. Adamatzky, V. Erokhin, M. Grube, Th. Schubert, A. Schumann, "Physarum Chip Project: Growing Computers From Slime Mould," *International Journal of Unconventional Computing*, vol. 8, no. 4, pp. 319–323, 2012.

[2] A. Adamatzky, *Physarum Machines: Computers from Slime Mould* (World Scientific Series on Nonlinear Science, Series A). World Scientific Publishing Company, 2010.

[3] J. Łukasiewicz, *Aristotle's Syllogistic From the Standpoint of Modern Formal Logic*. Oxford Clarendon Press, 2nd edition, 1957.

[4] T. Nakagaki, H. Yamada, A. Toth, *Maze-solving by an amoeboid organism. Nature,* vol. 407, pp. 470–470, 2000.

[5] T. Nakagaki, H. Yamada, and A. Tothm, "Path finding by tube morphogenesis in an amoeboid organism," *Biophysical Chemistry,* vol. 92, pp. 47–52, 2001.

[6] T. Nakagaki, M. Iima, T. Ueda, Y. Nishiura, T. Saigusa, A. Tero, R. Kobayashi, K. Showalter, "Minimum-risk path finding by an adaptive amoeba network," *Physical Review Letters,* vol. 99, pp. 68–104, 2007.

[7] T. Saigusa, A. Tero, T. Nakagaki, Y. Kuramoto, "Amoebae Anticipate Periodic Events," *Phys. Rev. Lett,* vol. 100, no. 1, 2008.

[8] A. Schumann, L. Akimova, "Simulating of Schistosomatidae (Trematoda: Digenea) Behavior by Physarum Spatial Logic," *Annals of Computer Science and Information Systems, Volume 1. Proceedings of the 2013 Federated Conference on Computer Science and Information Systems.* IEEE Xplore, 2013, pp. 225–230.

[9] A. Schumann, "On Two Squares of Opposition: the Lesniewski's Style Formalization of Synthetic Propositions," *Acta Analytica,* vol. 28, pp. 71–93, 2013.

[10] A. Schumann, "Two Squares of Opposition: for Analytic and Synthetic Propositions," *Bulletin of the Section of Logic,* vol. 40, no. 3/4, pp. 165–178, 2011.

[11] A. Schumann, "Payoff Cellular Automata and Reflexive Games," *Journal of Cellular Automata,* vol. 9, no. 4, pp. 287–313, 2014.

[12] A. Schumann, A. Adamatzky, "Physarum Spatial Logic," *New Mathematics and Natural Computation,* vol. 7, no. 3, pp. 483–498, 2011.

[13] A. Schumann, K. Pancerz, "Towards an Object-Oriented Programming Language for Physarum Polycephalum Computing," in *Proceedings of the Workshop on Concurrency, Specification and Programming (CS&P'2013),* Warsaw, Poland, September 25-27, pp. 389–397, 2013.

[14] A. Schumann, K. Pancerz, A. Adamatzky, and M. Grube, "Bio-Inspired Game Theory: The Case of Physarum Polycephalum," in *8th International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS). ICST,* 2015, doi: http://dx.doi.org/10.4108/icst.bict.2014.257869

[15] T. Shirakawa, Y.-P. Gunji, and Y. Miyake, "An associative learning experiment using the plasmodium of Physarum polycephalum," *Nano Communication Networks,* vol. 2, pp. 99–105, 2011.

[16] R. Smith, "Completeness of an ecthetic syllogistic," *Notre Dame J. Formal Logic,* vol. 24, no. 2, pp. 224–232, 1983.

[17] A. Tero, T. Nakagaki, K. Toyabe, K. Yumiki, R. Kobayashi, "A Method Inspired by Physarum for Solving the Steiner Problem," *IJUC,* vol. 6, no. 2, pp. 109–123, 2010.

[18] S. Tsuda, M. Aono, and Y.P. Gunji, "Robust and emergent Physarum-computing," *BioSystems,* vol. 73, pp. 45–55, 2004.

[19] Sh. Watanabe, A. Tero, A. Takamatsu, T. Nakagaki, "Traffic optimization in railroad networks using an algorithm mimicking an amoeba-like organism, Physarum plasmodium," *Biosystems,* vol. 105, no. 3, pp. 225–232, 2011.

# 8th Workshop on Computer Aspects of Numerical Algorithms

NUMERICAL algorithms are widely used by scientists engaged in various areas. There is a special need of highly efficient and easy-to-use scalable tools for solving large scale problems. The workshop is devoted to numerical algorithms with the particular attention to the latest scientific trends in this area and to problems related to implementation of libraries of efficient numerical algorithms. The goal of the workshop is meeting of researchers from various institutes and exchanging of their experience, and integrations of scientific centers.

## Topics

- Parallel numerical algorithms
- Novel data formats for dense and sparse matrices
- Libraries for numerical computations
- Numerical algorithms testing and benchmarking
- Analysis of rounding errors of numerical algorithms
- Languages, tools and environments for programming numerical algorithms
- Numerical algorithms on GPUs
- Paradigms of programming numerical algorithms
- Contemporary computer architectures
- Heterogeneous numerical algorithms
- Applications of numerical algorithms in science and technology

## Event Chairs

**Bylina, Jaroslaw,** Maria Curie-Sklodowska University, Poland

**Bylina, Beata,** Maria Curie-Sklodowska University, Poland

**Stpiczyński, Przemysław,** Maria Curie-Sklodowska University, Poland

## Program Committee

**Amodio, Pierluigi,** Università di Bari, Italy

**Anastassi, Zacharias,** Qatar University, Qatar

**Banaś, Krzysztof,** AGH University of Science and Technology, Poland

**Barán, Benjamín,** Universidad Nacional del Este

**Brugnano, Luigi,** Universita' di Firenze, Italy

**Czachorski, Tadeusz,** IITiS

**Filippone, Salvatore,** University Rome Tor Vergata, Italy

**Filote, Constantin**

**Fourneau, Jean-Michel**

**Gansterer, Wilfried,** University of Vienna, Austria

**Georgiev, Krassimir,** IICT - BAS, Bulgaria

**Gimenez, Domingo,** University of Murcia, Spain

**Gravvanis, George,** Democritus University of Thrace, Greece

**Knottenbelt, William,** Imperial College London, United Kingdom

**Kozielski, Stanislaw**

**Kucaba-Pietal, Anna,** Politechnika Rzeszowska, Poland

**Lirkov, Ivan,** Institute of Information and Communication Technologies , Bulgarian Academy of Sciences, Bulgaria

**Maksimov, Vyacheslav,** Institute of Mathematics and Mechanics, Russia

**Marowka, Ami,** Bar-Ilan University, Israel

**Petcu, Dana,** West University of Timisoara, Romania

**Pultarova, Ivana,** Czech Technical University in Prague, Czech Republic

**Satco, Bianca-Renata,** Stefan cel Mare University of Suceava, Romania

**Sedukhin, Stanislav,** The University of Aizu, Japan

**Sergeichuk, Vladimir,** Institute of Mathematics of NAS of Ukraine, Ukraine

**Shishkina, Olga,** Max Planck Institute for Dynamics and Self-Organization, Germany

**Srinivasan, Natesan,** Indian Institute of Technology, India

**Szadkowski, Zbigniew,** University of Lodz, Poland

**Szajowski, Krzysztof,** Institute of Mathematics and Computer Science, Poland

**Trivedi, Kishor S.,** Duke University, United States

**Tudruj, Marek,** Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland

**Tůma, Miroslav,** Academy of Sciences of the Czech Republic, Czech Republic

**Ustimenko, Vasyl,** Marie Curie-Sklodowska University, Poland

**Vazhenin, Alexander,** University of Aizu, Japan

**Wójcik, Grzegorz M.,** Institute of Computer Science, Maria Curie-Sklodowska University, Poland

**Wyrzykowski, Roman,** Czestochowa University of Technology, Poland

DOI: 10.15439/2015F354
ACSIS, Vol. 5

# Strategies of parallelizing nested loops on the multicore architectures on the example of the WZ factorization for the dense matrices

Beata Bylina, Jarosław Bylina
Marie Curie-Skłodowska University,
Institute of Mathematics,
Pl. M. Curie-Skłodowskiej 5,
20-031 Lublin, Poland
Email: {beata.bylina,jaroslaw.bylina}@umcs.pl

***Abstract*—In the WZ factorization the outermost parallel loop decreases the number of iterations executed at each step and this changes the amount of parallelism in each step. The aim of the paper is to present four strategies of parallelizing nested loops on multicore architectures on the example of the WZ factorization.**

**For random dense square matrices with the dominant diagonal we report the execution time, the performance, the speedup of the WZ factorization for these four strategies of parallelizing nested loops and we investigate the accuracy of such solutions.**

**It is possible to shorten the runtime when utlilizing the appropriate strategies with the use of good scheduling.**

**Keywords: linear system, WZ factorization, matrix factorization, matrix computations, multicore architecture, parallel nested loops, OpenMP**

## I. Introduction

MULTICORE computers with shared memory are used to solve the computational science problems. One of more important computational problems is solution of linear systems, the form of which is the following:

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \qquad \text{where} \quad \mathbf{A} \in \mathbb{R}^{n\times n}, \quad \mathbf{b} \in \mathbb{R}^{n}. \tag{1}$$

One of the direct methods of solving a dense linear system (1) is to factorize the matrix $\mathbf{A}$ into some simpler matrices — that is its decomposition into factor matrices of a simpler structure — and then solving simpler linear systems.

Such factorization is hard to compute and that is why it is worth applying different optimization techniques and simultaneously using parellelism of contemporary computers.

The implementation of the factorization contains nested loops. The reasearch of the parallelization of nested loops have been undertaken by different scientistcs.

In the work [6], the authors study five different models for nested parallel loops execution on shared-memory multiprocessors and show a simulation based performance comparison of different techniques using real application. The possibility to take advantage of the parallelism in nested parallel loops with the use of good scheduling and synchronization algorithms is described.

An automatic mechanism to dynamically detect the best way to exploit the parallelism when having nested parallel loops is presented in the study [3]. This mechanism is based on a number of threads, size of the problem, number of iterations in a loop and its was implemented inside IBM XL runtime library. This paper examined (among other) an LU kerner, which decomposes the matrix $\mathbf{A}$ into the matrices: $\mathbf{L}$ (lower triangular matrix) and $\mathbf{U}$ (upper triangular matrix).

An algorithm for finding good distributions of threads to tasks is provided and the implementation of nested parallelism in OpenMP is discussed in the paper [1].

The main focus of [5] was to investigate the possibility of dynamically choosing, at runtime, the loop which best utilises the available threads.

To implement parallel programs on multicore systems with shared-memory, programmers usually use the OpenMP standard [8]. The programming model provides a set of directives to explicitly define parallel regions in applications. The compliator translates these directives. One of its most interesting features in the language is the support for nested parallelism.

This work investigate the issue of the parallelizing nested loops in OpenMP. The OpenMP standard supports loop paralelism. For OpenMP standard, it is done by the utilization of the directive `#pragma omp parallel for`, which provides a shourtcut for specifying a parallel region that contains a single `#pragma omp parallel`.

Parallelism of the nested loops in the WZ factorization is the aim of the work. In the WZ kerner the outermost parallel loop decreases the amount of iterations executed at each step and this changes the number of parallelism in each step. In this paper we investigate the time, the scalability, the speedup and the acuraccy for four different nested loops parallelism strategies for the WZ factorization.

The paper deals with the following issues. In Section II the idea of the WZ factorization [2], [7] and the way the matrix $\mathbf{A}$ is factorized to a product of matrices $\mathbf{W}$ and $\mathbf{Z}$ are described. Such a factorization exists for every nonsingular matrix (with pivoting) which was shown in [2].

Section III provides information about some strategies of parallelizing nested loops and their application to the orginal the WZ factorization. Section IV presents the results of our

978-83-60810-66-8/$25.00©2015, IEEE

experiments. The time, the speedup, the performance of WZ factorization for different strategies on the two platforms are analysed. The influence of the size of the matrix on the achieved numerical accuracy is studied as well. Section V is a summary of our experiments.

## II. WZ Factorization (WZ)

The chapter presents the WZ factorization usage to solve (1). The WZ factorization is described in [2], [4].

Let us assume that the $\mathbf{A}$ is a square nonsingular matrix of an even size (it is somewhat easier to obtain formulas for even sizes than for odd ones). We are to find matrices $\mathbf{W}$ and $\mathbf{Z}$ that fulfill $\mathbf{WZ} = \mathbf{A}$ and the matrices $\mathbf{W}$ and $\mathbf{Z}$ consist of the following rows $\mathbf{w}_i^T$ and $\mathbf{z}_i^T$ respectively:

$$\begin{aligned}
\mathbf{w}_1^T &= (1, \underbrace{0, \ldots, 0}_{n-1}) \\
\mathbf{w}_i^T &= (w_{i1}, \ldots, w_{i,i-1}, 1, \underbrace{0, \ldots, 0}_{n-2i+1}, w_{i,n-i+2}, \ldots, w_{in}) \\
&\quad \text{for} \quad i = 2, \ldots, \tfrac{n}{2}, \\
\mathbf{w}_i^T &= (w_{i1}, \ldots, w_{i,n-i}, \underbrace{0, \ldots, 0}_{2i-n-1}, 1, w_{i,i+1}, \ldots, w_{in}) \\
&\quad \text{for} \quad i = \tfrac{n}{2} + 1, \ldots, n-1, \\
\mathbf{w}_n^T &= (\underbrace{0, \ldots, 0}_{n-1}, 1) \\
\mathbf{z}_i^T &= (\underbrace{0, \ldots, 0}_{i-1}, z_{ii}, \ldots, z_{i,n-i+1}, 0, \ldots, 0) \\
&\quad \text{for} \quad i = 1, \ldots, \tfrac{n}{2}, \\
\mathbf{z}_i^T &= (\underbrace{0, \ldots, 0}_{n-i}, z_{i,n-i+1}, \ldots, z_{ii}, 0, \ldots, 0) \\
&\quad \text{for} \quad i = \tfrac{n}{2} + 1, \ldots, n.
\end{aligned} \tag{2}$$

After the factorization we can solve two linear systems:

$$\left\{ \begin{array}{lcl} \mathbf{Wy} & = & \mathbf{b} \\ \\ \mathbf{Zx} & = & \mathbf{y} \end{array} \right. \tag{3}$$

(where $\mathbf{y}$ is an auxiliary intermediate vector) instead of one (1).

In this paper we are interested only in obtaining the matrices $\mathbf{Z}$ and $\mathbf{W}$. The first part of the algorithm consists of setting succesive parts of columns of the matrix $\mathbf{A}$ to zeros. In the first step we do that with the elements in the 1st and $n$th columns — from the 2nd row to the $(n-1)$st row. Next we update the matrix $\mathbf{A}$.

More formally we can describe the first step of the algorithm the following way.

1) For every $i = 2, \ldots, n-1$ we compute $w_{i1}$ and $w_{in}$ from the system:

$$\left\{ \begin{array}{lclcl} a_{11} w_{i1} & + & a_{n1} w_{in} & = & -a_{i1} \\ a_{1n} w_{i1} & + & a_{nn} w_{in} & = & -a_{in} \end{array} \right.$$

and we put them in the matrix of the form:

$$\mathbf{W}^{(1)} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ w_{21} & 1 & \ddots & \vdots & w_{2n} \\ \vdots & 0 & \ddots & 0 & \vdots \\ w_{n-1,1} & \vdots & \ddots & 1 & w_{n-1,n} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

2) We compute:

$$\mathbf{A}^{(1)} = \mathbf{W}^{(1)} \mathbf{A}.$$

After the first step we get a matrix of the form:

$$\mathbf{A}^{(1)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2,n-1}^{(1)} & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{n-1,2}^{(1)} & \cdots & a_{n-1,n-1}^{(1)} & 0 \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & a_{nn} \end{bmatrix}, \tag{4}$$

where (for $i, j = 2, \ldots, n-1$):

$$a_{ij}^{(1)} = a_{ij} + w_{i1} a_{1j} + w_{in} a_{nj}. \tag{5}$$

Then, we proceed analogously — but for the inner square matrices — $\mathbf{A}^{(1)}$ of size $(n-2)$ and so on.

So, the whole algorithm is the following.

For $k = 1, 2, \ldots, \frac{n}{2} - 1$:

1) For every $i = k+1, \ldots, n-k$ we compute $w_{ik}$ and $w_{i,n-k+1}$ from the system:

$$\left\{ \begin{array}{lcl} a_{kk}^{(k-1)} w_{ik} & + & a_{n-k+1,k}^{(k-1)} w_{i,n-k+1} \\ & = & -a_{ik}^{(k-1)} \\ \\ a_{k,n-k+1}^{(k-1)} w_{ik} & + & a_{n-k+1,n-k+1}^{(k-1)} w_{i,n-k+1} \\ & = & -a_{i,n-k+1}^{(k-1)} \end{array} \right.$$

and we put them in a matrix of the form shown in Figure 1.

2) We compute:

$$\mathbf{A}^{(k)} = \mathbf{W}^{(k)} \mathbf{A}^{(k-1)}.$$

After $(\frac{n}{2} - 1)$ such steps we get the matrix

$$\mathbf{Z} = \mathbf{A}^{(\frac{n}{2} - 1)}.$$

Moreover, we know that:

$$\mathbf{W}^{(\frac{n}{2}) - 1} \cdot \ldots \cdot \mathbf{W}^{(1)} \cdot \mathbf{A} = \mathbf{Z},$$

so we get

$$\mathbf{A} = \{\mathbf{W}^{(1)}\}^{-1} \cdot \ldots \cdot \{\mathbf{W}^{(\frac{n}{2}) - 1}\}^{-1} \cdot \mathbf{Z} = \mathbf{WZ}.$$

Algorithm 1 shows the WZ algorithm.

The complexity of Algorithm 1 can be expressed by the formule

$$\sum_{k=1}^{\frac{n}{2}-1} \left( 3 + \sum_{i=k+1}^{n-k} \left( 8 + \sum_{j=k+1}^{n-k} 4 \right) \right) = \frac{4n^3 - 7n - 18}{6}. \tag{6}$$

So, this algorithm requires $O(n^3)$ arthmetic operations.

$$\mathbf{W}^{(k)} = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & w_{k+1,k} & \ddots & & w_{k+1,n-k+1} & & \\ & & \vdots & & \ddots & \vdots & & \\ & & w_{n-k,k} & & \ddots & w_{n-k,n-k+1} & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix}$$

Fig. 1. The matrix $\mathbf{W}^{(k)}$ in $k$th step.

**Algorithm 1** Outline of the WZ factorization algorithm (WZ)

```
Require: A
Ensure: W, Z
for $k = 1$ to $n/2 - 1$ do
    $k2 \leftarrow n - k + 1$
    $det \leftarrow a_{kk} * a_{k2k2} - a_{k2k} * a_{kk2}$
    for $i = k + 1$ to $k2 - 1$ do
        $w_{ik} \leftarrow (a_{k2k2} * a_{ik} - a_{k2k} * a_{ik2})/det$
        $w_{ik2} \leftarrow (a_{kk} * a_{ik2} - a_{kk2} * a_{ik})/det$
        for $j = k + 1$ to $k2 - 1$ do
            $a_{ij} \leftarrow a_{ij} - w_{ik} * a_{kj} - w_{ik2} * a_{k2j}$
        end for
    end for
end for
$Z \leftarrow A$
```

## III. Nested loops parallelism strategies

An application with nested loops can be performed parallely in different ways depending on compilators, hardware and run-time system support available. Nested loops require from of a programmer taking a decision concerning details of parallelism.

In this work we deal with the following parallelization strategies for nested loops:

1) *outer*
2) *inner*
3) *nested*
4) *split*

While all variables used in a parallel region are by default `shared`, in each strategy we declare explicitly all variables as `private` or `shared` for all directives respectively. Using the `private` clause, we specify that each thread has its own copy of variables.

To ensure load balancing for all threads we use the schedule clause, which specifies how the iterations of the loop are assigned to the threads. In the clause `schedule` of the directive `#pragma omp parallel for` we set the value `static`, because the computational cost of the tasks is known.

### A. Outer

*Outer* — the simplest parallelization strategy of nested loops is parallel execution of the most outer loop. All inner loops are executed in a sequence. This approach gives good results if the number of iterations in a loop is big and the iteration's granularity is coarse enough, which happens exactly in case of the WZ factorization. Algorithm 2 presents outer strategy for WZ factorization. The outermost $k$-loop cannot be parallelized, however, we can parallelize the $i$-loop. In this simple parallelization strategy the loop is divided equally between threads, so every thread performs the same amount of work, which ansure regular distribution of work beetwen threads.

### B. Inner

Another strategy of paralelizing nested loops involves executing the inner loops in parallel on all processors, but the outer loop is executed in a sequence. Clearly, in case of WZ factorization blocking barrier is used at the end of each parallel

**Algorithm 2** Outline of the WZ factorization algorithm (WZ) — outer strategy

```
Require: A
Ensure: W, Z
for $k = 1$ to $n/2 - 1$ do
    $k2 \leftarrow n - k + 1$
    $det \leftarrow a_{kk} * a_{k2k2} - a_{k2k} * a_{kk2}$
    #pragma omp parallel for private(i) shared(k, k2, w, a, det,j)
    for $i = k + 1$ to $k2 - 1$ do
        $w_{ik} \leftarrow (a_{k2k2} * a_{ik} - a_{k2k} * a_{ik2})/det$
        $w_{ik2} \leftarrow (a_{kk} * a_{ik2} - a_{kk2} * a_{ik})/det$
        for $j = k + 1$ to $k2 - 1$ do
            $a_{ij} \leftarrow a_{ij} - w_{ik} * a_{kj} - w_{ik2} * a_{k2j}$
        end for
    end for
end for
$Z \leftarrow A$
```

loop, which prevents incorect results. Parallelizing the inner loop will potentially provide smaller pieces of work so they can be distributed evenly between the available threads but it has more overhead due to work distribution and synchronization beetwen threads. This overhead may be high if the loop granularity is too fine. Algorithm 3 presents inner strategy for WZ factorization, in which the $j$-loop are parallelized.

### C. Nested

The third strategy of execution of nested loops paralelization is exploiting the paralelism on each level — nested parallelism. Standard OpenMP (from 2.5 version) makes it possible to nest parallel loops, however, it must be switched on by means of the environment variable `OMP_NESTED` or the function `omp_set_nested`. Each task needs at least one thread to its own disposal. Algorithm 4 presents the nested strategy.

This algorithm shows how a 2-level parallelism can be implemented in OpenMP based on the directives. Nesting parallel loops is a way to use more threads in a computation. This can easily create a large number of threads as their number is the product of the number of threads forked at each level of nested loops.

### D. Split

The final strategy concernes division of $i$-loops into two separate loops. Such a split facilites presentation of $k$th step in the form of a dag (directed acyclic graph), which shows the order of the task execution. The dag represents computational solutions in which the nodes represent tasks to be executed and edges represent precedence among the tasks. In the figure 2 a dag for $k$th step and shows, which part of the matrix is processed in a particular task is presented. By Task 1 we understand determining of valiables $k2$ and $det$ (lines 2 and 3 in Algorithm 1. Task 2 is the computation of $k$th and $k2$nd column of the matrix $\mathbf{W}$ (lines 4, 5 and 6 in Algorithm 1). Task 3 is the computation of values in the matrix $\mathbf{A}$ (lines 4, 7 and 8 in Algorithm 1).

Algorithm 5 shows the *split* strategy for WZ factorization. The first loop is parallelized. The second loop is nested loop and we use outer version to parallelize this loop.

## IV. Numerical Experiments

In this section we tested the time, the performance, the speedup and the absolute accuracy of the WZ factorization. Our intention was to investigate diffrent nested loops parallelization strategies for the WZ factorization on multicore architecuters. We examined five versions algorithms of the WZ factorization:

- *sequential* (Algorithm 1),
- *outer* (Algorithm 2),
- *inner* (Algorithm 3),
- *nested* (Algorithm 4),
- *split* (Algorithm 5).

Here we used experiments, based on information collected at runtime, to decide whether a loop should execute clause `static` or `dynamic` and we chose `static`.

The input matrices are generated (by the authors). They are random, dense, square matrices with a dominant diagonal of even sizes (1000, 2000,. . . , 9000)

We used two hardware platforms for testing: E5-2660 and X5650. Their details specifications are presented in Table I.

The algorithms *sequential*, *outer*, *inner*, *nested* and *split* were implemented with the use of the C language with the use of the double precision. Our codes were compiled by INTEL C Compiler (icc) with optimization flag `-O3`. Additionally, all algorithms were linked with the OpenMP library.

### A. The Time

All the processing times are reported in seconds. The time is measured with an OpenMP function `open_get_wtime()`. They were tested in the double precision.

In Figures 3 and 4 we have compared the average running time of the four parallel WZ decomposition algorithms and the sequential version on two platforms.

**Algorithm 3** Outline of the WZ factorization algorithm (WZ) — inner strategy

```
Require: A
Ensure: W, Z
for $k = 1$ to $n/2 - 1$ do
    $k2 \leftarrow n - k + 1$
    $det \leftarrow a_{kk} * a_{k2k2} - a_{k2k} * a_{kk2}$
    for $i = k + 1$ to $k2 - 1$ do
        $w_{ik} \leftarrow (a_{k2k2} * a_{ik} - a_{k2k} * a_{ik2})/det$
        $w_{ik2} \leftarrow (a_{kk} * a_{ik2} - a_{kk2} * a_{ik})/det$
        #pragma omp parallel for private(j) shared(k, k2, w, a, det,i)
        for $j = k + 1$ to $k2 - 1$ do
            $a_{ij} \leftarrow a_{ij} - w_{ik} * a_{kj} - w_{ik2} * a_{k2j}$
        end for
    end for
end for
$Z \leftarrow A$
```

**Algorithm 4** Outline of the WZ factorization algorithm (WZ) — nested strategy

```
Require: A
Ensure: W, Z
for $k = 1$ to $n/2 - 1$ do
    $k2 \leftarrow n - k + 1$
    $det \leftarrow a_{kk} * a_{k2k2} - a_{k2k} * a_{kk2}$
    #pragma omp parallel for private(i) shared(k, k2, w, a, det)
    for $i = k + 1$ to $k2 - 1$ do
        $w_{ik} \leftarrow (a_{k2k2} * a_{ik} - a_{k2k} * a_{ik2})/det$
        $w_{ik2} \leftarrow (a_{kk} * a_{ik2} - a_{kk2} * a_{ik})/det$
        #pragma omp parallel for private(j) shared(k, k2, w, a, det)
        for $j = k + 1$ to $k2 - 1$ do
            $a_{ij} \leftarrow a_{ij} - w_{ik} * a_{kj} - w_{ik2} * a_{k2j}$
        end for
    end for
end for
$Z \leftarrow A$
```

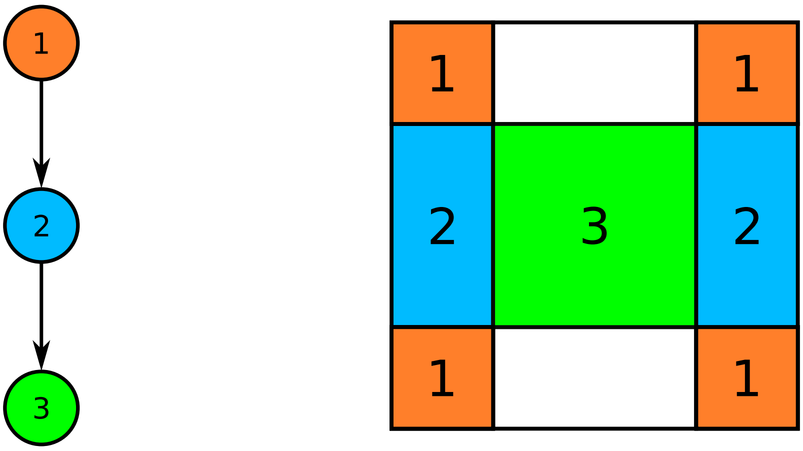


Fig. 2. The dag of the tasks (left). The sequence of calculations in the matrix in the WZ factorization in every step (right).

**Algorithm 5** Outline of the WZ factorization algorithm (WZ) — split strategy

```
Require: A
Ensure: W, Z
for $k = 1$ to $n/2 - 1$ do
  $k2 \leftarrow n - k + 1$
  $det \leftarrow a_{kk} * a_{k2k2} - a_{k2k} * a_{kk2}$
  #pragma omp parallel for private(i) shared(k, k2, w, a, det)
  for $i = k + 1$ to $k2 - 1$ do
    $w_{ik} \leftarrow (a_{k2k2} * a_{ik} - a_{k2k} * a_{ik2})/det$
    $w_{ik2} \leftarrow (a_{kk} * a_{ik2} - a_{kk2} * a_{ik})/det$
  end for
  #pragma omp parallel for private(i) shared(k, k2, w, a, det)
  for $i = k + 1$ to $k2 - 1$ do
    for $j = k + 1$ to $k2 - 1$ do
      $a_{ij} \leftarrow a_{ij} - w_{ik} * a_{kj} - w_{ik2} * a_{k2j}$
    end for
  end for
end for
$Z \leftarrow A$
```

TABLE I
SOFTWARE AND HARDWARE PROPERTIES OF E5-2660 AND X5650 SYSTEMS

| | E5-2660 System | X5650 System |
|---|---|---|
| CPU | 2x Intel Xeon E5-2660 (20M Cache, 2.20 GHz, 8 cores with HT) | 2x Intel Xeon X5650 (12M Cache, 2.66 GHz, 6 cores with HT) |
| CPU memory | 48GB DDR3 | 48GB DDR3 |
| Operating system | CentOS 5.5 (Linux 2.6.18-164.el5) | Debian (GNU/Linux 7.0) |
| Libraries | OpenMP, Intel Composer XE 2013 | OpenMP, Intel Composer XE 2013 |
| Compilers | Intel | Intel |

Figure 3 shows the dependence of the time on the number of threads for the matrix of the size 9000 on two platforms (X5650 on the right side, E5-2660 on the left side).

Figure 4 shows the dependence of the time on the matrix size for 12 threads for X5650 system (the right side) and 16 threads for E5-2660 system (the left side).

Using obtained results we conclude that:

- For a growing number of threads E5-2660 architecture outperforms X5650, due to the fact that the latter one is its older. We were expecting this result.
- The time is the shortest for 12 threads on the X5650 system and for 16 threads on the E5-2660 system. For bigger number of threads the time is the same as for 12 threads on the X5650 system nad for 16 threads on E5-2660, which proves weakness of the hyperthrading technology.
- If the size matrix is increased, then the runtime is increased too and it becomes more profitable to use a big number of the threads.
- *split* and *outer* algorithms achieve very similar execution time, which is the shortest compared with other algorithms.
- The worse execution time was achieved by the *nested* algorithm and for E5-2660 it is even worse than sequential algorithm.

### B. The Performance

Figures 5 and 6 compare the performance (in Gflops) results obtained for those five algorithms (*sequential*, *outer*, *inner*, *nested*, *split*) — in the double precision on two platforms. The performance is based on the number of floating-point operations in the WZ factorization (6).

Figure 5 shows dependence of the performance on the number of threads (maximum number of the threads is 24) for the matrix of the size 9000 for two platforms (X5650 — the right, E5-2660 — the left).

Figure 6 shows dependence of the performance on matrix size for 12 threads for X5650 system (the right side) and 16 threads for E5-2660 (the left side).

We can see the best performance (about 5.5 Gflop/s) achieved by *split* algorithm for the matrix of the size 9000 for 16 threads on E5-2660 system, and worst (less than 1 Gflop/s) is for *nested* version and *sequential* algorithm for all matrix sizes. On X5650 system we obtain worse performance for all tested algorithms than on E5-2660 System. The performance is very low for all algorithms on X5650 system and almost the same for inner, outer and *split* algorithms on X5650 system.

### C. The Speedup

Figures 7 and 8 present the speedup results obtained for four algorithms implementations — in the double precision on two platforms. Figure 7 shows dependence of the speedup on the number of threads (maximum number of the threads is

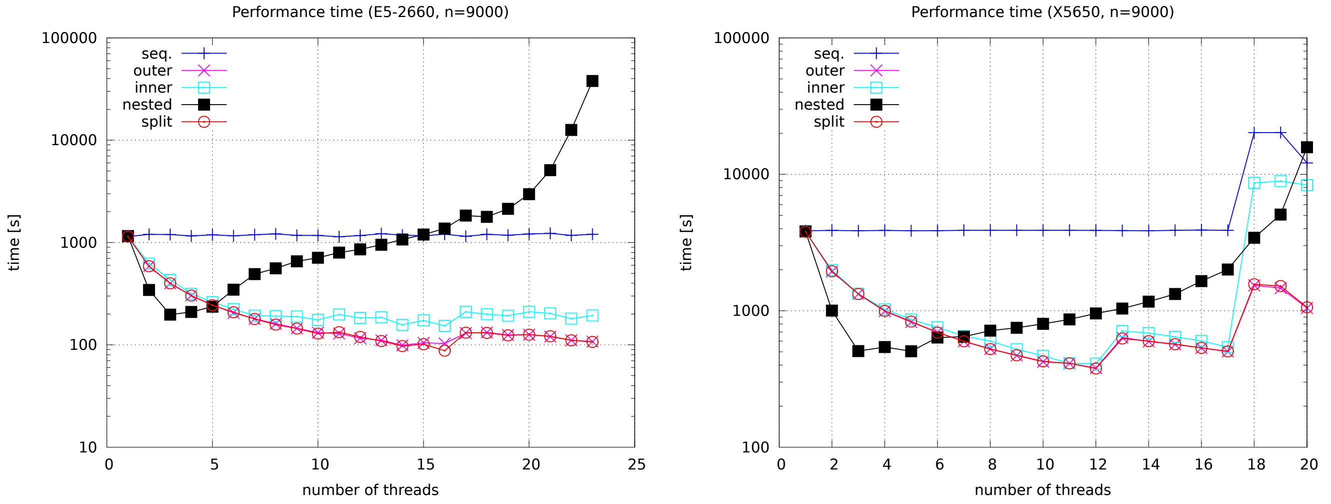


Fig. 3. The average running time of the WZ matrix decomposition as a function of the number of threads — for the five algorithms using the double precision on two platforms (E5-2660 on the left side and X5650 on the right side) for the matrix of the size 9000 (logarithmic $y$-axis).

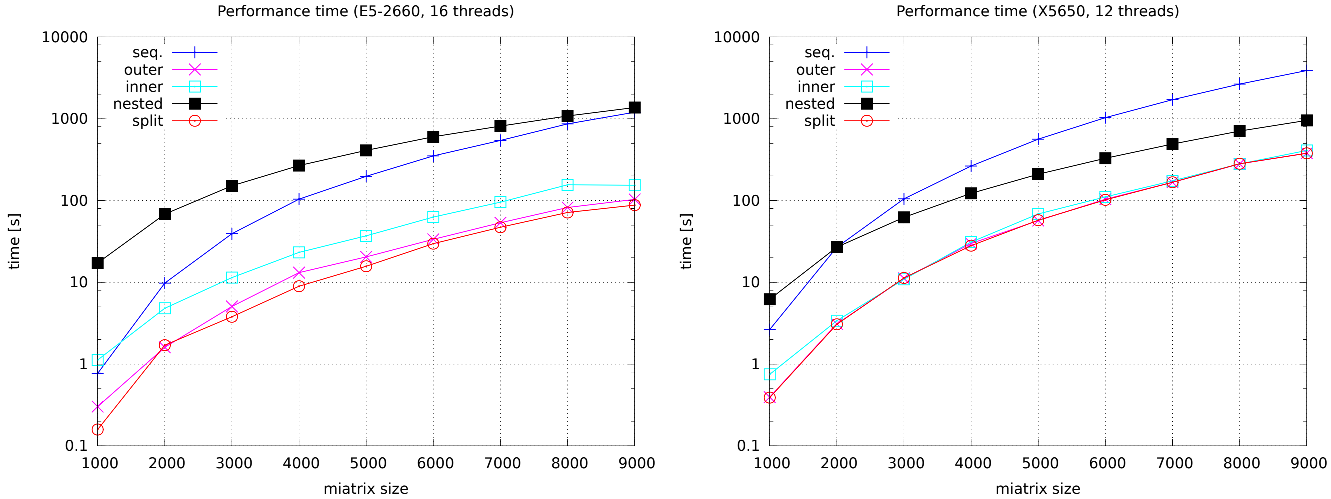


Fig. 4. The average running time of the WZ matrix decomposition as a function of the matrix size — for 16 threads on E5-2660 system (the left side) and for 12 threads on X5650 system (the right side) (logarithmic $y$-axis ).

23) for the matrix of the size 9000 for two platforms (X5650 — the right, E5-2660 — the left).

Figure 8 shows dependence of the performance on the matrix size for 12 threads for X5650 system (the right side) and 16 threads for E5-2660 (the left side).

Note that:

- All algorithms scale well with the size of a matrix; moreover,the bigger the matrix, the better the speedup.
- The speedup increases steadily until 12 threads on E5-2660 System and 16 threads on X5650 system, before it starts to level off.
- *Split* algorithm has the better speedup, even value up to 14 for 16 threads on E5-2660 system.
- On the X5650 system *split* and outer algorithm have similary seedup, but on E5-2660 System *split* algorithm has higher speedup than *split* algorithm.

### D. Numerical Accuracy

The purpose of this section is not to accomplish a full study of the numerical stability and accuracy of the WZ factorization, but justify experimentally that our implementation of the WZ algorithm can be used in practice.

As a measure of accuracy we took the following expression (where $||\mathbf{M}||$ is the Frobenius norm of the matrix $\mathbf{M}$) based on the absolute error:

$$||\mathbf{A} - \mathbf{WZ}||.$$

Table II illustrates the accuracy (given as the norms $||\mathbf{A} - \mathbf{WZ}||$) of the WZ factorization. The norms on both platforms (E5-2660 and X5650) are the same for appropriate matrix

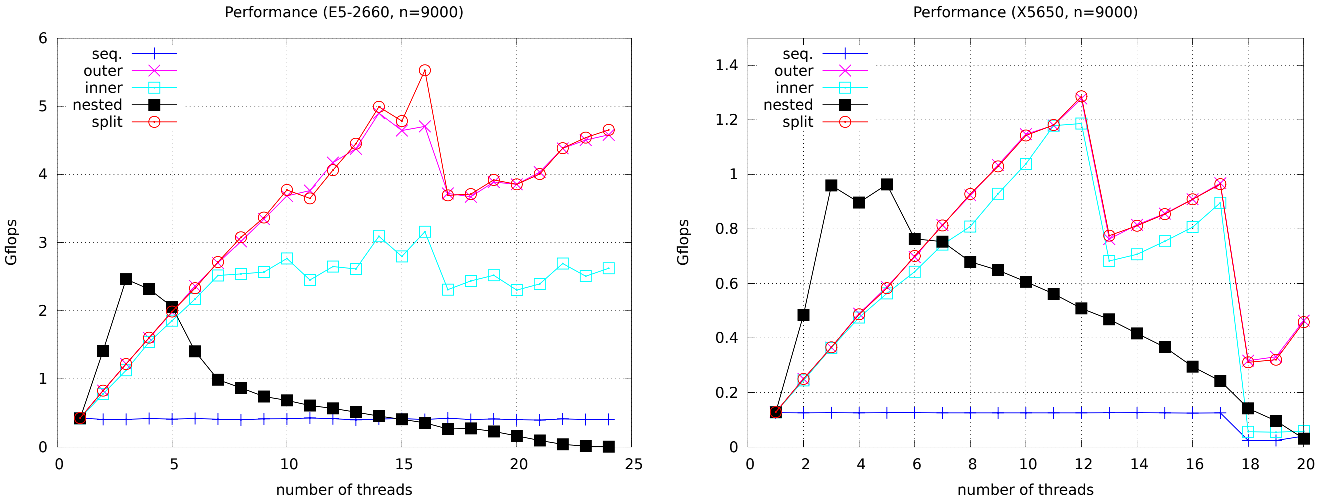


Fig. 5. The performance results for the WZ factorization — using the double precision on two platforms (E5-2660 — the left side; X5650 — the right side) for the five algorithms as the function of number of threads.

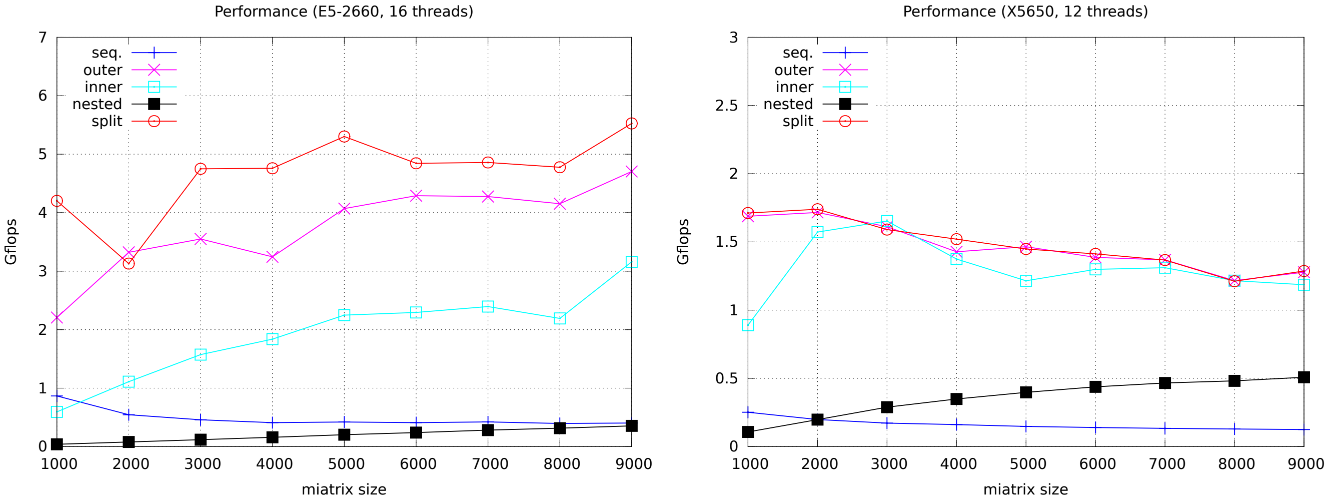


Fig. 6. The performance results for the WZ factorization — using the double precision for 16 threads on E5-2660 system (the left side) and for 12 threads on X5650 system (the right side) for the five algorithms as a function of the matrix size.

sizes. Values of the norm do not depend on the number of the threads and do not depend on a choice of algorithms (for all algorithms the norm depends only on the matrix size).

## V. Conclusion

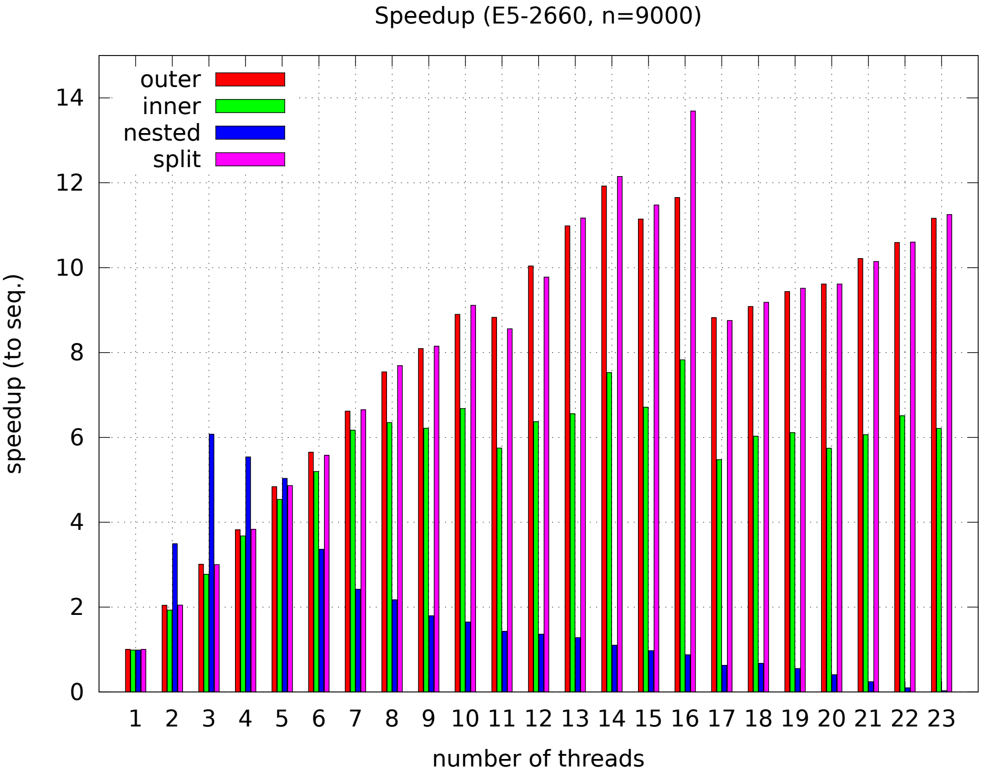
In this paper we examined several practical aspect of nested parallel loop execution. We used four different strategies for executing nested parallel loops on the examples of the WZ factorization. All proposed approaches usually accelerate sequential computations, except the *nested* algorithm.

*Nested* algorithm for a small number of threads proved to be the fastest, but for a big numer of threads its execution took longer time even than for a *sequential* algorithm. We may explain that creating any parallel region will cause the overhead. Overhead from nesting of parallel regions may cause overheads greater than necessary if, for example, an outer region could simply employ more threads in a computation. The appliaction lost the time on scheduling threads. OpenMP allows the specification of nested parallel loops, but for WZ factorization does not acquire satisfactory results. OpenMP uses nesting poorly.

The available number of threads exploited both outer and *split* algorithms best. *Split* approach achievs the best speedup. The speedup of 14 was achived for 16 threads on the E5-2660 system. We find this result very satisfactory.

The implementation had no impact on the accuracy of the factorization — the accuracy depended only on the size of the matrix what is quite self-evident.

The implementation of the *split* algorithm presented in this paper achieves high performance results, which has a direct



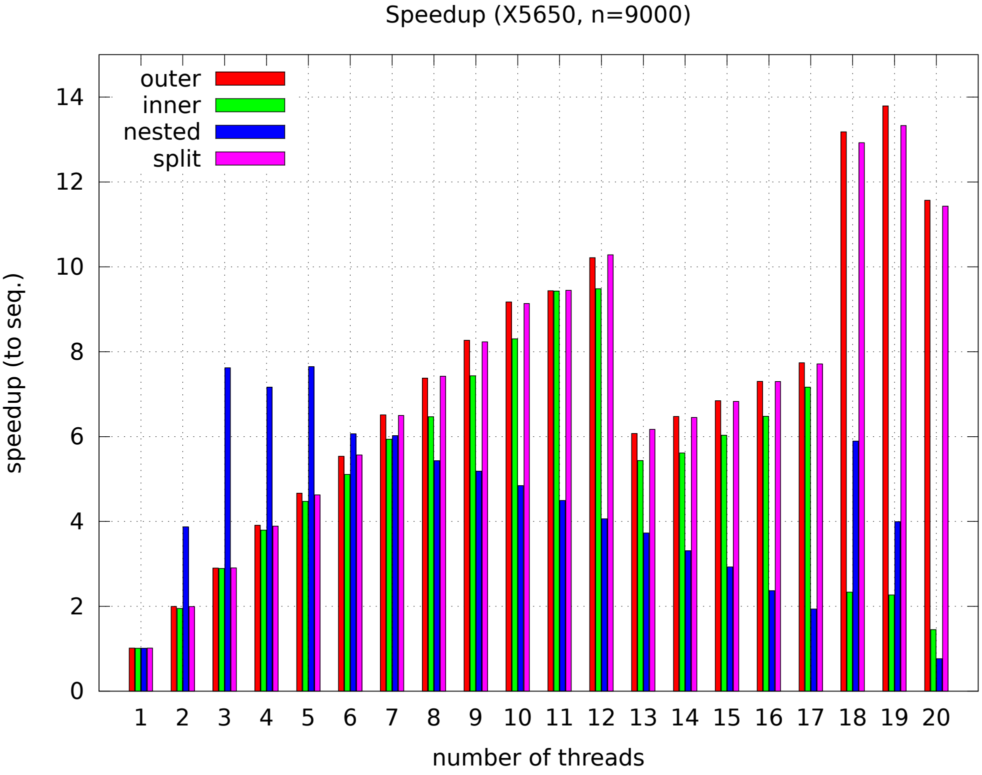


Fig. 7. The speedup results for the WZ factorization — using the double precision on two platforms (E5-2660 — top; X5650 — bottom) for the four algorithms as the function of number of threads.

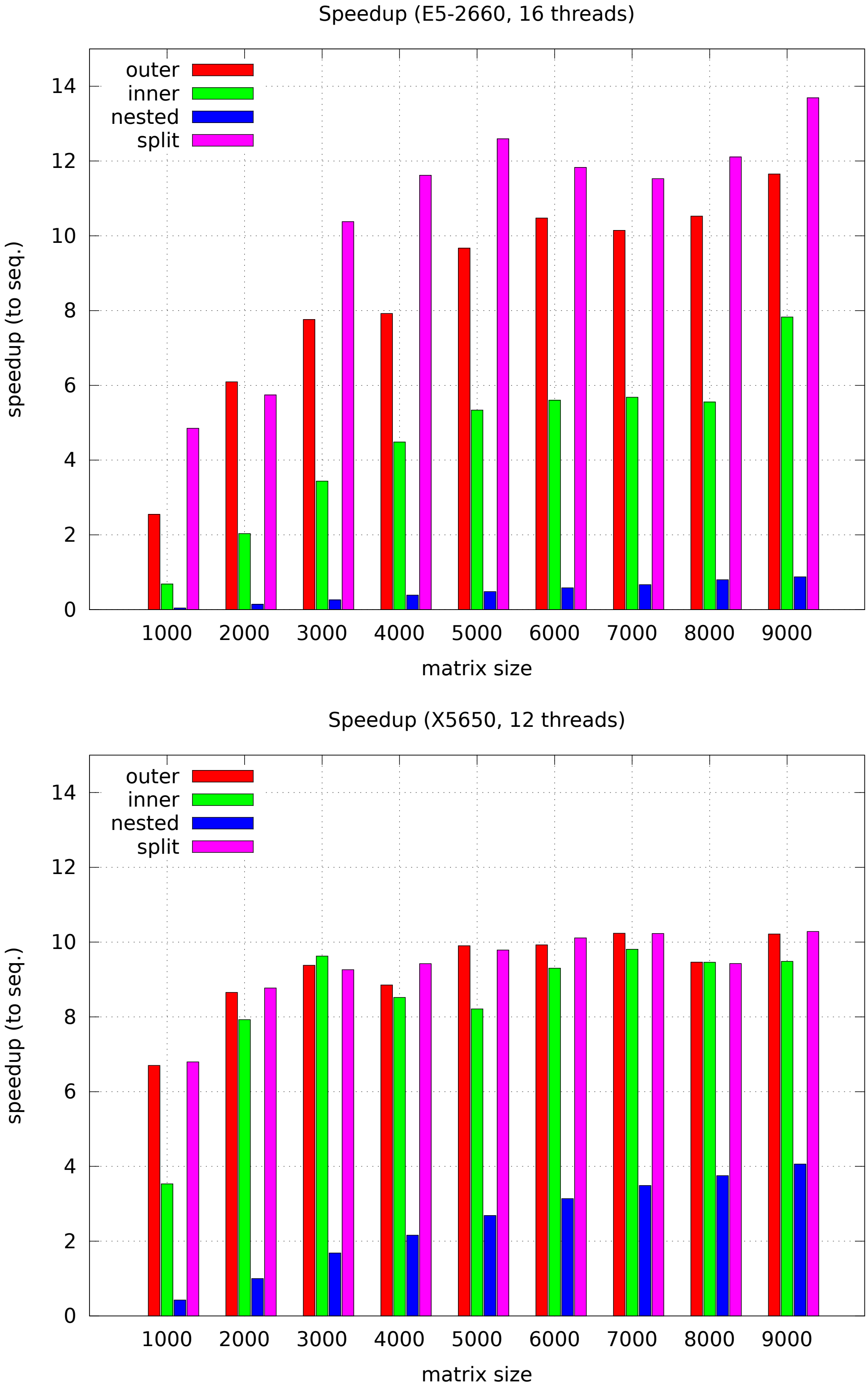


Fig. 8. The speedup results for the WZ factorization — using the double precision for 16 threads on E5-2660 System (top) and 12 threads on X5650 System (bottom) for the four algorithms as a function of the matrix size.

TABLE II
THE NORMS FOR THE WZ FACTORIZATIONS IN DOUBLE PRECISION ON E5-2660 SYSTEM AND X5650 SYSTEM FOR ALL THE ALGORITHMS IN DOUBLE PRECISION

| matrix size | $\|\|\mathbf{A}-\mathbf{WZ}\|\|$ |
|---|---|
| 1000 | $2.89 \cdot 10^{-22}$ |
| 2000 | $1.18 \cdot 10^{-21}$ |
| 3000 | $2.97 \cdot 10^{-21}$ |
| 4000 | $5.59 \cdot 10^{-21}$ |
| 5000 | $9.32 \cdot 10^{-21}$ |
| 6000 | $1.40 \cdot 10^{-20}$ |
| 7000 | $2.06 \cdot 10^{-20}$ |
| 8000 | $2.83 \cdot 10^{-20}$ |
| 9000 | $3.78 \cdot 10^{-20}$ |

impact on the solution of linear systems.

This paper is another example of the succesful use of OpenMP for solving scientific appliactions.

## REFERENCES

[1] R. Blikberg, T. Sørevik: "Load balancing and OpenMP implementation of nested parallelism", *Parallel Computing* 31, Elsevier, 2005, pp. 984–998.

[2] S. Chandra Sekhara Rao: "Existence and uniqueness of WZ factorization", *Parallel Computing* 23, (1997), pp. 1129–1139.

[3] A. Duran, R. Silvera, J. Corbalan, J. Labarta: "Runtime adjustment of parallel nested loops", *Proceedings of the 5th international conference on OpenMP Applications and Tools: shared Memory Parallel Programming with OpenMP*, Houston, 2004, pp. 137–147.

[4] D. J. Evans, M. Hatzopoulos: "The parallel solution of linear system", *Int. J. Comp. Math.* **7** (1979), pp. 227–238.

[5] A. Jackson, O. Agathokleous: "Dynamic Loop Parallelisation", arXiv: 1205.2367v1, 10 May 2012.

[6] A. Sadun, W. W. Hwu: "Executing nested parallel loops on shared-memeory multiprocessors", *Proceedings of the 21st Annual International Conference on Parallel Processing*, 1992.

[7] P. Yalamov, D. J. Evans: "The WZ matrix factorization method", *Parallel Computing* 21, 1995, pp. 1111–1120.

[8] OpenMP, http://openmp.org/wp/, April 2015.

DOI: 10.15439/2015F286
ACSIS, Vol. 5

# A $K$-iterated scheme for the first-order Gaussian recursive filter with boundary conditions

Salvatore Cuomo
University of Naples Federico II
Department of Mathematics and Applications "R. Caccioppoli", Italy
Email: salvatore.cuomo@unina.itl

Ardelio Galletti
University of Naples "Parthenope"
Department of Science and Technology, Italy
Email: ardelio.galletti@uniparthenope.it

Raffaele Farina
Institute for high performance computing and networking
CNR, Italy
Email: raffaele.farina@na.icar.cnr.it

Livia Marcellino
University of Naples "Parthenope"
Department of Science and Technology, Italy
Email: livia.marcellino@uniparthenope.it

***Abstract*—Recursive Filters (RFs) are a well known way to approximate the Gaussian convolution and are intensively used in several research fields. When applied to signals with support in a finite domain, RFs can generate distortions and artifacts, mostly localized at the boundaries of the computed solution. To deal with this issue, heuristic and theoretical end conditions have been proposed in literature. However, these end conditions strategies do not consider the case in which a Gaussian RF is applied more than once, as often happens in several realistic applications. In this paper, we suggest a way to use the end conditions for such a $K$-iterated Gaussian RF and propose an algorithm that implements the described approach. Tests and numerical experiments show the benefit of the proposed scheme.**

## I. Introduction

Recursive filters (RFs) have achieved a central role in several research fields, as in data assimilation for operational three-dimensional variational analysis schemes (3Dvar) [4], [10], and in Electrocardiogram (ECG) denoising [1], [2]. Among RFs, the Gaussian RFs are particularly suitable for digital image processing [13] and applications of the scale-space theory [8], [15]. Gaussian RFs are an efficient computational tool for approximating Gaussian-based convolutions [3], [14], [10], [11], [12]. Gaussian RFs are mainly derived in three different ways: the Deriche strategy uses an approximation of the Gaussian function in the space domain [5]; the approximation procedure of Jin et al. is carried out in the $z$-domain, i.e. it is based on an approximation of the $z$-transform of the Gaussian function by means of rational polynomial functions [9]; and the approach followed by Vliet et al. [12], [16] approximates the Gaussian function in the Fourier domain. The latter approach is particularly attractive since the Fourier transform of a Gaussian function is a Gaussian.

From a mathematical point of view, a Gaussian RF consists of two infinite sequences of equations (forward and backward equations) that involve the entries of both the input and output signals. However, algorithms that implement RFs need to consider only a finite number of these equations, i.e. they take into account a finite number of input and output signal samples. Without additional assumptions, such a reduction (from the infinite to the finite) introduces distortions and artifacts on the computed finite output signals. In particular, these solutions present an error that affects the entries near to the boundaries. This phenomenon has been recognized as an edge effect in [3] and some authors, such as Purser et al. [10] and Triggs et al. [14], suggest how to avoid this by simulating the effect of the continuation of the neglected equations. This results in inserting the so-called boundary conditions, or end conditions, i.e. in modifying some of the filter equations so that the backward equations can be primed. These strategies are based on some heuristic assumptions on the input signal and seem to fix the edge effect problem.

In some real applications [3], [6], [7], Gaussian RFs are used iteratively. This relies on a more general definition of RFs, named $K$-iterated RFs, which have been formally introduced in [3]. So far, a comprehensive study of the conditions for the $K$-iterated RFs is still lacking. Moreover, as noticed in [3], and recalled in this work (see Section 3), edge effects reappear when the number $K$ of filter iterations increases, even though the classic RF end conditions are used. In this context, the purpose of this work is to provide an algorithm which combines classic end conditions with a suitable oversizing and reduction of both the input and the output signals, so that the edge effects are strongly mitigated. The paper is organized as follows. In Section 2, we recall the definitions of the discrete Gaussian convolution and Gaussian recursive filters, then we derive classic end conditions in a general way and show their specific features for the first-order Gaussian RF used in [3], [6], [7]. The benefit of the end conditions is shown through some numerical examples. In Section 3, we give the definition of the $K$-iterated RFs and point out how the filter coefficients of a $K$-iterated first-order RF have to be taken. Finally, we propose a numerical scheme that implements a $K$-iterated first-order RF with end

978-83-60810-66-8/$25.00©2015, IEEE

conditions and a strategy for preventing the occurrence of edge effects. In Section 4, we report some experiments to confirm the reliability of the proposed algorithm and give suggestions on how to set the parameters of the proposed $K$-iterated scheme. Finally, Section 5 contains some concluding remarks.

## II. Mathematical Background

Let

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

be the Gaussian function and

$$s^{(0)} = \{s_j^{(0)}\}_{j=-\infty}^{+\infty} = \left(\dots, s_{-2}^{(0)}, s_{-1}^{(0)}, s_0^{(0)}, s_1^{(0)}, s_2^{(0)}, \dots\right)$$

a signal (i.e. a complex function defined on the set of integers **Z**). The discrete Gaussian convolution of $s^{(0)}$ gives rise to an output signal $s^{(g)}$ whose entries are defined by:

$$s_j^{(g)} \overset{\text{def}}{=} \left(g * s^{(0)}\right)_j = \sum_{t=-\infty}^{+\infty} g(j-t) s_t^{(0)}. \qquad (1)$$

The entries $s_j^{(g)}$ of $s^{(g)}$ in (1) can be approximated by means of Gaussian recursive filters. The $n$-order Gaussian RF filter gives an output signal $s$, which is an approximation of $s^{(g)}$, whose entries solve the infinite sequences of equations:

$$p_j = \beta_j s_j^{(0)} + \sum_{t=1}^{n} \alpha_{j,t} p_{j-t}, \qquad j = -\infty, \dots, +\infty, \qquad (2)$$

$$s_j = \beta_j p_j + \sum_{t=1}^{n} \alpha_{j,t} s_{j+t}, \qquad j = -\infty, \dots, +\infty. \qquad (3)$$

The equations in (2) and (3) are conveniently referred to as the advancing and backing filters, respectively, since in the former the index $j$ must be treated as in increasing order while, in the latter, it must be treated in decreasing order [10]. The values $\alpha_{j,t}$ and $\beta_j$ are called *smoothing coefficients* and satisfy:

$$\beta_j = 1 - \sum_{t=1}^{n} \alpha_{j,t}.$$

The smoothing coefficients depend on $\sigma$, $n$ and, in a more general setting, even on the index $j$. Hereafter, we only consider the homogeneous case, i.e. we set:

$$\beta_j \equiv \beta, \qquad \alpha_{j,t} \equiv \alpha_t, \qquad (4)$$

where the advancing and backing filters take the form:

$$p_j = \beta s_j^{(0)} + \sum_{t=1}^{n} \alpha_t p_{j-t}, \qquad j = -\infty, \dots, +\infty, \qquad (5)$$

$$s_j = \beta p_j + \sum_{t=1}^{n} \alpha_t s_{j+t}, \qquad j = -\infty, \dots, +\infty. \qquad (6)$$

If we assume that the support of the input signal $s^{(0)}$ is in the grid $\{1, 2, \dots, N\}$, then equations (5) and (6) can be implemented in an algorithm in which the index $j$ increases from 1 to $N$, for the advancing filter, and decreases from $N$ to 1, for the backing filter. This scheme needs to prime the advancing filter, by setting the values $p_j$, for $j \in \{0, -1, \dots, 1-n\}$, and the backing filter, by setting the values $s_j$ for $j \in \{N+1, N+2, \dots, N+n\}$. For example, a common choice is to set at zero the required $p_j$ and $s_j$ values, i.e.:

$$\begin{aligned} p_{-n+1} = p_{-n+2} = \dots = p_0 = 0; \\ s_{N+1} = s_{N+2} = \dots = s_{N+n} = 0. \end{aligned} \qquad (7)$$

However, this assumption gives rise to a well-known edge effect, already noticed in [14], and discussed in detail in [3]. An outline of such a scheme, implementing (5), (6) and (7), is provided in **Algorithm 1**.

**Algorithm 1** Scheme of an $n$-order Recursive Filter with zero end constraints

```
Input: $s^{(0)}$, $\sigma$    Output: $s$
set $\beta, \alpha_1, \dots, \alpha_n$        % smoothing coefficient precomputation
for $j = 1, 2, \dots, n$        % left zero end conditions
    $p_{j-n} := 0$
endfor
for $j = 1, 2, \dots, N$        % advancing filter
    $p_j := \beta s_j^{(0)}$
    for $t = 1, 2, \dots, n$
        $p_j := p_j + \alpha_t p_{j-t}$
    endfor
endfor
for $j = 1, 2, \dots, n$        % right zero end conditions
    $s_{N+j} := 0$
endfor
for $j = N, \dots, 1$        % backing filter
    $s_j := \beta p_j$
    for $t = 1, 2, \dots, n$
        $s_j := s_j + \alpha_t s_{j+t}$
    endfor
endfor
```

**Example 1.** Now we provide some insights into the edge effect through the following example. Let $s^{(0)}$ be the input signal with entries:

$$s_j^{(0)} = \begin{cases} 0 & \text{for } j \leq 0; \\ 1 & \text{for } 1 \leq j \leq N = 30; \\ 0 & \text{for } j \geq N+1 = 31; \end{cases} \qquad (8)$$

In Figure 1, we report the results obtained by applying the first-order RF and third-order RF to the signal $s^{(0)}$. In particular, the red line represents the signal $s^{(g)}$, i.e. the discrete Gaussian convolution of $s^{(0)}$ with $\sigma = 4$. Conversely, the white squares and gray diamonds are the values of the output signals computed by means of the first-order RF and third-order RF, respectively. Notice that both computed solutions differ from $s^{(g)}$ mostly on the boundary entries. In particular, the third-order filter seems to not well approximate the entries of $s^{(g)}$ close to the right boundary. This phenomenon can be better

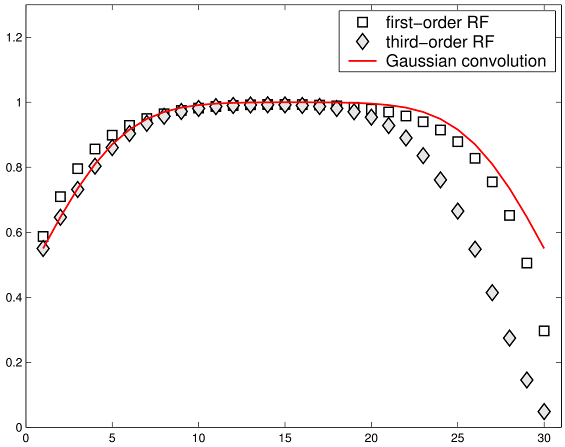


Fig. 1. Red line: discrete Gaussian. White squares: first-order RF solution. Gray diamonds: third-order RF solution.

understood by looking at the relative errors

$$e_j = \frac{|s_j - s_j^{(g)}|}{|s_j^{(g)}|}. \tag{9}$$

The value of the errors $e_j$, for the first-order RF and third-order RF, are shown in Figure 2. Observing the order of magnitude of the values $e_j$, we can conclude that, in general, the output signals present a larger error at the boundaries.

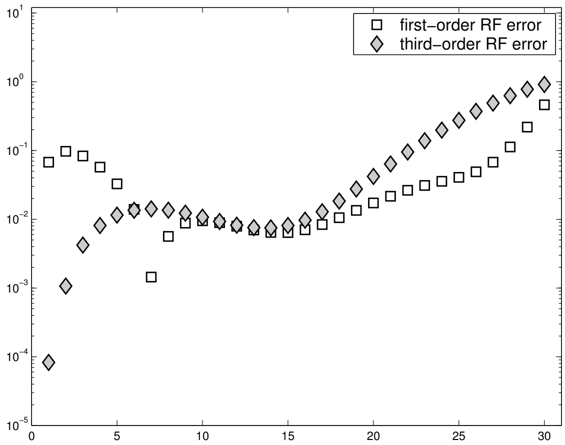


Fig. 2. White squares: first-order RF relative errors. Gray diamonds: third-order RF relative errors.

### A. *Introducing end conditions*

The edge effect, shown in the previous example, can be partly explained by observing that, in the transition from the infinite sequences of equations to the finite algorithm, the computed solution $s$ has been constrained to assume the value zero on the right off-grid points $N+1, N+2, \ldots, N+n$, while this property is not true for $s^{(g)}$. Hence, assumption (7) introduces a sort of perturbation error on the output signal $s$ of the RF. This drawback can be avoided by simulating, in the finite setting, the effect of the continuation of the neglected equations (for $j > N$). To achieve this aim, we adapt to our setting and notations the derivation of the end conditions (e.c.) described in [14].

We consider an $n$-order recursive filter with smoothing coefficients $\alpha_1, \ldots, \alpha_n$ and $\beta_j$. Let $C$ and $A$ be the following $n \times n$ matrices:

$$\mathbf{A} = \begin{pmatrix} \alpha_1 & \ldots & \alpha_{n-1} & \alpha_n \\ 1 & \ldots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 1 & 0 \end{pmatrix} \tag{10}$$

$$\mathbf{C} = \begin{pmatrix} \beta & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 \end{pmatrix} \tag{11}$$

Now, by setting:

$$\begin{aligned} \vec{p}_t &= \left(p_t, p_{t-1}, \ldots, p_{t-n+1}\right)^T, \\ \vec{s}_t &= \left(s_t, s_{t+1}, \ldots, s_{t+n-1}\right)^T, \\ \vec{s}_t^{(0)} &= \left(s_t^{(0)}, s_{t+1}^{(0)}, \ldots, s_{t+n-1}^{(0)}\right)^T, \end{aligned}$$

the forward and backward filters in (5) and (6), take the form:

$$\vec{p}_{t+1} = C\vec{s}_{t+1}^{(0)} + A\vec{p}_t, \tag{12}$$

$$\vec{s}_t = C\vec{p}_t + A\vec{s}_{t+1}. \tag{13}$$

Then, applying recursively (12) and (13), we obtain:

$$\vec{p}_{t+k} = \sum_{l=0}^{k-1} A^l C \vec{s}_{t+k-l}^{(0)} + A^k \vec{p}_t, \tag{14}$$

$$\vec{s}_t = \sum_{l=0}^{k-1} A^l C \vec{p}_{t+l} + A^k \vec{s}_{t+k}. \tag{15}$$

Assuming that the support of the input signal $s^{(0)}$ is in $\{1, 2, \ldots, N\}$, the equation (14), for $k > 0$ and $t \geq N$, becomes:

$$\vec{p}_{t+k} = A^k \vec{p}_t. \tag{16}$$

Now, using (16) in (15), and taking the limit for $l \to \infty$, provided that $s$ is bounded, we obtain:

$$\vec{s}_t = M\vec{p}_t, \qquad \text{with} \qquad M \overset{\text{def}}{=} \sum_{l=0}^{\infty} A^l C A^l. \tag{17}$$

For $t = N$, equation (17) provides a way of priming the backing filter. In other words, (17) behaves as a sort of *turning* condition that accounts for the infinite sequence of neglected equations and allows us to skip from the advancing filter to the backing filter. Indeed, the values $s_N, s_{N+1}, \ldots, s_{N+n-1}$ are linear functions of the values $p_N, p_{N-1}, \ldots, p_{N-n+1}$ which are computed by the advancing filter. In this approach, we need to precompute the matrix $M$. This can be done in several ways: in [14], it is observed that the recursive definition of $M$ implies that it solves the equation:

$$M = C + AMA;$$

otherwise, one could compute an approximation of $M$ as the partial sum $M_k = \sum_{l=0}^{k} A^l C A^l$, provided that $A^{k+1} C A^{k+1}$ is negligible. Observe that, for the first-order RF with smoothing coefficients $\alpha \equiv \alpha_1$ and $\beta = 1 - \alpha$, we simply have:

$$A = \alpha, \quad C = \beta, \quad M = \sum_{l=0}^{\infty} \beta \cdot \alpha^{2l} = \frac{\beta}{1-\alpha^2} = \frac{1}{1+\alpha}$$

and for $t = N$ in (17),

$$s_N = \frac{1}{1+\alpha} p_N.$$

An exact expression of the matrix $M$, in terms of the coefficients $\alpha_1$, $\alpha_2$, $\alpha_3$ and $\beta$, can be derived for the third-order RF [14] and will be used in the example below. In order to take into account the end conditions, steps 11. 12. and 13. of **Algorithm 1** must be replaced with the computation of $M$ and the computation of $\vec{s}_N$. Moreover, the last loop (steps 14.-19.) must start from $N - 1$.

**Example 2.** To see the effect of the e.c. on the RFs, we consider the input signal $s^{(0)}$ in (8). The results obtained by applying the first-order and the third-order RFs, with and without e.c., are plotted in Figure 3 and Figure 4, respectively.

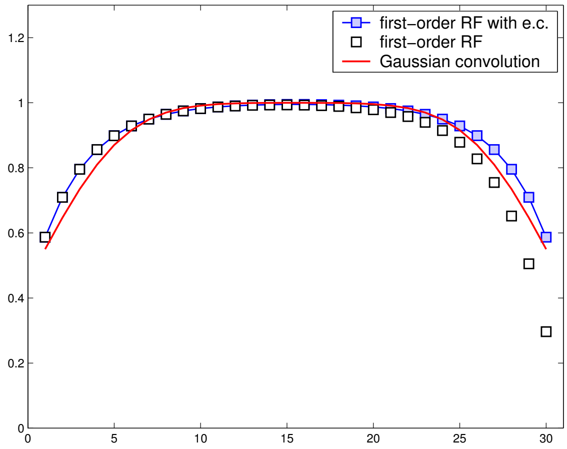


Fig. 3. Red line: discrete Gaussian. White squares: first-order RF without end conditions. Blue squares: first-order RF with end conditions.

In Figure 3 the solutions of the first-order RFs with and without e.c. are compared; analogously, Figure 4 deals with the third-order RFs. In both cases, one can notice the improvement of the computed solutions, especially on their right boundary entries, when the e.c. are considered. In particular, the output signal of the third-order RF with e.c. is very close to the Gaussian convolution output $s^{(g)}$.

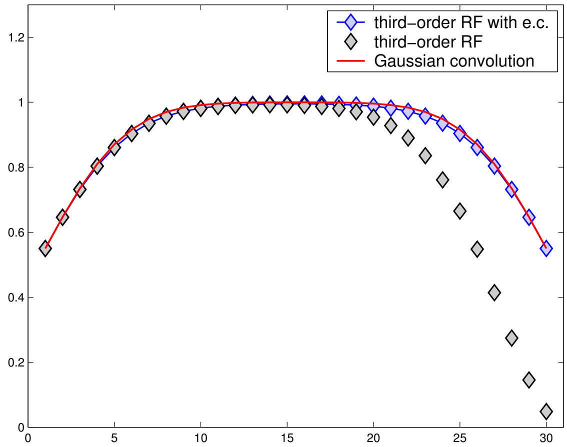


Fig. 4. Red line: discrete Gaussian. Gray diamonds: third-order RF without end conditions. Blue diamonds: third-order RF with end conditions.

## III. $K$-ITERATED GAUSSIAN RECURSIVE FILTERS

The latter example highlights that the third-order filter mimics very well the effect of the discrete Gaussian convolution. Conversely, the output signal of the first-order RF, even with e.c., is still a poor approximation of $s^{(g)}$. Purser et al. [10] suggest that applying several times the same RF can improve the accuracy of the approximation. This fact could be seen as a consequence of the central limit theorem applied to the Gaussian convolution. This idea has been discussed and implemented in [3], [6] and [7] where the authors iteratively used the first-order RF (without e.c.) to solve a three-dimensional variational data assimilation problem. Here, we are interested in formalizing such an iterative approach.

A $K$-iterated $n$-order Gaussian RF filter computes the output signal $s^{(K)}$, i.e. the $K$-iterated approximation of $s^{(g)}$, as follows:

$$p_j^{(k)} = \beta s_j^{(k-1)} + \sum_{t=1}^{n} \alpha_t p_{j-t}^{(k)}, \qquad j = -\infty, \ldots, +\infty, \quad (18)$$

$$s_j^{(k)} = \beta p_j^{(k)} + \sum_{t=1}^{n} \alpha_t s_{j+t}^{(k)}, \qquad j = -\infty, \ldots, +\infty. \quad (19)$$

The filter iteration counter $k$ goes from 1 to $K$, where $K$ is the total number of filter iterations. For $K > 1$, the problem of triggering the advancing filter at iteration $k = 2, \ldots, K$ can be faced with the same strategy used to generate the (right) end conditions for a classic RF ($K = 1$). Assuming indeed that, at the iteration $k$:

$$\begin{aligned} p_{-n+1}^{(k)} &= p_{-n+2}^{(k)} = \ldots = p_0^{(k)} = 0, \\ s_{N+1}^{(k)} &= s_{N+2}^{(k)} = \ldots = s_{N+n}^{(k)} = 0, \end{aligned} \quad (20)$$

and rearranging the equations (12) and (13), it results:

$$\vec{p}_1^{(k)} = M \vec{s}_1^{(k-1)}, \quad (21)$$

with $M$ as in (17). Similarly, the backing filter, for iterations $k = 2, \ldots, K$, can be triggered by setting

$$\vec{s}_N^{(k)} = M \vec{p}_N^{(k)}. \quad (22)$$

The smoothing coefficients depend on $\sigma$, $n$ and even on $K$. The correct setting of the smoothing coefficients is crucial for the convergence. We remark that, in the Fourier domain, the Gaussian convolution in (1) becomes

$$S^{(g)}(\omega) = G(\omega, \sigma) \cdot S^{(0)}(\omega), \tag{23}$$

where

$$S^{(g)}(\omega) = \Big(\mathcal{F}\big(s^{(g)}\big)\Big)(\omega), \qquad S^{(0)}(\omega) = \Big(\mathcal{F}\big(s^{(0)}\big)\Big)(\omega)$$

and

$$G(\omega, \sigma) = \Big(\mathcal{F}(g)\Big)(\omega) = \exp\left(-\frac{\omega^2\sigma^2}{2}\right)$$

are the Fourier Transforms of the signals $s^{(g)}$, $s^{(0)}$ and $g$, respectively. Rewriting (23) as:

$$S^{(g)}(\omega) = \left(G\left(\omega, \frac{\sigma}{\sqrt{K}}\right)\right)^K \cdot S^{(0)}(\omega), \tag{24}$$

we obtain, in the signal domain, that $s^{(g)}$ can be seen as the result of $K$ successive Gaussian convolutions with an identical Gaussian function with a standard deviation:

$$\sigma_K = \frac{\sigma}{\sqrt{K}}.$$

This argument suggests that $\sigma_K$ must replace $\sigma$ in the expression of the smoothing coefficients of a $K$-iterated RF. For example, the smoothing coefficients $\alpha$ and $\beta$ of the (one-iterated) first-order Gaussian RF, in the homogeneous case, are set as:

$$\alpha = 1 + E_\sigma - \sqrt{E_\sigma(E_\sigma + 2)}, \qquad \beta = \sqrt{E_\sigma(E_\sigma + 2)} - E_\sigma. \tag{25}$$

with

$$E_\sigma = \frac{1}{\sigma^2}.$$

Then, for the $K$-iterated first-order Gaussian RF, the value of $E_\sigma$ must be replaced by:

$$E_{\sigma_K} = \frac{1}{\sigma_K^2} = \frac{K}{\sigma^2}.$$

**Example 3.** To see the behaviour of such a $K$-iterated filter, we apply it to a random input signal with support in $\{1, 2, \dots, 30\}$. In Figure 5, the RF output signals for three different values of the number iterations ($K = 12, 25, 50$) are reported. The results show that, despite the fact that the accuracy on the central entries does not seem to change, the edge effects reappear when the number of filter iterations increases. In particular (see Figure 6), the relative errors $e_j$, defined as in (9), increase both in the left and right boundary entries as $K$ increases. Conversely, the convergence to the Gaussian convolution output signal may be observed by looking at the central entries of the errors in Figure 6. We highlight that, in the previous test, the end conditions (21) and (22) have been used. The fact that the e.c. do not work as expected, mostly at the edges, can be simply explained as a consequence of the wrong assumptions in (20).

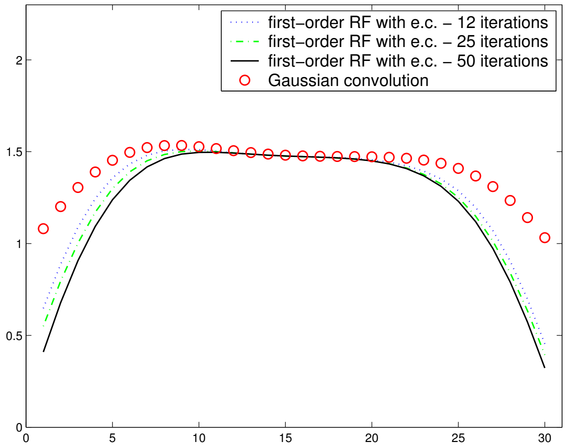


Fig. 5. Discrete Gaussian (red circles) versus first-order RF solutions with 12 iterations (dotted blue line), 25 iterations (dashdot green line) and 50 iterations (solid black line).

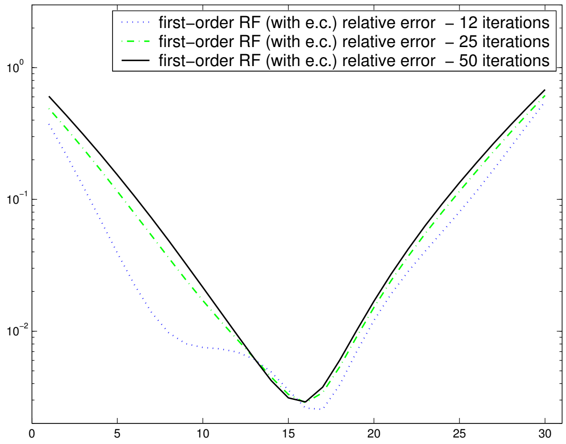


Fig. 6. First-order RF relative errors for 12 iterations (dotted blue line), 25 iterations (dashdot green line) and 50 iterations (solid black line).

Here, we suggest a scheme that allows the use of the $K$-iterated first-order Gaussian RF and prevents the edge effect. Our idea consists in three steps:

(i) *extending* the given input signal $s^{(0)}$, with support in $\{1, 2, \dots, N\}$, by adding artificial zero entries at the left and right boundaries. More specifically, we introduce the extended signal:

$$s^{(0),m} = \big(0, \dots, 0, s_1^{(0)}, \dots, s_N^{(0)}, 0, \dots, 0\big), \tag{26}$$

which is obtained by placing $m$ zeros before $s_1^{(0)}$ and $m$ zeros after $s_N^{(0)}$;

(ii) applying the $K$-iterated first-order Gaussian RF to $s^{(0),m}$;

(iii) reducing the output signal $s^{(K),m}$, by removing its first and last $m$ entries.

The underlying idea of that scheme is to shift the edge effects on the artificially added entries. Steps (i)-(iii) are summarized in the following algorithm. For the sake of simplicity, we consider the case of the first-order RF. Nevertheless, the same algorithm could be easily modified for any $K$-iterated $n$-order RF.

**Algorithm 2** Scheme of the $K$-iterated first-order recursive filter with end conditions

```
Input: $s^{(0)}, \sigma, m, K$      Output: $s^{(K)}$
extend $s^{(0)}$ to $s^{(0),m}$
set $s^{(0)} := s^{(0),m}$; $\beta, \alpha$ as in (25); $M := 1/(1+\alpha)$
for $k = 1, 2, \dots, K$                          % filter loop
    compute $p_1^{(k)} := M s_1^{(k-1)}$          % left end conditions
    if $k = 1$ then
        $p_1^{(k)} := \beta s_1^{(k-1)}$
    end
    for $j = 2, \dots, N$                         % advancing filter
        $p_j^{(k)} := \beta s_j^{(k-1)} + \alpha p_{j-1}^{(k)}$
    endfor
    compute $s_N^{(k)} := M p_N^{(k)}$            % right end conditions
    for $j = N-1, \dots, 1$                       % backing filter
        $s_j^{(k)} := \beta p_j^{(k)} + \alpha s_{j+1}^{(k)}$
    endfor
endfor
reduce $s^{(K)}$ as described in step (iii)
```

In the next section, through some numerical examples, it will be pointed out how the parameter $m$ has to be set, in order to obtain a satisfactory accuracy on the computed RF output signals, even at their boundary entries.

## IV. Numerical experiments

In this section, we present two experiments. The aim of the first test is to prove the effectiveness of **Algorithm 2** in removing the edge effects. In the second test, we show that a suitable value of $m$ can guarantee a negligible edge error, regardless of the filter number iterations $K$ and the size $N$ of the input signal.

### A. *Convergence at the edges*

We consider the random input signal of **Example 3.**. We use **Algorithm 2** varying both the number of iterations and the size of $s^{(0),m}$. More precisely, in Figure 7, we report the output signals obtained with $K = 12$ iterations and by extending the input signal $s^{(0)}$ with $m = 1, 3, 6$ zeros at each boundary. Figure 8 shows the results obtained with the same values of $m$ and increasing the number of iteration to $K = 25$.

Looking at Figure 7 and Figure 8, we observe that the edge effects seem to disappear as $m$ increases. Notice that, comparing the behaviour of the blue curves ($m = 1$), the results worsen as $K$ increases. This drawback can be avoided by suitably increasing the value of $m$. Figure 9 shows pictorially that, taking $m = 3\sigma = 12$, the accuracy is preserved also at the boundaries. Moreover, Figure 10 indicates that the component-wise convergence holds as $K$ increases.

### B. *Suitable choice of* $m$

In this test we measure the error $\|s^{(g)} - s^{(K)}\|_2$, where $s^{(K)}$ is the output signal obtained by applying the **Algorithm 2** to the random input signal $s^{(0)}$ of **Example 3.** varying the values of $m$ and $K$. The results summarized in Table I, for $\sigma = 4$, show that the convergence is reached, as $K$ increases, provided that $m \geq 2\sigma$. For $m \leq \sigma$ the convergence is not achieved due to the edge effects. We also note that increasing $m$ beyond $3\sigma$ does not improve significantly the accuracy.

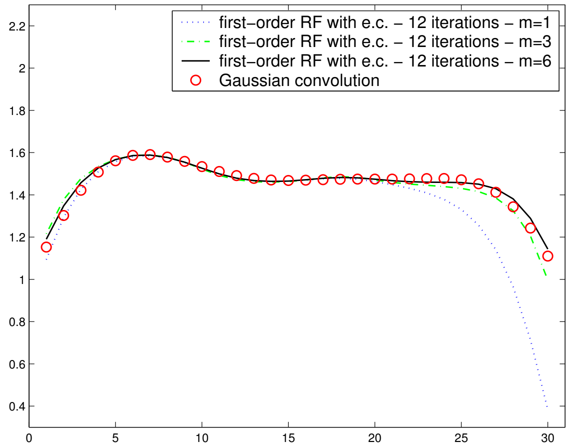


Fig. 7. Discrete Gaussian (red circles) versus first-order RF solutions with 12 iterations and $m = 1$ (dotted blue line), $m = 3$ (dashdot green line) and $m = 6$ (solid black line).

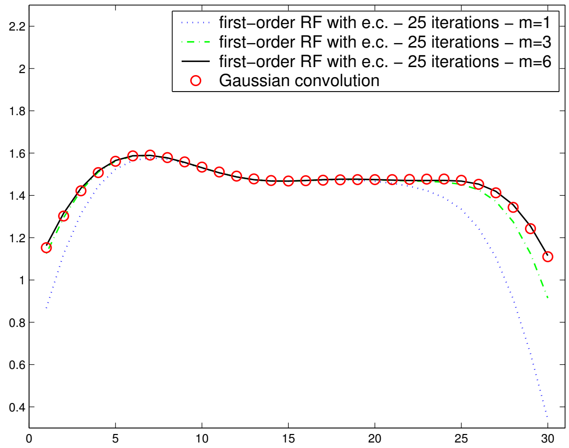


Fig. 8. Discrete Gaussian (red circles) versus first-order RF solutions with 25 iterations and $m = 1$ (dotted blue line), $m = 3$ (dashdot green line) and $m = 6$ (solid black line).

TABLE I
NORMS $\|s^{(g)} - s^{(K)}\|_2$ FOR SEVERAL VALUES OF $K$ AND $m$.
$\sigma = 4, N = 30$.

| $K \backslash m$ | $0.25\sigma$ | $0.5\sigma$ | $\sigma$ | $2\sigma$ | $3\sigma$ | $6\sigma$ |
|---|---|---|---|---|---|---|
| 5 | 9.85e-01 | 4.70e-01 | 1.00e-01 | 1.23e-01 | 1.42e-01 | 1.16e-01 |
| 15 | 1.14e+00 | 5.66e-01 | 7.75e-02 | 4.96e-02 | 6.00e-02 | 5.73e-02 |
| 30 | 1.19e+00 | 6.07e-01 | 8.14e-02 | 3.41e-02 | 3.48e-02 | 3.00e-02 |
| 50 | 1.23e+00 | 6.18e-01 | 8.61e-02 | 2.48e-02 | 2.49e-02 | 2.54e-02 |
| 100 | 1.26e+00 | 6.40e-01 | 9.09e-02 | 1.71e-02 | 1.70e-02 | 1.70e-02 |

Table II shows that the above results about convergence and accuracy hold true also for different values of $\sigma$. We remark

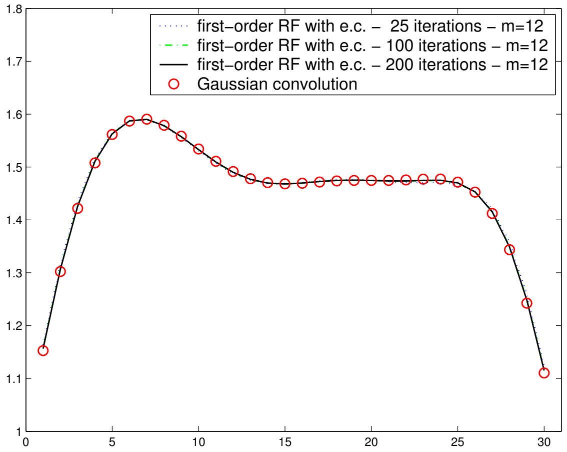


Fig. 9. Discrete Gaussian (red circles) versus first-order RF solutions with 25 iterations (dotted blue line), 100 iterations (dashdot green line) and 200 iterations (solid black line). $m$ is set to 12 and the computed solutions can not be distinguished.

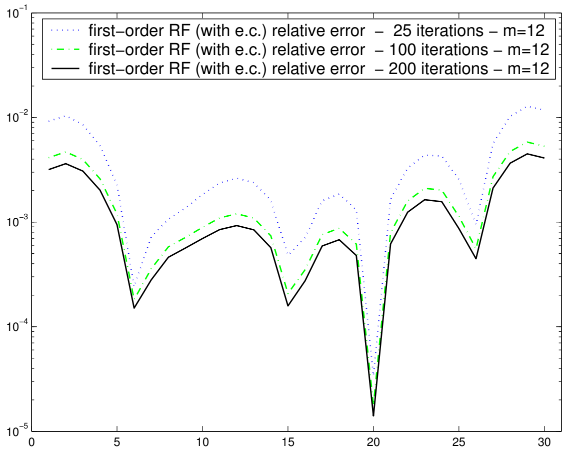


Fig. 10. First-order RF relative errors for 25 iterations (dotted blue line), 100 iterations (dashdot green line) and 200 iterations (solid black line).

that, for any fixed value of $K$, the value $m = 3\sigma$ represents a good trade-off between accuracy and efficiency.

TABLE II
NORMS $\|s^{(g)} - s^{(K)}\|_2$ FOR SEVERAL VALUES OF $K$ AND $m$.
$\sigma = 6, N = 30$.

| $K\backslash m$ | $0.25\sigma$ | $0.5\sigma$ | $\sigma$ | $2\sigma$ | $3\sigma$ | $6\sigma$ |
|---|---|---|---|---|---|---|
| 5 | 1.02e+00 | 4.88e-01 | 1.01e-01 | 1.13e-001 | 1.61e-001 | 1.18e-01 |
| 15 | 1.13e+00 | 5.63e-01 | 7.71e-02 | 4.57e-002 | 4.98e-002 | 5.43e-02 |
| 30 | 1.20e+00 | 6.02e-01 | 8.31e-02 | 3.24e-002 | 3.65e-002 | 3.48e-02 |
| 50 | 1.23e+00 | 6.21e-01 | 8.62e-02 | 2.31e-002 | 2.41e-002 | 2.30e-02 |
| 100 | 1.26e+00 | 6.42e-01 | 9.08e-02 | 1.64e-002 | 1.74e-002 | 1.65e-02 |

The results in Table III prove that the above conclusions do not depend on the size of the input signal.

## V. CONCLUSIONS

In this paper, we have discussed a way to overcome the edge effect in $K$-iterated Gaussian RFs, by a suitable choice of the end conditions. We have introduced an algorithm that implements the described approach. We have shown by means of several numerical experiments the effectiveness of the proposed $K$-iterated scheme. Finally, we have discussed in detail several issues related to the suitable choice of the parameter of our method.

TABLE III
NORMS $\|s^{(g)} - s^{(K)}\|_2$ FOR SEVERAL VALUES OF $K$ AND $m$.
$\sigma = 4, N = 2000$.

| $K\backslash m$ | $0.25\sigma$ | $0.5\sigma$ | $\sigma$ | $2\sigma$ | $3\sigma$ | $6\sigma$ |
|---|---|---|---|---|---|---|
| 5 | 1.13e+00 | 5.84e-01 | 3.07e-01 | 4.68e-01 | 2.87e-01 | 3.47e-01 |
| 15 | 1.20e+00 | 6.00e-01 | 1.39e-01 | 1.33e-01 | 1.25e-01 | 1.32e-01 |
| 30 | 1.25e+00 | 6.25e-01 | 1.17e-01 | 8.23e-02 | 9.08e-02 | 9.11e-02 |
| 50 | 1.29e+00 | 6.44e-01 | 1.03e-01 | 6.27e-02 | 5.90e-02 | 6.48e-02 |
| 100 | 1.32e+00 | 6.67e-01 | 1.01e-01 | 4.28e-02 | 4.29e-02 | 4.28e-02 |

## REFERENCES

[1] S. Cuomo, G. De Pietro, R. Farina, A. Galletti, and G. Sannino - *Novel O(n) Numerical Scheme for ECG Signal Denoising*. Procedia Computer Science , Volume 51, pp. 775784, doi: 10.1016/j.procs.2015.05.198, 2015.

[2] S. Cuomo, G. De Pietro, R. Farina, A. Galletti, and G. Sannino - *A framework for ECG denoising for mobile devices*. PETRA 15 ACM. ISBN 978-1-4503-3452-5/15/07, doi: 10.1145/2769493.2769560, 2015.

[3] S. Cuomo, R. Farina, A. Galletti, L. Marcellino -*An error estimate of Gaussian Recursive Filter in 3Dvar problem*, Federated Conference on Computer Science and Information Systems, FedCSIS 2014, doi: 10.15439/2014F279, pp. 587-595, 2014.

[4] L. D'Amore, R. Arcucci, L. Marcellino, A. Murli- *HPC computation issues of the incremental 3D variational data assimilation scheme in OceanVarsoftware*. Journal of Numerical Analysis, Industrial and Applied Mathematics, 7(3-4), pp. 91-105, 2013.

[5] R. Deriche - *Recursively implementing the Gaussian and its derivatives*. INRIA Research Report RR-1893, 1993.

[6] S. Dobricic, N. Pinardi - *An oceanographic three-dimensional variational data assimilation scheme*. Ocean Modeling 22, pp. 89-105, doi: 10.1016/j.ocemod.2008.01.004, 2008.

[7] Farina, R., Dobricic, S., Storto, A., Masina, S., Cuomo, S. -*A revised scheme to compute horizontal covariances in an oceanographic 3D-VAR assimilation system*. Journal of Computational Physics, 284, pp. 631-647, doi: 10.1016/j.jcp.2015.01.003, 2015.

[8] T. Lindeberg *Scale-space theory in computer vision*. Dordrecht: Kluwer Academic Publishers; 1994.

[9] J. S. Jin, Y. Gao - *Recursive implementation of Log Filtering*. Real-Time Imaging pp. 59-65, doi: 10.1006/rtim.1996.0045, 1997.

[10] R.J. Purser, W.-S. Wu, D.F. Parrish, N.M. Roberts - *Numerical Aspects of the Application of Recursive Filters to Variational Statistical Analysis. Part I: Spatially Homogeneous and Isotropic Gaussian Covariances*. Monthly Weather Review 131, pp. 1524-1535, doi: 10.1175//2543.1, 2003.

[11] C. Hayden, R. Purser - *Recursive filter objective analysis of meteorological field: applications to NESDIS operational processing*. Journal of Applied Meteorology 34, pp. 3-15, doi: 10.1175/1520-0450-34.1.3, 1995.

[12] L.V. Vliet, I. Young, P. Verbeek - *Recursive Gaussian derivative filters*. International Conference Recognition, pp. 509-514, doi: 10.1109/ICPR.1998.711192, 1998.

[13] L.J. van Vliet, P.W. Verbeek - *Estimators for orientation and anisotropy in digitized image*. Proc. ASCI'95, Heijen , pp. 442-450, 1995.

[14] B. Triggs, M. Sdika - *Boundary conditions for Young-van Vliet recursive filtering*. IEEE Transactions on Signal Processing, 54 (6 I), pp. 2365-2367, doi: 10.1109/TSP.2006.871980, 2006.

[15] A. Witkin - *Scale-space filtering*. Proc. Internat. Joint Conf. on Artificial Intelligence, Karlsruhe, germany, pp. 1019-1021, 1983.

[16] I.T. Young, L.J. van Vliet - *Recursive implementation of the Gaussian filter*. Signal Processing 44, pp. 139-151, doi: 10.1016/0165-1684(95)00020-E, 1995.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 649–654

DOI: 10.15439/2015F177
ACSIS, Vol. 5

# Using Random Butterfly Transformations in Parallel Schur Complement-Based Preconditioning

Marc Baboulin
Université Paris-Sud
91405 Orsay, France
Email: baboulin@lri.fr

Aygul Jamal
Université Paris-Sud
91405 Orsay, France
Email: jamal@lri.fr

Masha Sosonkina
Old Dominion University
Norfolk, VA, 23529, United States
Email: msosonki@odu.edu

***Abstract*—We propose to use a randomization technique based on Random Butterfly Transformations (RBT) in the Algebraic Recursive Multilevel Solver (ARMS) to improve the preconditioning phase in the iterative solution of sparse linear systems. We integrated the RBT technique into the parallel version of ARMS (pARMS). The preliminary experimental results on some matrices from the Davis' collection show an improvement of the convergence and accuracy of the results when compared with existing implementations of the pARMS preconditioner.**

## I. Introduction

With the evolution of recent computer architectures, the growing gap between communication and computation efficiency makes communication very expensive (at a cost of one communication we can generally perform thousands of arithmetical operations). This requires the rethinking of most of numerical libraries in order to take advantage of current parallel architectures which are commonly based on multicore processors [1], possibly with accelerators [2], such as Graphics Processing Units (GPU) or Intel Xeon Phi.

In this work we are concerned with the solution of linear systems $Ax = b$ where $A$ is an $n \times n$ real matrix (dense or sparse), $b$ is a real n-vector and $x$ is the n-vector of unknowns. This operation is at the heart of many applications in high-performance computing (HPC) and is usually solved using either direct or iterative methods.

Direct methods [3] usually solve a linear system of equations $Ax = b$ using factorization techniques depending on the properties of the original matrix $A$. For a general system, we compute an $LU$ factorization of $A$ that decomposes the input matrix $A$ into the product $L \times U$, where $L$ is a lower triangular matrix and $U$ is an upper triangular matrix. When $A$ is positive definite, then we decompose the matrix $A$ into the product $A = L \times L^T$ (Cholesky decomposition, which requires half the number of flops of the $LU$ factorization). In both cases ($LU$ or Cholesky), the solution is then obtained by solving successively 2 triangular systems.

Another possibility to solve $Ax = b$ is to use an iterative method to compute an approximate solution. These methods involve passing from one iteration to the next one by modifying one or a few components of an approximate vector solution at a time. Classical examples of iterative methods are the Jacobi, Gauss-Seidel, Successive Over-Relaxation (SOR), or Krylov subspace methods [4].

The Algebraic Recursive Multilevel Solver (ARMS) is one of the solvers which applies the iterative Krylov subspace methods in sparse linear systems, it relies on multilevel partial elimination. The preconditioning separates the entries into two parts, the first part called fine set which is composed of block independent set, and the second part called coarse set which contains the rest of the entries. The coarse set can be used to built the Schur complement, which allows us to perform a block $LU$ factorization. The inter-level $LU$ factorization can be built from the upper level $LU$ factorization and the fine set, up to the first level.

Parallel ARMS (pARMS) is a distributed-memory implementation of ARMS, which relies on distributed group independent sets. It provides a set of standard preconditioners such as Additive Schwartz, Schur complement and Block Jacobi, which allow to run performance tests.

When solving square linear systems $Ax = b$ using Gaussian elimination (e.g., in $LU$ factorization), we commonly use partial pivoting to avoid having zero or too-small numbers on the diagonal. This technique is implemented in current linear algebra libraries and ensures stability [5]. However, partial pivoting requires communication (search for pivots, swapping of rows). For example, on a hybrid CPU/GPU system, the $LU$ algorithm in the MAGMA library [2] spends more than 20% of the factorization time in pivoting even for a large random matrix of size 10, 000 × 10,000 [6].

As an alternative to pivoting, an approach based on randomization called Random Butterfly Transformation (RBT) [7] was recently revisited. Following the RBT method, $A$ is transformed into a matrix that would be sufficiently random to avoid pivoting (with a probability close to 1). RBT is a random transformation of $A$ which can avoid pivoting and then can reduce the amount of communication. We can obtain

This work used resources of the National Energy Research Scientific Computing Center (NERSC), supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Sosonkina was supported in part by the Air Force Office of Scientific Research under the AFOSR award FA9550-12-1-0476, by the National Science Foundation grants NSF/OCI—0941434, 0904782, 1047772, and by the U.S. Department of Energy, Office of Advanced Scientific Computing Research, through the Ames Laboratory, operated by Iowa State University under contract No. DE-AC02-07CH11358.

978-83-60810-66-8/$25.00©2015, IEEE

satisfying accuracy with an additional computational cost, which is negligible compared to the cost of factorization. This method has been successfully applied to dense linear systems for either general [6] or symmetric indefinite [8] systems, in the context of direct methods based on matrix factorization.

In this work we want to study the possibility of using RBT in iterative linear system solvers based on Krylov Subspace methods, which are widely used in physical and industrial applications.

This paper is organized as follows. Section II presents the preconditioned Krylov subspace method (PKSM) and the parallel Algebraic Recursive Multilevel Solver (pARMS) for solving sparse linear systems. Section III explains how randomization through Random Butterfly Transformation can be integrated into pARMS. For the obtained solver, Section IV proposes performance and accuracy results. Conclusions are presented in Section V.

## II. Preconditioned Krylov Methods and the pARMS Solver

### A. Preconditioned Krylov Methods

A preconditioned Krylov subspace method (PKSM) is used to solve the linear system $Ax = b$, where $A$ is square non-symmetric matrix, in general. If $M$ is a preconditioning matrix, then the right-preconditioned system is may be expressed as:

$$AM^{-1}y = b, \text{ where } y = Mx \ , \tag{1}$$

which is solved instead of the original system $Ax = b$. To solve this system by using iterative methods, first, we compute the residual $r_0 = b - Ax_0$ [9] after initializing $x_0$, then we may use a right-preconditioned Krylov subspace method to find an approximate solution from the affine subspace [10]:

$$x_m = x_0 + \text{span}\{r_0,\ AM^{-1}r_0, \ldots, (AM^{-1})^{m-1}r_0\} \ , \tag{2}$$

which satisfies certain conditions. For instance, the GMRES algorithm [4] requires that the residual $r_m = b - Ax_m$ has a minimal 2-norm. The flexible GMRES is abbreviated as FGMRES [4]. Its implementation differs from that of GMRES mainly in storing the preconditioned vectors $z_j = M_j^{-1}v_j$ because the relation $AZ_m = V_{m+1}\bar{H}_m$ is used instead of a simpler one $(AM^{-1})V_m = V_{m+1}\bar{H}_m$ from GMRES.

One way to obtain the preconditioning matrix $M$ is to use an incomplete $LU$ (ILU) factorization. ILU is constructed by performing an approximate Gaussian Elimination (GE) [11] on a sparse matrix $A$ and dropping certain nonzero entries of the factorization according to different dropping strategies. A dropping strategy that relies on levels of the matrix fill-in results in a factorization called $ILU(K)$.

The preconditioner $ILU(0)$ is obtained by performing the $LU$ factorization of $A$ and dropping all fill-in elements generated during the process. Conversely, if the nonzeros are dropped according to their numerical value magnitudes, then the resulting factorization is called $ILU$ with the threshold or—if combined with the dropping strategy based on the number of remaining nonzero—with dual threshold ($ILUT$) and is performed as follows. In the algorithm $ILUT(k, \tau)$, there are two important rules. (1) If an element is less than relative tolerance $\tau_i$ ($\tau$ × the norm of the $i$th row), it is dropped. (2) Keep only the $k$ largest elements in the $L$ and $U$ parts of the row along with the diagonal element.

In this work, we use a preconditioner called Algebraic Recursive Multilevel Solver (ARMS) [12], which is based on a block incomplete $LU$ factorization with different dropping strategies. This block factorization consists of an approximate GE process separating the unknowns into two sets; and an idea of independent or "group independent" set is exploited to define the separation. Hence, the original linear system $Ax = b$ is permuted into the form:

$$\begin{pmatrix} B & F \\ E & C \end{pmatrix} \times \begin{pmatrix} u \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} , \tag{3}$$

where the submatrix $B$ corresponds to group-independent set reorderings, thereby generating a block-diagonal matrix $B$ [13]. Thus, it is convenient to eliminate the $u$ variable to obtain a system with only $y$ variable. The coefficient matrix for the resulting "reduced system" is the Schur complement $S = C - EB^{-1}F$ [14]. A recursion can now be exploited, such that dropping is applied to $S$ to limit the fill-ins followed by the reordering of the resulting reduced system into the form (3) by using the group-independent set reordering again. This process is repeated for several levels of recursion until the Schur-complement system is small enough or until a maximum number of recursion levels is reached. Then, the last Schur complement may be solved by a direct or an iterative solver. Note that the sparsification of the Schur complement may be undertaken at each level of recursion, to keep down the preconditioning costs.

In this paper, we are interested in parallelizing the iterative methods rather than direct methods. There are two reasons can explicate our choices. First, the direct methods are scale poorly with problem size, when the problem size augment rapidly, the iterative methods are the only choice, which can compute the approximate solution of linear system $Ax = b$. Second, it is hard to parallelize the direct methods which need more space and time to compute, while iterative methods involve passing from one iteration to the next one by modifying one or a few components of an approximate vector solution at a time and it is easy to parallelize.

### B. Parallel Implementation of ARMS

Figure 1 outlines distributed linear system solution using pARMS [15]. First, the initial matrix $A$ is distributed among the processors, using a graph partitioning method. In Figure 1, each column of blocks depicts one processor, hence there are five processors shown. Second, each processor solves its part of the system in parallel to construct its portion of the global preconditioner. Then FGMRES solves the preconditioned system with a given accuracy.

When considering the parallel implementation, it is important to specify how the matrix is distributed and handled in parallel. In particular, our pARMS implementation partitions

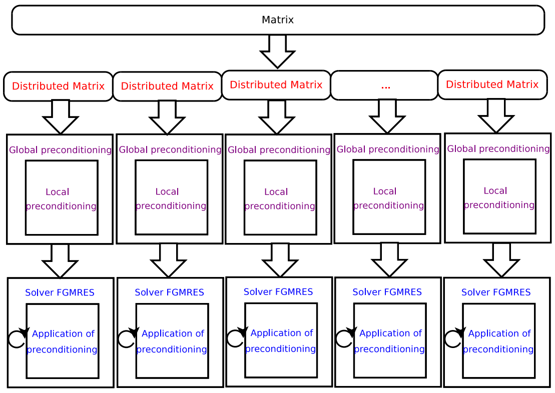


Fig. 1. Sketch of the distributed linear system solution using pARMS on five processors.

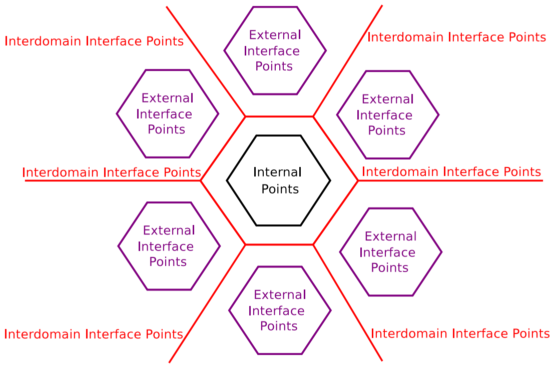

Interdomain Interface Points
External Interface Points
Interdomain Interface Points
External Interface Points
External Interface Points
Interdomain Interface Points
Internal Points
Interdomain Interface Points
External Interface Points
External Interface Points
External Interface Points
Interdomain Interface Points
Interdomain Interface Points

Fig. 2. Per-subdomain view of equation variables-points.

the whole matrix on a single processor using a distributed site expansion (DSE) technique, which is rather simple yet effective in constructing well-balanced subdomains with small interfaces [16]. Although partitioning the entire matrix by a single processor lacks scalability, we note here that this is done by the driver routine, which may be adapted to an application matrix size and format at hand. Given a distributed matrix, Figure 2 shows the per-subdomain division of variables into internal, interdomain interface, and external (residing on the neighboring processors) sets.

We outline now three global preconditioners types available in pARMS: Block-Jacobi preconditioner (BJ), Schur complement preconditioner (SCHUR), and Schur-complement based Restrictive Additive Schwartz preconditioner (SchurRAS). BJ is the simplest global preconditioner because it does not take into account the interface information between neighboring subdomains [17]. SCHUR relates equations associated with the local and interdomain interface points [18]. SchurRAS is constructed from the local ARMS preconditioners in each subdomain using an overlap similar to a standard RAS preconditioner [19] and acting on the Schur complement system as shown in [20]. Specifically, for each of the three preconditioner types, the following algorithms may be implemented in each subdomain.

**BJ preconditioner**:
1. Update local residual: $r_i = (b - Ax)_i$,
2. Solve: $A_i \delta_i = r_i$,
3. Update local solution: $x_i = x_i + \delta_i$.

**SCHUR preconditioner**:
1. From (3) compute: $g_i' = g_i - E_i B_i^{-1} f_i$,
2. Solve: $S_i y_i + \sum_{j \in N_i} E_{ij} y_j = g_i'$, where $S_i = C_i - E_i B_i^{-1} F$ and $N_i$ is a set of neighboring subdomains,
3. Back substitute: $u_i$ with $B_i u_i = f_i - E_i y_i$.

**SchurRAS preconditioner**:
1. Compute local right-hand side $g_i'$.
2. Solve local Schur-complement system extended with rows for all external variables $y_{i,ext}$.
3. Back substitute: $u_i$ with $B_i u_i = f_i - E_i y_i$.

Note that the local solves in step 2 of BJ, SCHUR, and SchurRAS may be accomplished using incomplete $LU$ or ARMS procedures, mentioned in section II-A. In this work, we apply ARMS enhanced with Recursive Butterfly Transformations (RBT) in step 2 of SCHUR to alleviate the extra work associated with pivoting that may be required in the Schur-complement matrix $S_i$ due to its poor conditioning.

## III. Overview of Random Butterfly Transformations and Implementation

In this section we recall the main definitions related to RBT and how it can be applied to pARMS.

### A. Randomization

Random Butterfly Transformation (RBT) is a randomization technique initially described by Parker [7] and recently revisited in [6] for general dense systems and [8] for symmetric indefinite systems. It has also been applied recently to a sparse direct solver in a preliminary paper [21]. The procedure to solve $Ax = b$, where $A$ is a general matrix, using a random transformation and the $LU$ factorization is summarized in Algorithm 1. The random matrices $U$ and $V$ are chosen among a particular class of matrices called *recursive butterfly matrices*. A *butterfly matrix* is a *random* $n \times n$ matrix of the form

$$B^{<n>} = \frac{1}{\sqrt{2}} \begin{bmatrix} R_0 & R_1 \\ R_0 & -R_1 \end{bmatrix},$$

where $R_0$ and $R_1$ are random diagonal $\frac{n}{2} \times \frac{n}{2}$ matrices. A *recursive butterfly matrix* of size $n$ and depth $d$ is defined recursively as

$$W^{<n,d>} = \begin{bmatrix} B_1^{<n/2^{d-1}>} & & \\ & \ddots & \\ & & B_{2^{d-1}}^{<n/2^{d-1}>} \end{bmatrix} \cdot W^{<n,d-1>}$$

with $W^{<n,1>} = B^{<n>}$ where the $B_i^{<n/2^{d-1}>}$ are butterflies of size $n/2^{d-1}$, and $B^{<n>}$ is a butterfly of size $n$.

In the original work by Parker, $d = \log_2 n$; it is proved that, given two recursive butterfly matrices $U$ and $V$, the matrix $U^T AV$ can be factored into $LU$ without pivoting with probability 1 in exact arithmetic, or with probability $1 - O(2^{-t})$ using $t$-bit floating point numbers. RBT was extensively studied for dense matrices and it was shown in [6] that in practice, $d = 1$ or 2 is enough to obtain a satisfactory accuracy (in most cases a few steps of iterative refinement can recover the digits that have been lost). It has been showed that random butterfly matrices are cheap to store and to apply ($O(nd)$ and $O(dn^2)$ respectively) and they proposed implementations using the dense linear algebra PLASMA and MAGMA. As was demonstrated in the related papers, the preprocessing by RBT can be easily parallelized and provides good scalability.

**Algorithm 1** Random Butterfly Transformation Algorithm

```
  Generate recursive butterfly matrices $U$ and $V$
  Perform randomization to update the matrix $A$ and obtain
  the randomized matrix $A_r = U^T AV$
  Factorize the randomized matrix with no pivoting [22]
  Compute $U^T b$ and solve $A_r y = U^T b$, then $x = Vy$
```

### B. Integration of RBT into pARMS

We describe in this section how Random Butterfly Transformations can be integrated into pARMS. Our goal is to find the last level of preconditioning and then replace the original $ILUT$ factorization by the RBT pre-processing. Note that RBT usually concerns dense linear systems, while ARMS addresses sparse linear systems. So we have to convert the last Schur complement which is a sparse matrix into a dense format, and after that we can use RBT. Then after randomizing the last Schur complement $A$ with recursive butterfly matrices $U$ and $V$, the dense matrix is factorized using a Lapack-like [23] routine that performs Gaussian elimination without pivoting, followed by two triangular solves. Note that RBT requires the size of the matrix to be a power of 2, which can be obtained by "augmenting" the matrix $A$ with additional 1's on the diagonal.

The pARMS solver manages the parallel part by using global preconditioning with MPI instructions, while the local part of the code, more precisely the local preconditioning phase does not use MPI instructions. Then the parallelism is entirely managed by pARMS. The local preconditioning can be based on $ilu0$, $iluk$, $ilut$ or $arms$. The essential part resides in the last Schur complement, where we implemented RBT and the preconditioned matrix is then used in FGMRES in order to solve the linear system.

## IV. Numerical Experiments

This section describes preliminary results obtained by integrating RBT into the pARMS solver. The experiments have been carried out using one node (2 twelve-core AMD MagnyCours Opteron 6172 processors running at 2.10GHz) of the Hopper machine located at NERSC[1]. In these experiments, we used matrices from the Davis' collection [24] to test the performance of different preconditioners. The first matrix (`Sherman5`) is a real non-symmetric matrix of size $3,312$ ($nnz$ = 20,793). `Sherman5` arises from a three dimensional simulation model on a $n_x \times n_y \times n_z$ grid using a seven-point finite-difference approximation with $n_c$ equations and unknowns per grid block, where $n_x$ is 16, $n_y$ is 23, $n_z$ is 3, $n_c$ is 3. The second matrix (`Raefsky3`) is a real non-symmetric matrix of size $21,200$ ($nnz$ =1,488,768), which arises from a fluid structure interaction turbulence problem. The third matrix (`Cant`) is a real symmetric matrix that comes from a 2D/3D FEM problem, of size 62,451 ($nnz$ = 2,034,917). For the three matrices, we study the results obtained when using a global Schur complement-based preconditioner with the following local preconditioners: $ilu0$, $iluk$, $ilut$, $arms$, or $arms\_rbt$. The pARMS parameters are chosen for these matrices following the guidance for the local ARMS preconditioner, as explained in [12], for example. Certain parameters influence considerably the size and density of the last Schur Complement, which, in turn, affects greatly the performance of RBT. Since the RBT for dense matrices is used in this work, it is desirable that the last Schur Complement remains dense while being

[1] http://www.nersc.gov

relatively small. Hence, parameter values for the number of ARMS levels and the ARMS independent block size were chosen such that a small Schur Complement is obtained. In particular, the former parameter was small (equals two) while the latter was large (allowing to form the blocks up to the entire local matrix size). At the same time, the drop tolerance for the last Schur Complement was kept quite low (0.001) as well as all the other intermediate-level drop tolerances, so that there is close to none sparsification of the Schur Complement.

The experiments are performed using 4 to 12 cores, and we use one MPI process per core and no multi-threading. In Figure 3, we compare the number of iterations required for convergence. We observe that, for matrices `Sherman5` (fig. 3(a)), `Raefsky3` (fig. 3(b)), $arms\_rbt$ performs better than the other local preconditioners. For `Cant`, $arms\_rbt$ converges in fewer iterations than the other preconditioners do so when using up to eight cores. This observation suggests that $arms\_rbt$ may be a more versatile preconditioner to use for obtaining superior convergence. Note that, for different numbers of subdomains (one per MPI process) in a given matrix, the obtained parallel preconditioning varies leading to the differences in the number of iterations to converge. Note that for these preliminary tests, $arms\_rbt$ requires more time to solve the system since, in our preliminary implementation, a dense-matrix solver was used to solve the last Schur complement system in pARMS. In the future, we plan to develop a sparse RBT solver based on a sparse direct solver, such as SuperLU [25]. Figure 4 represents the residual obtained with the five local preconditioners. We observe that these preconditioners provide us with a similar accuracy, $arms\_rbt$ being more accurate for the matrix `Raefsky3`.

## V. Conclusion and Future Work

We have investigated the feasibility of using RBT randomization in the pARMS solver and how RBT may enhance the iterative convergence. Most of our experiments showed an improvement in the number of iterations and accuracy of results. However, our integration of RBT in pARMS necessitates an implementation that may adjust the sparsity of the last Schur complement matrix based on the available memory and on the performance characteristics of its (direct) solver at hand. As a future work, we will integrate a sparse RBT direct solver based on SuperLU, which will also enable us to solve large-scale sparse linear systems.

## References

[1] A. Buttari, J. Langou, J. Kurzak, J. Dongarra, A class of parallel tiled linear algebra algorithms for multicore architectures, Parallel Comput. 35 (1) (2009) 38–53.

[2] S. Tomov, J. Dongarra, M. Baboulin, Towards dense linear algebra for hybrid GPU accelerated manycore systems, Parallel Computing 36 (5&6) (2010) 232–240.

[3] T. A. Davis, Direct Methods for Sparse Linear Systems (Fundamentals of Algorithms 2), Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2006.

[4] Y. Saad, Iterative Methods for Sparse Linear Systems, 2nd Edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2003.

[5] N. J. Higham, Accuracy and Stability of Numerical Algorithms, 2nd Edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002.

[6] M. Baboulin, J. Dongarra, J. Herrmann, S. Tomov, Accelerating linear system solutions using randomization techniques, ACM Trans. Math. Softw. 39 (2) (2013) 8:1–8:13.

[7] D. S. Parker, Random butterfly transformations with applications in computational linear algebra, Tech. Rep. CSD-950023, University of California Los Angeles, CA USA (1995).

[8] M. Baboulin, D. Becker, G. Bosilca, A. Danalis, J. Dongarra, An efficient distributed randomized algorithm for solving large dense symmetric indefinite linear systems, Parallel Computing 40 (7) (2014) 213–223.

[9] M. Arioli, J. Demmel, I. Duff, Solving sparse linear systems with sparse backward error, SIAM Journal on Matrix Analysis and Applications 10 (2) (1989) 165–190.

[10] B. N. Bond, L. Daniel, Guaranteed stable projection-based model reduction for indefinite and unstable linear systems, in: Proceedings of the 2008 IEEE/ACM International Conference on Computer-Aided Design, ICCAD '08, IEEE Press, Piscataway, NJ, USA, 2008, pp. 728–735.

[11] S. Donfack, J. Dongarra, M. Faverge, M. Gates, J. Kurzak, P. Luszczek, I. Yamazaki, A Survey of Recent Developments in Parallel Implementations of Gaussian Elimination, Concurrency and Computation: Practice and Experience (2014) 18.

[12] Y. Saad, B. Suchomel, ARMS: an algebraic recursive multilevel solver for general sparse linear systems, Numerical Linear Algebra with Applications 9 (5) (2002) 359–378.

[13] Y. Bai, W. N. Gansterer, R. C. Ward, Block tridiagonalization of "effectively" sparse symmetric matrices, ACM Trans. Math. Softw. 30 (3) (2004) 326–352.

[14] Z.-H. Cao, Constraint schur complement preconditioners for nonsymmetric saddle point problems, Appl. Numer. Math. 59 (1) (2009) 151–169.

[15] Z. Li, Y. Saad, M. Sosonkina, pARMS: a parallel version of the algebraic recursive multilevel solver, Numerical Linear Algebra with Applications 10 (5-6) (2003) 485–509.

[16] Y. Saad, M. Sosonkina, Non-standard parallel solution strategies for distributed sparse linear systems, in: P. Z. *et al.* (Ed.), Parallel Computation: 4th International ACPC Conference, Vol. 1557 of Lecture Notes in Computer Science, Springer-Verlag, 1999, pp. 13–27.

[17] B. F. Smith, P. E. Bjørstad, W. D. Gropp, Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations, Cambridge University Press, New York, NY, USA, 1996.

[18] Y. Saad, M. Sosonkina, Distributed Schur Complement techniques for general sparse linear systems, SIAM J. Scientific Computing 21 (1999) 1337–1356.

[19] X.-C. Cai, M. Sarkis, A restricted additive schwarz preconditioner for general sparse linear systems, SIAM J. Sci. Comput. 21 (2) (1999) 792–797.

[20] Z. Li, Y. Saad, Schurras: A restricted version of the overlapping schur complement preconditioner, SIAM J. Sci. Comput. 27 (5) (2005) 1787–1801.

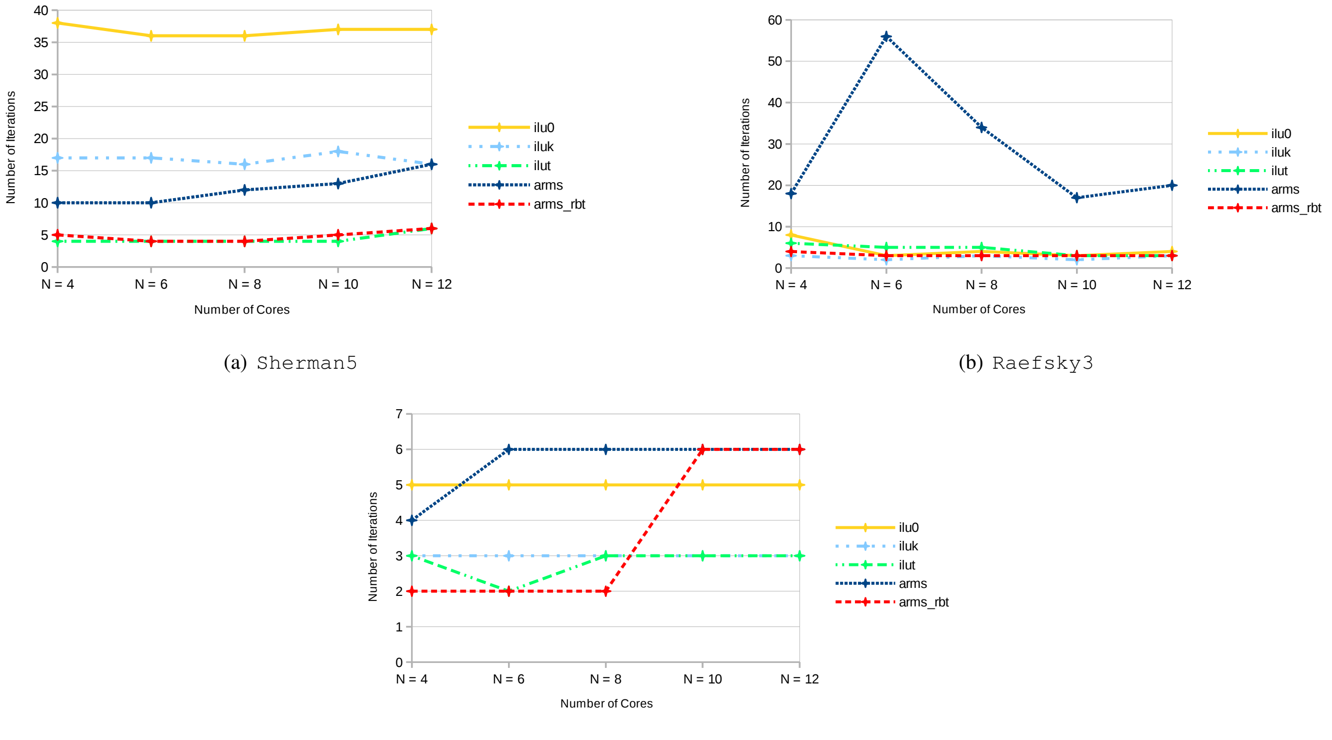
[21] M. Baboulin, X. S. Li, F.-H. Rouet, Using random butterfly transformations to avoid pivoting in sparse direct methods, in: Proceedings of VECPAR 2014, 2014.

[22] M. Baboulin, S. Donfack, J. Dongarra, L. Grigori, A. Rémy, S. Tomov, A class of communication-avoiding algorithms for solving general dense linear systems on cpu/gpu parallel machines, in: International Conference on Computational Science (ICCS 2012), Vol. 9 of Procedia Computer Science, Elsevier, 2012, pp. 17–26.

[23] E. Andersen, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchev, D. Sorensen, LAPACK user's guide, 3rd. Ed. 1999 SIAM, Philadelphia (1999).

[24] T. A. Davis, Y. Hu, The University of Florida Sparse Matrix Collection, ACM Trans. Math. Softw. 38 (1) (2011) 1–25.

[25] X. S. Li, An overview of SuperLU: Algorithms, implementation, and user interface, ACM Transactions on Mathematical Software 31 (3) (2005) 302–325.

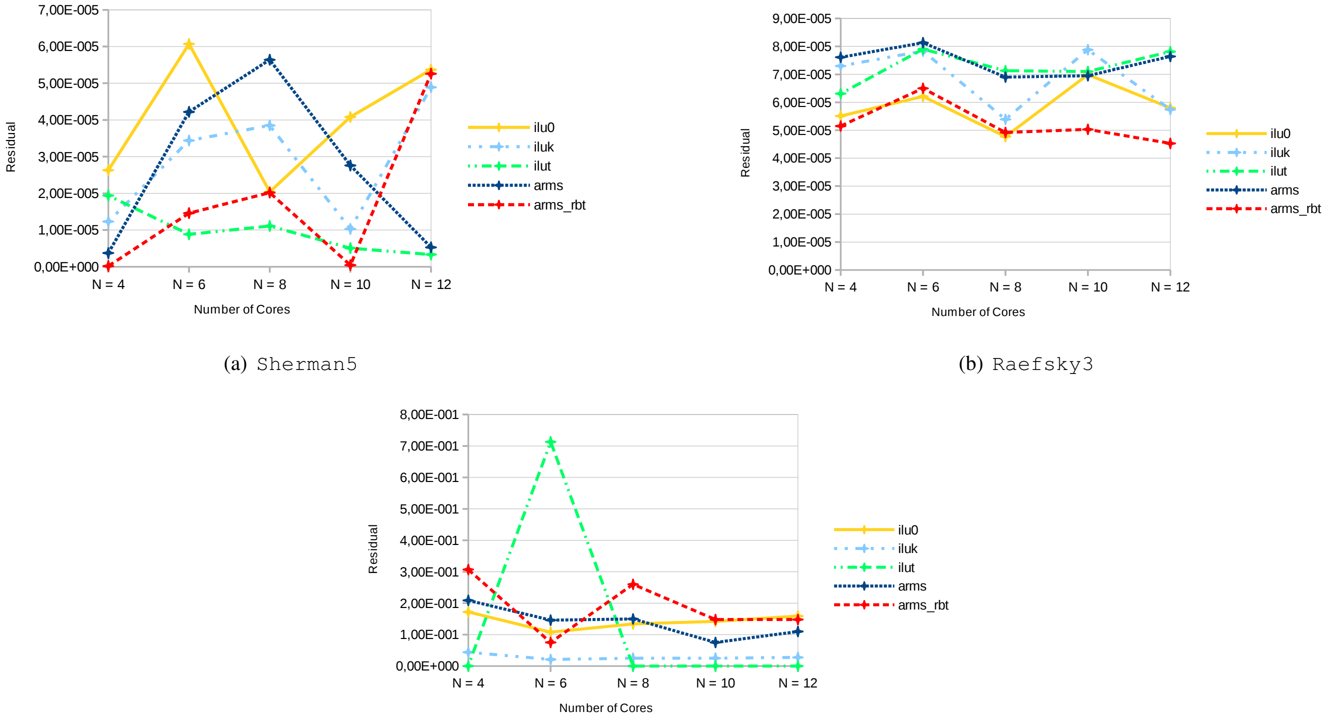


(a) Sherman5

(b) Raefsky3

(c) Cant

Fig. 3. Iterations required for convergence with five choices of local preconditioner.



(a) Sherman5

(b) Raefsky3

(c) Cant

Fig. 4. Residual for test problems with five choices of local preconditioner.

DOI: 10.15439/2015F87
ACSIS, Vol. 5

# Block Preconditioned Conjugate Gradient Method for Extraction of Natural Vibration Frequencies in Structural Analysis

Sergiy Fialko
Tadeusz Kościuszko Cracow
University of Technology
ul. Warszawska 24 St., 31-155 Kraków, Poland
*Email: sergiy.fialko@gmail.com*

Filip Żegleń
Tadeusz Kościuszko
Cracow University of Technology
ul. Warszawska 24 St., 31-155 Kraków, Poland
*Email: filipzeglen@hotmail.com*

***Abstract*—The block preconditioned conjugate gradient method for extraction of eigenfrequencies and eigenmodes is presented for finite element software in structural analysis. The proposed approach is focused on multi-core desktops and laptops and allows us to effectively analyze large design models, when classical methods based on the factoring of stiffness matrix, significantly reduce performance by intensive use of disk memory. The main attention is paid to proper construction of preconditioning, application of shift technique and creation of the block algorithm allowing the improvement of computing stability and multithreaded parallelization.**

## I. Introduction

CURRENTLY, the vast majority of methods for determining the frequency and modes of natural oscillations in finite element software, usually based on factorization of stiffness matrix **K** or matrix $\mathbf{K}_\sigma = \mathbf{K} - \sigma\mathbf{M}$, where **M** is a mass matrix and σ is a shift.

However, if the dimension of the problem becomes large, the factored matrix does not fit in memory and is written to disk. Thus on each iteration in Lanczos method or in subspace iteration method it is necessary to read the factored matrix from a disk twice when performing forward and back substitutions. The performance of the specified methods considerably degrades because the task is carried out at an intensive use of a disk. Especially sharply this problem becomes when desktops and laptops with usually small amount of core memory are used.

It would seem the alternative approach consists in application of conjugate gradient method or steepest descent method for solution of the generalized algebraic eigenvalue problem to which the natural vibration problem is reduced when using the finite element method. The [13] presents survey of some other approaches using preconditioning technique. However, application of such methods to problems of structural mechanics usually results in to the slow convergence of the iterative process. Often there is a locking of convergence of numerical solution owing to a wide scattering of stiffness in design model, existence of a large number of close natural vibration frequencies etc. This is confirmed by the fact that the widely used commercial FEA software offer a wide selection of iterative solvers for static analysis, which leads to solving systems of linear algebraic equations. However for the solution of the generalized algebraic eigenvalue problem as a rule are used the methods based on factorization of a stiffness matrix. The exception is ANSYS in which the preconditioned conjugate gradient method is applied for generation of Lanczos vectors. The survey of existing software packages devoted to solution of algebraic eigenvalue problems in technical and scientific applications is in [10].

In this article, we confine ourselves to consideration of the methods applying to a class of problems, difficult for numerical realization, – to the large problems of structural mechanics arising when the finite element method is used to modeling of tall buildings and constructions. We present the preconditioned conjugate gradient (PCG) method for solution of partial generalized algebraic eigenvalue problem – extraction of natural vibration frequencies and modes

$$\mathbf{K}\mathbf{v}_i - \lambda_i \mathbf{M}\mathbf{v}_i = 0 \tag{1}$$

where **K** is a symmetrical positive definite stiffness matrix, **M** is a mass matrix, $\lambda_i$ and $\mathbf{v}_i$ is an eigenpair, $i \in [1, n]$, $n << N$, $N$ is dimension of problem.

For ensuring sustainable convergence of method, we paid the main attention to designing of an effective preconditioning, use of shift technique and creation of a block algorithm of PCG method. In addition, we used the multithreaded parallelization to accelerate computations.

## II. Block Preconditioned Conjugate Gradient Solver

### A. Sparse incomplete Cholesky conjugate gradient preconditioning

The problems of structural mechanics often demonstrate the slow convergence due to using of different types of finite elements, thin-walled finite elements of floors, roofs and walls, considerable scattering of stiffness etc. [15]. Therefore, it is crucial to construct the effective preconditioning for accelerating of convergence of the PCG method.

We use an incomplete Cholesky factoring "by value" approach, based on sparse matrix technique [6], allowing to keep a small value of drop parameter ψ ($\psi \in [10^{-9}, 10^{-20}]$). The small entries $\mathbf{H}_{ij} < \psi \cdot \mathbf{H}_{ii} \cdot \mathbf{H}_{jj}$ erasing on each step of

978-83-60810-66-8/$25.00©2015, IEEE

triangular solution procedure practically does not depend on number of threads. The duration of triangular solution for considered class of problem of structural mechanics, requiring a preconditioner of high ability to improve of convergence, is essentially larger than duration of **Kv** and other remaining procedures. Therefore, namely this procedure should be accelerated firstly. However, it is a large problem on SMP multicore computers [6].

Table IX depicts results obtained by non-block sequential PCG eigenvalue solver presented in [3], [4].

TABLE IX
PROBLEM 1. NAÏVE SEQUENTIAL APPROACH. COMPUTER A (AMD OPTERON 6276, 64 GB RAM)

| np | Time of K*v | Time of B*z = r → z | Time of iterations, s | Nos of iterations |
|---|---|---|---|---|
| PCG with shift correction over each 100 iterations ($k_{max}$ = 100) | | | | |
| 1 | 326 | 6107 | 7601 | 1144 |

This method uses naïve algorithms of matrix vector product and triangular solution, taking into account only symmetry of sparse matrices. Besides, the loop unrolling is used in triangular solution procedure.

Comparison between Tables VIII and IX show that the application of the matrix vector product procedure in the Intel MKL runs up to 3 times faster. However, the procedure of a triangular solution of Intel MKL is slower than naïve. This leads to the fact that the naïve sequential PCG method solves the problems of this class faster.

The problem with acceleration of triangular solution procedure does inefficient the approach based on internal parallelization of leading procedures of PCG method.

## III. CONCLUSION

The block conjugate gradient method with shifts in sparse incomplete Cholesky factorization preconditioning is proposed for extraction of lower eigenpairs applying to natural vibration problems arising due to application of finite element method to problems of structural mechanics.

On examples of typical problems of structural mechanics, it is shown that on achievement of high computing stability of a method the specific construction of a preconditioning, introduction of shifts to a preconditioning, and iterations in the block have a crucial importance.

The comparison with shifted block Lanczos method based on parallel sparse direct solver PARFES shows that proposed SBPCG method may be very efficient on computers with restricted amount of core memory, when factorized stiffness matrix block-by-block is stored on disk. In such a situation, the proposed SBPCG method in contrast to Lanczos method runs in core memory.

## REFERENCES

[1] V. E. Bulgakov, M. E. Belyi and K. M. Mathisen, "Multilevel aggregation method for solving large-scale generalized eigenvalue problems in structural dynamics," *Int. J. Numer. Methods Eng.,* vol. 40. pp. 453–471, 1997, http://DOI: 10.1002/(SICI)1097-0207(19970215)40:33.0.CO;2-2.

[2] Y. T. Feng and D. R. J. Owen, "Conjugate gradient methods for solving the smallest eigenpair of large symmetric eigenvalue problems," *Int. J. Numer. Methods Eng.,* vol. 39. pp. 2209 – 2229, 1996, http://DOI: 10.1002/(SICI)1097-0207(19960715)39:13<2209::AID-NME951>3.0.CO;2-R.

[3] S. Yu. Fialko, "Natural vibrations of complex bodies," *Int. Applied Mechanics,* vol. 40, no. 1, pp. 83 – 90, 2004, http://DOI:10.1023/B:INAM.0000023814.13805.34.

[4] S. Fialko, "Aggregation Multilevel Iterative Solver for Analysis of Large-Scale Finite Element Problems of Structural Mechanics: Linear Statics and Natural Vibrations", in *PPAM 2001*, R. Wyrzykowski et al. (Eds.), *LNCS* 2328, Springer-Verlag Berlin Heidelberg, 2002, pp. 663–670, http://DOI: 10.1007/1-4020-5370-3_41.

[5] S. Yu. Fialko, E. Z. Kriksunov and V. S. Karpilovskyy, "A block Lanczos method with spectral transformations for natural vibrations and seismic analysis of large structures in SCAD software," in *Proc. CMM-2003 – Computer Methods in Mechanics*, Gliwice, Poland, 2003, pp. 129 —130.

[6] S. Yu. Fialko, "Iterative methods for solving large-scale problems of structural mechanics using multi-core computers," Archieves of Civil and Mechanical Engineering, vol. 14, pp. 190 – 203, 2014, http:// doi:10.1016/j.acme.2013.05.009.

[7] S. Yu. Fialko, "PARFES: A method for solving finite element linear equations on multi-core computers," Advances in Engineering software, vol. 40, no. 12, pp. 1256-1265, 2010, http:// doi:10.1016/j.advengsoft.2010.09.002.

[8] G. Gambolati, G. Pini and F. Sartoretto, "An improved iterative optimization technique for the leftmost eigenpairs of large symmetric matrices," *J. Comp. Phys.,* no 74, pp. 41 – 60, 1988, http://doi: 10.1016/0021-9991(88)90067-8.

[9] C. K. Gan, P. D. Haynes and M. C. Payne, "Preconditioned conjugate gradient method for sparse generalized eigenvalue problem in electronic structure calculations," *Computer Physics Communications*, vol 134, nr. 1, pp. 33 – 40, 2001, http://DOI: 10.1016/S0010-4655(00)00188-0.

[10] V. Hernbadez, J. E. Roman, A. Tomas and V. Vidal, "A survey a software for sparse eigenvalue problems," Universitat Politecnica De Valencia, SLEPs technical report STR-6, 2009.

[11] G. Karypis and V. Kumar, "METIS: Unstructured Graph Partitioning and Sparse Matrix Ordering System,". Technical report, Department of Computer Science, University of Minnesota, Minneapolis, 1995.

[12] A. V. Knyazev and K. Neymayr, "Efficient solution of symmetric eigenvalue problem using multigrid preconditioners in the locally optimal block conjugate gradient method," *Electronic Transactions on Numerical Analysis*, vol. 15, pp. 38 – 55, 2003.

[13] R. B. Morgan, "Preconditioning eigenvalues and some comparison of solvers," *Journal of* computational *and applied mathematics*, vol. 123, pp. 101 – 115, 2000, http://doi: 10.1016/S0377-0427(00)00395-2.

[14] M. Papadrakakis, "Solution of partial eigenproblem by iterative methods," *Int. J. Num. Meth Eng*., vol. 20. pp. 2283—2301, 1984, http://DOI: 10.1002/nme.1620201209.

[15] A. V. Perelmuter, S. Yu. Fialko, "Problems of computational mechanics relate to finite-element analysis of structural constructions," *International Journal for Computational Civil and Structural Engineering*, vol. 1, no 2, 2005, pp. 72 – 86.

[16] Y. Saad, *Numerical methods for large eigenvalue problems, Revised edition, Classics in applied mathematics*. SIAM, 2011, http://dx.doi.org/10.1137/1.9781611970739.

[17] S. Tomov, J. Langou, A. Canning, Lin-Wang Wang, J. Dongarra, "Conjugate-gradient eigenvalue solver in computing electronic properties of nanostructure architecture," Int. J. Computational Science and Engineering, vol. 2, nr. 3-4, pp. 205 – 212, 2006.

[18] Intel Math Kernel Library Reference Manual.
URL: ttps://software.intel.com/sites/products/documentation/doclib/iss/2013/mkl/mklman/index.htm (Last access: 16.04.2015).

DOI: 10.15439/2015F150
ACSIS, Vol. 5

# Suppression of Radio Frequency Interferences Using the Adaptive FIR Filter Based on the Linear Prediction.

Dariusz Głas
University of Łódź
Department of Physics and Applied Informatics,
Faculty of High-Energy Astrophysics,
90-236 Łódź, Pomorska 149, Poland
Email: dariusz.glas@uni.lodz.pl

***Abstract*—The Linear Predictor (LP) is a finite impulse response adaptive filter using linear prediction to suppress radio frequency interferences (RFI) in the Auger Engineering Radio Array (AERA). AERA focus on the electromagnetic part of the Extensive Air Showers. The electromagnetic part of the shower produces radio signals in geomagnetic radiation and charge excess processes. Due to the reflection of the atmosphere AERA radio stations can observe these signals in the frequency band 30 - 80 MHz. This frequency range is contaminated by narrow-band and other human-made RFI. To suppress these contaminations AERA uses two kind of filters: the Median filter and the infinite impulse response - notch filter, however both of them have disadvantages. LP is a new approach in real-time signal filtering. Laboratory and pampas tests show fast adaptation, acceptable power consumption and very good efficiency of the LP filter.**

## I. Introduction

THE radio emission detection of extensive air showers is widely used in many experiments all over the world [1] [2]. The experiment, which focus on this task within the Pierre Auger Observatory (PAO) [3] is the Auger Engineering Radio Array (AERA) [4].

When ultra-high energy cosmic ray (UHECR) reaches the Earth's atmosphere it produces avalanches of secondary particles, called air showers. The study of the air shower allows understanding the mass of primary particle, its energy and incoming direction. These data help to understand emission mechanisms and help to find sources of the UHECR.

The Earth's magnetic field can deflect the trajectory of electrons and positrons generated in the air shower. This process generates a synchrotron radiation, which can be detected by radio stations [5] [6] [7]. Charged particles are concentrated in a thin disk in front of the shower. The coherent emission from this disk is up to 100 MHz. The frequency band from 30 to 80 MHz was chosen due to the reflection of the atmosphere. The noise in this frequency range consists of galactic noise, human made continuous contamiantions and machine generated transients.

Radio detectors, installed in AERA radio stations have high angular resolution and their duty cycle is nearly 100%. Moreover, they are sensitive to the evolution of longitudal air-showers and the total cost of their instalation and work is relatively low. Data taken by radio stations are next digitized by 12bit 200 MHz analog-to-digital converter and processed in real time by FPGA chip. The triggering depends on a signal shape, so it requires good-quality data. The filtering of the signal is crucial.

## II. Mathematical background

The Linear Predictor (LP) is an adaptive finite impulse response (FIR) filter based on the linear prediction. If the periodic signal is hidden in data, we can use several data samples (stages of the filter) to predict how it would look like in the future [8]. Next, we can subtract the predicted periodic signal from the original one, to get the clean data. The graphical presentation of this process is shown on fig. 1.

If the $s(i)$ represents the continuous stream of data samples, the $a_n$ are the filter coefficients and $e(i)$ is cleaned signal, we can describe the LP filter as:

$$e(i) = s(i) - \sum_{n=1}^{p} a_n s(i - D - n), \tag{1}$$

where $p$ is the number of used data samples and simultaneously number of the coefficients of the LP filter, $D$ is the delay between data we used for prediction and data we want to predict. This delay is necessary to minimize the distortion of the transient signals. It would be discussed later. The predicted signal $\hat{s}(i)$ can be described as:

$$\hat{s}(i) = \sum_{n=1}^{p} a_n s(i - D - n). \tag{2}$$

We can find the optimal solution of the coefficients by

This work was supported by the National Science Centre (Poland) under NCN Grant No. UMO-2013/08/M/ST9/00322

Dariusz Głas is the PhD student with the University of Łódź, Department of Physics and Applied Informatics, Faculty of High-Energy Astrophysics, 90-236 Łódź, Poland

978-83-60810-66-8/$25.00©2015, IEEE

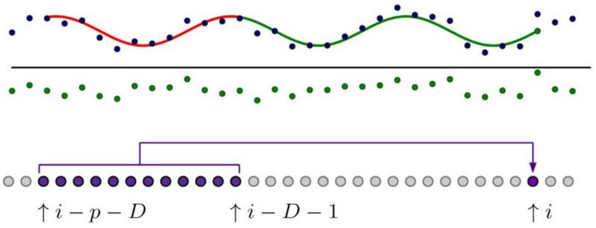


Fig. 1. Graphical illustration of algorithm of the LP filter. Several stages of data are used to predict how the periodic signal hidden in data will look like in the future (above the line). Filtered signal (below the line) is received after subtraction predicted signal form the original one.

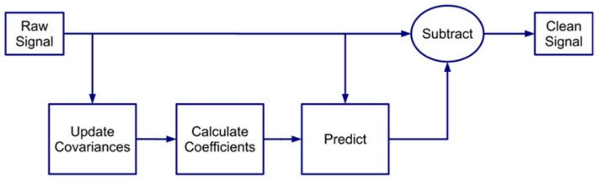


Fig. 2. Schematics showing the LP filter data flow in FPGA. After updating covariance matrix, coefficients can be found by solving a system of linear equations. Prediction step is just summing the products of the coefficients and delayed data. This predicted signal is used for subtraction from original data samples, to get filtered signal.

assuming the Gaussianity and calculating the estimated mean square error:

$$E = \frac{1}{N}\sum_{i=0}^{N-1} e^2(i) = \frac{1}{N}\sum_{i=0}^{N-1}(s(i)-\hat{s}(i))^2, \qquad (3)$$

where $N$ is number of data samples used for calculation of the covariance matrix. In our tests we used $N = 1024$. The next step in finding the best values of coefficients is the minimization of the mean square error, by derivating it with respect to every coefficient and comparing the result to zero:

$$\frac{\partial}{\partial a_i}E = 0. \qquad (4)$$

As a result we get the system of $p$ equations:

$$\sum_{i=0}^{N-1} s(i-D-n)s(i) = \sum_{i=0}^{N-1}\sum_{m=1}^{p} a_m s(i-n)s(i-m). \qquad (5)$$

We can rewrite this equation in vectorial form:

$$\overrightarrow{r^*} = \boldsymbol{R}\vec{a}, \qquad (6)$$

where $\vec{a}$ is vector of the coefficients and $\overrightarrow{r^*}$ is defined as follows:

$$r^*(n) = \sum_{i=0}^{N-1} s(i-D-n)s(i). \qquad (7)$$

**R** is the covariance matrix described as:

$$R(m,n) \equiv \sum_{i=0}^{N-1} s(i-n)s(i-m). \qquad (8)$$

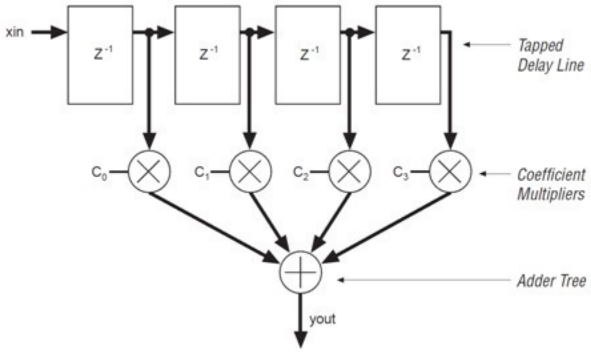


Fig. 3. Schematics of prediction step. Delayed data are multiplied by proper coefficients. Products are summed by several steps of accumulation.

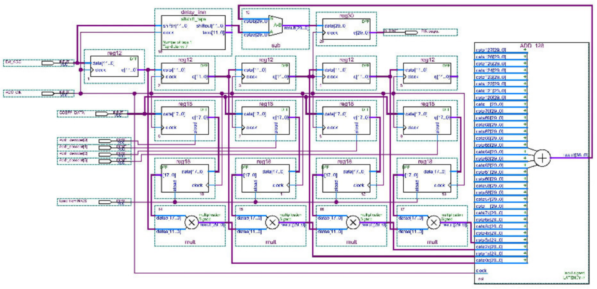


Fig. 4. Fragment of fast logic block connections required for the prediction step.

This matrix is symmetric and diagonal-constant (Toeplitz matrix). The system of linear equations consisting on the Toeplitz matrix can be solved using Levinson procedure instead of Gauss-Jordan elimination. The Levinson procedure algorithm runs in $o(n^2)$ time, which is significant improvement in comparison to Gaussian algorithm, which runs in $o(n^3)$.

## III. FPGA IMPLEMENTATION

Fig. 2 shows the FPGA data flow of the LP filter. The 1024 ADC samples are used to build the covariance matrix. This step can be done either in the virtual processor NIOS® or in fast logic blocks. The creation of the matrix and the $(\overrightarrow{r^*})$ vector in fast logic requires the delay line to get $s(i-D-n)$, the multiplier to get the partial terms and the accumulator. The covariance matrix can be created in fast logic blocks in time of 5 $\mu$s, when in the NIOS® it can be done in several tens of ms. The main reason of this is relatively slow data transfer to the NIOS®. When virtual processor NIOS® receives the data, it sends the signal, that it is ready to receive the next data and the address of this data in DPRAM. This process is very time consuming. Moreover, the NIOS® processor can work with maximal frequency only 100 MHz (80 MHz for tests). The fast logic works with 200MHz clock and can be used in situations, when time is an important factor.

The calculation of the LP coefficients is the most time consuming step in the refreshing of the LP filter. It cannot be done in the fast logic due to a complexity of the task,

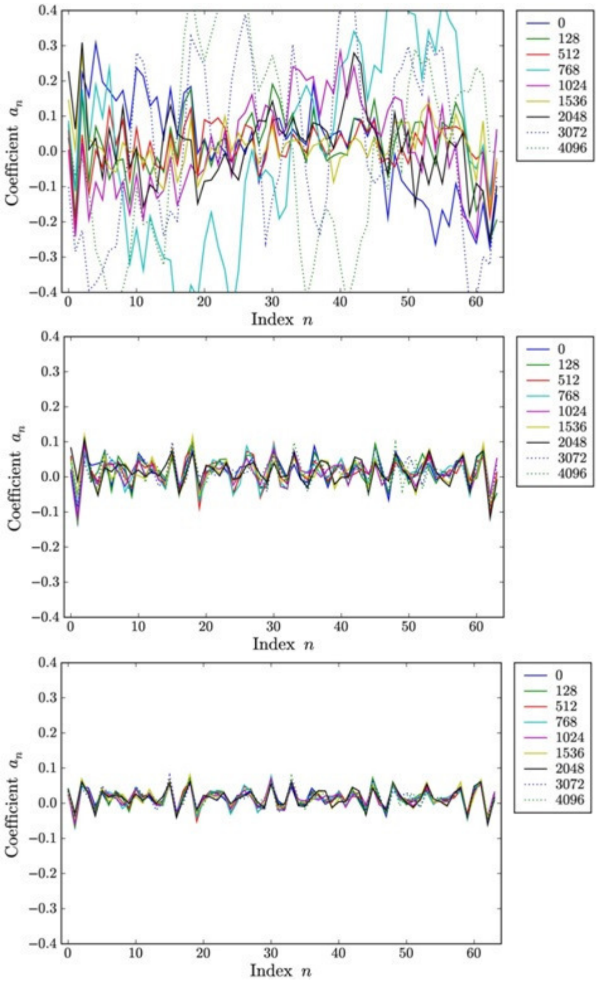


Fig. 5. Stability of the calculated coefficients. Upper plot shows the situation when the determinant of the covariance matrix is close to zero. After introducing the fudge factor $f$ coefficients are stabilized slightly (middle graph $f$=0.1). For $f$=1 (bottom plot) the stabilization is the best.

but it can be done in the virtual processor NIOS®. This step consists in solving the vectorial equation (eq. 6) and highly depends on the number of coefficients and the method used for solving. The Levinson method needs 191 ms for solving the 64x64 covariance matrix, when the Gauss elimination for the same task requires almost 2 seconds. For solving the 128x128 covariance matrix Levinson procedure needs 760 ms and Gauss elimination almost 15 seconds.

Results of both methods are the same. After the calculation, coefficients are transferred from the NIOS® to fast logic temporary registers one by one, which requires at least $p$ clock cycles, where $p$ is the number of the coefficients. When all coefficients are transferred to temporary registers, they are reloaded in one clock cycle to destination registers. This step is required to avoid the situation, when some of coefficients are ”old” and some are ”new”.

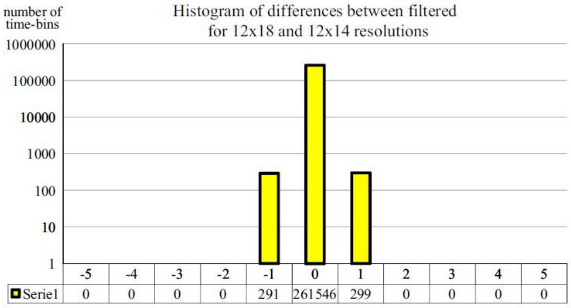


Fig. 6. Histogram of differences between 12x14 and 12x18 resolutions. Differences shows only on the LSB and are on the level of 0.01%.

The prediction step consists in summing multiplications of coefficients and delayed data (eq. 2). This step is done in FPGA fast logic. Data are delayed to get $s(i-D-n)$ and next they are multiplied by an appropriate coefficient. This product is next added in several steps to get the predicted, periodical signal, hidden in the original data. Fig. 3 shows the schema of this process and fig. 4 shows fast logic blocks connections required for calculations. The prediction step can work continuously, because it does not require the gap to update the coefficients. After the prediction we just have to subtract the predicted, periodic signal from the original one to get the clean data.

## IV. Optimization of the LP filter

The stability of the coefficients in time is shown on the fig. 5. The instability of the coefficients seen in the first plot is caused by the specific situation, when determinant of the covariance matrix is close to zero. To avoid this situations we introduce the fudge factor ($f$) and change the covariance matrix:

$$\tilde{R}_{nn} = (1+f)\cdot R_{nn}$$

$$\tilde{R}_{mn} = R_{mn} \text{ for } m \neq n$$

Middle and bottom graphs show the stabilization of the coefficients for $f$=0.1 and $f$=1 respectively. The stabilization is the best for The next specific parameter is the width of coefficients data. The algorithm of the LP filter requires multiplications of coefficients and data from the ADC. The dedicated multiplier requires the different number of DSP blocks, depending on data width. Single DSP block can be used if ADC data width and coefficients width are maximally 9 bits, however data received from the ADC is 12 bits. This implicates the multiplier must use at least 2 DSP blocks. Using 2 DSP blocks allow to multiply 12 bit data with maximally 18 bit coefficient. The 18 bit accuracy of coefficients is not necessary, so it can be reduced to 14 bits without special loses in the efficiency. Differences between those two cases for the filtered signal are only on the LSB and appear to be on the level of 0.01% (fig. 6). Moreover, the reduction from 18 to 14 bits reduces by ~30mW the power consumption. The further reduction of the coefficients width (to 12 bits) increases differences to the level of 1%. The reduction of the power consumption is only

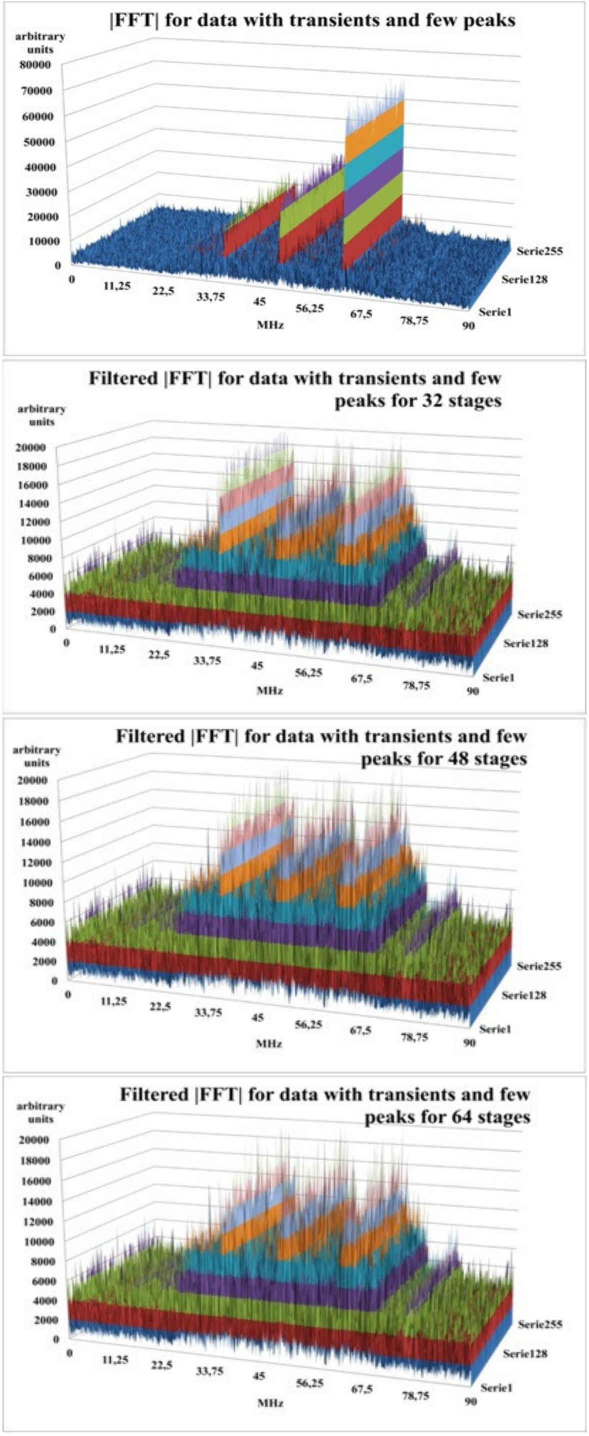


Fig. 7. Original and filtered data with transients and few peaks. LP filter variants with 32, 48 and 64 stages give almost the same result. Slightly contaminated data can be filtered using LP filter with 32 stages

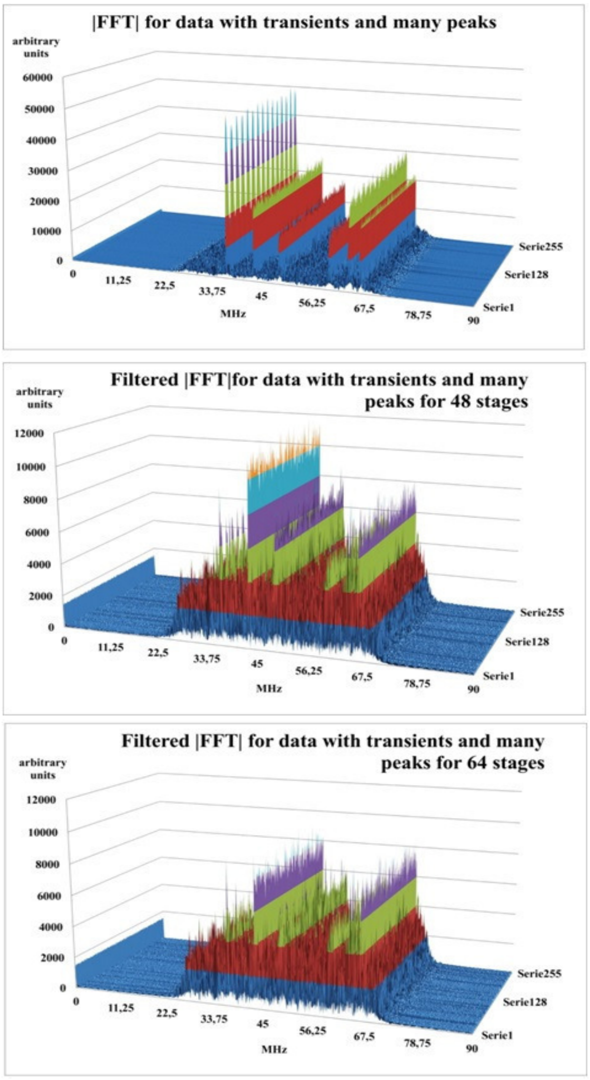


Fig. 8. Original and filtered data with transients and many peaks. In this case the suppression factor is much better in 64-stages variant of the LP filter, however variant with 48 stages also significantly suppresses the contamination. It is recommended to use 64 stages for strongly contaminated data.

~10mW, which is negligible. It was decided that width of the coefficients would be 14 bits.

The length of the filter is an important parameter, which has significant influence on the efficiency of the LP filter. The length of the filter is the number of data used for the prediction and it is the same as the number of coefficients of the filter.

Various lengths of the filter: 64, 48 and 32 have been tested. The filter with length 128 was decided not to be tested, because of the long time of the refreshment (almost 0.8 s). Results of tests are shown on fig. 7 and fig. 8. Fig. 7 shows the data contaminated only by few narrow RFI. The results of the filtering for this case are almost the same for 64, 48 and even for 32 stages of the filter. However, when data are contaminated by many RFI (fig. 8) the filtering using 64 coefficients is more efficient than using 48 stages of the

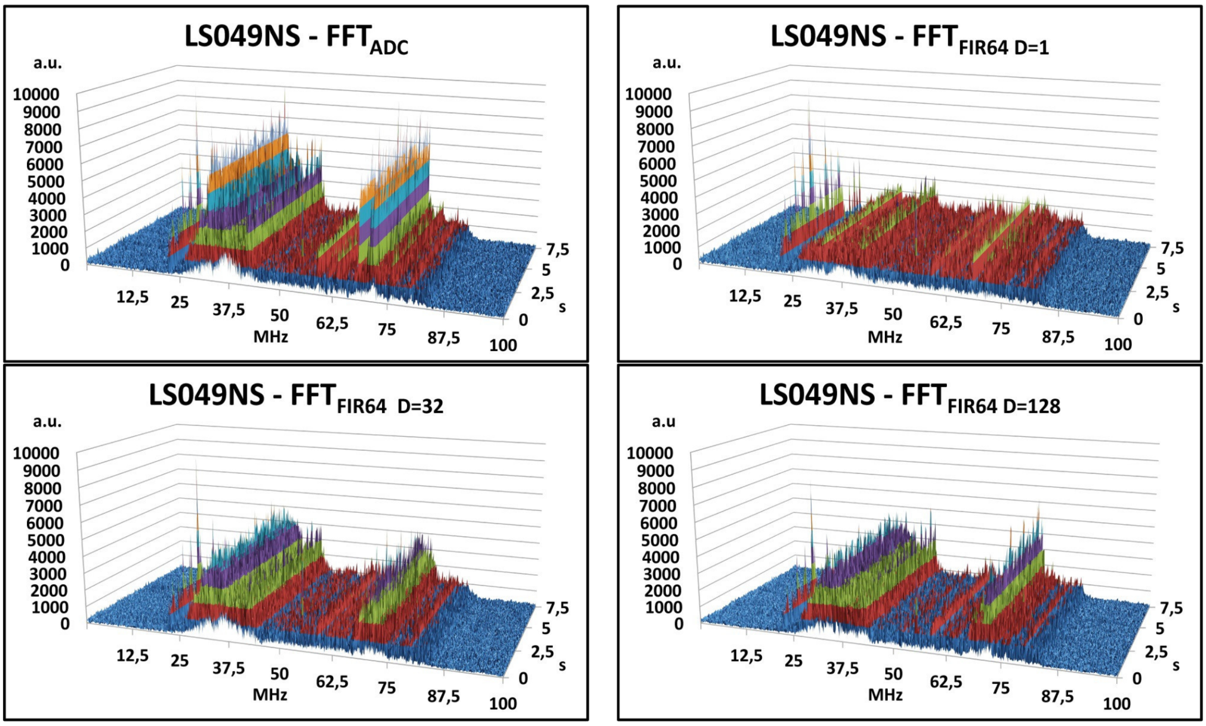


Fig. 9. |FFT| of original and filtered data from station LS049, polarization NS for several variants of delay line $D$. Filtering is the best, when the delay line $D = 1$. In rest of cases filtering is on the same, acceptable level.

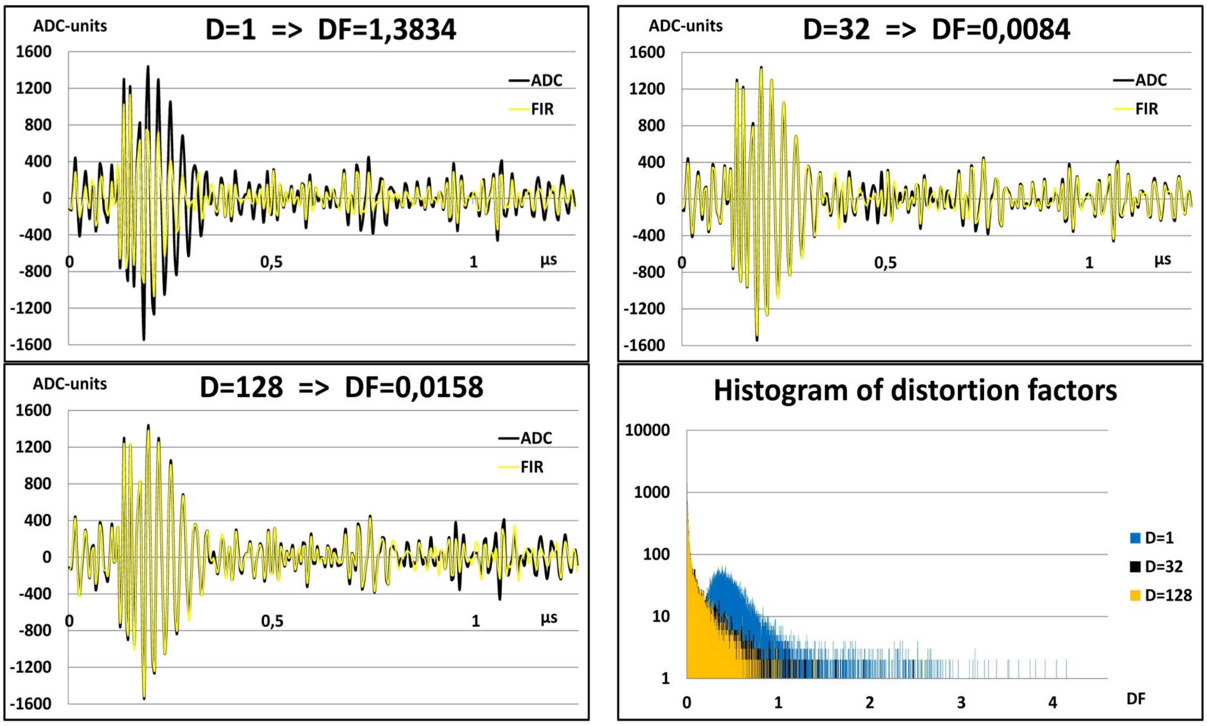


Fig. 10. Original and filtered data in time domain. Although suppression of the variant of the LP filter with delay-line $D = 1$ is the best, it distorts signal too much and has to be rejected. Both variants, with 32 and 128 length of the delay-line move the distortion made by signal appropriately 32 or 128 data samples away from the signal. In these cases signal remains almost unaffected.

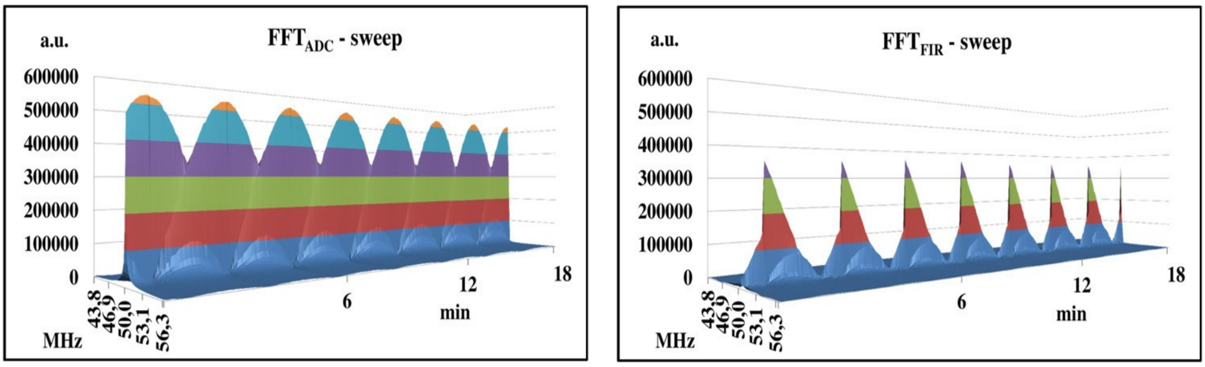


Fig. 11. |FFT| of original and filtered mono-carrier with 2 minutes drift in frequency from 50.0 MHz to 50.2 MHz. When frequency of the contamination is the same as in time of calculation coefficients for the LP filter, suppression level is very high. This result is repeated every 2 minutes.

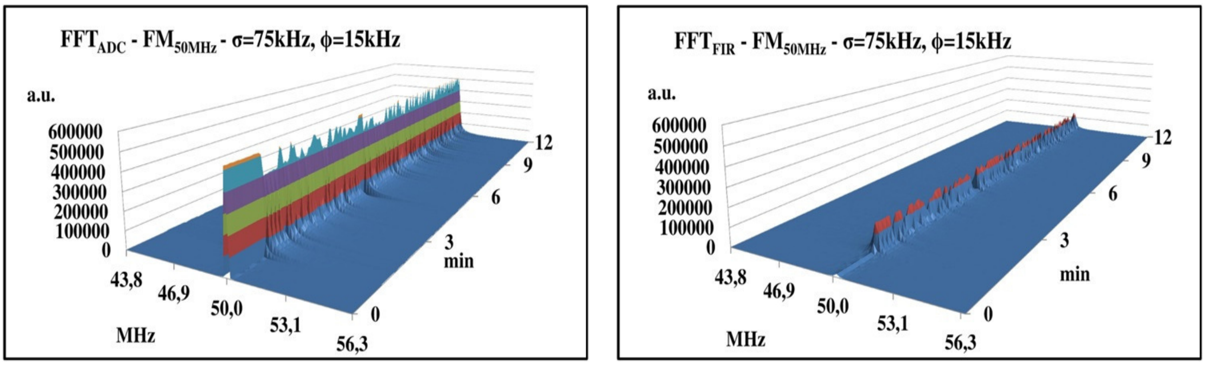


Fig. 12. |FFT| of original and filtered mono-carrier with restrictive FM modulation. The efficiency of the filtering is still very high.

filter. Shorter filters require less FPGA resources, use less power and can be calculated much faster ( ~2 times faster for 48 stages and ~4 times faster for 32 stages of the filter). Simulations show, the LP filter can be used with 32 stages in situations, where the contamination is not sophisticated. For very contaminated data we can use the LP filters with 48 or 64 stages.

The delay line $D$ between the data we used for prediction and data we want to predict was introduced to avoid the distortion of the signal. We can define the distortion factor ($DF$) to calculate how the filter affects data and the registered signal:

$$DF = \sum_{k=-16}^{16} \left( \frac{(x_{FIR})_k \; - \; (x_{ADC})_k}{(x_{ADC})_k} \right)^2 \qquad (9)$$

Fig. 9 shows filtered data for three variants of parameter $D$. For all values of the $D$ factor the LP filter shows good efficiency, but filtering is the best when data we want to predict is located just after data we used for prediction ($D$=1). However, this case introduce significant distortion of the signal (fig. 10 upper left plot) and has to be rejected. Histogram of distortion factors (fig. 10 bottom right plot) shows explicitly, that delay-line $D$=128 distorts signal minimally. $D$=128 moves the distortion made by signal 128 data samples away from the signal region. To keep this safety margin, it was decided to use delay-line $D$=128 for all tests.

## V. Laboratory Tests

All laboratory tests were made using Altera DK-DEV-5CEA7N development kit with Cyclone V FPGA. The generated signals were transferred from generators to Texas Instr. ADS4249EVM Evaluation Module with 2-channel 14-bits 250 MSps ADC. Connection between both modules was provided using the LVDS data transmission by Altera HSMC-ADC-BRIDGE (fig. 22).

The long-term stability of the filter was tested by using a single mono-carrier, which frequency drifted from 50.0 MHz to 50.2 MHz in 2 minutes. Coefficients of the LP filter had been calculated and had been loaded to the registers before this test. These coefficients were not changed during this test. Fig. 11 shows results of the test in time of several minutes. We can observe very good suppression when the frequency goes back to the frequency, which was used for calculation of LP filter coefficients. If data are contaminated by RFI which frequency slowly changes in time, coefficients of the LP filter do not have to be refreshed frequently.

Next test consists on checking the Hi-Fi FM configuration. The signal was contaminated by 50 MHz carrier with 75 kHz deviation, with acoustic modulation 15 kHz. This conditions

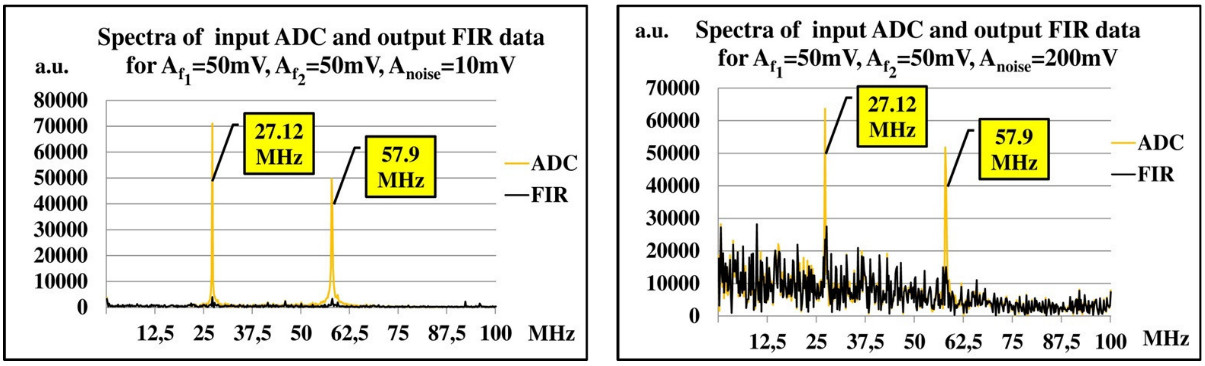


Fig. 13. |FFT| of original (ADC) and filtered (FIR) data of two mono-carriers with 2 variants of noise level. Suppression of carriers is on very good level, even if the amplitude of the noise is greater than the amplitude of carriers.

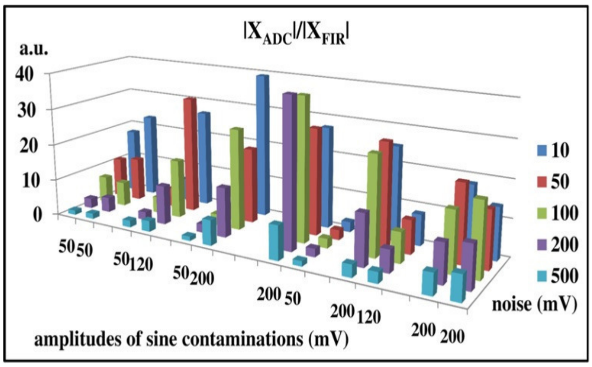


Fig. 14. Suppression levels of two sine contamination in dependancy of noise level and several variants of sine amplitudes. Noise level can greatly change the suppression factor. Contamination with higher amplitude is much more suppressed than the second one. This effect is not dependant on noise level.

are very restrictive and they should not appear in real conditions in detectors on pampas. Frequency band for Hi-Fi FM transmissions is 88 - 108 MHz and it is cut-off by already installed analog filters. The suppression of these restrictive contamination is very good (fig. 12).

The LP filter was also tested to check the suppression efficiency. The generated signal was two carriers (27.12 and 57.9 MHz) with noise. Suppression was almost total, when signals were much stronger than noise, but there was also a significant suppression when signals and noise were on the same level (fig. 13).

Several variants of the carriers and noise amplitudes had been tested. Fig. 14 shows the dependency of the suppression level on amplitudes of contamination. Higher noise level reduces efficiency of filtering. We can also observe, that if the amplitude of one carrier is much higher than the amplitude of the second carrier, the suppression level of the second carrier is much lower than the suppression level of the first carrier.

## VI. Comparison to the currently used filters

### A. Median filter

AERA uses two kinds of filters to improve the signal to noise ratio. The first kind is the Median filter. It is based on the Fast Fourier Transform (FFT) technique. Radio signals are transferred from the time domain to the frequency domain by FFT. In the next step, the Median filter suppresses narrow peaks, which correspond to stationary mono-frequent contaminations. This filtering does not affect the radio pulses from cosmic showers. After this step, the cleaned signal is transformed to the time domain again by using the inversed FFT. The whole process is shown on fig. 15. The Median filter has also several disadvantages. Theoretically, the chain of the FFT and inversed FFT should give same data. However, due to finite representation of data in the FPGA, this chain of transformations introduces approximation errors. There are two possible architectures of the FFT: streaming and variable streaming. The streaming architecture uses less resources and is faster than the variable streaming, but the approximation error is much bigger. In the variable streaming architecture the error is reduced to the last significant bit. Another problem is aliasing. When the pulses are located near the converted data blocks, the reconstructed signal may change. Aliasing elimination can be done using overlapping neighboring data blocks, which require over-clocking of the signal processing. This approach is much more energy and resource consuming. The Median filter also distorts the signal in the region near the peak, which is responsible for the trigger. Fig. 16 shows the distortion factors of the Median filter and the LP filter for comparison. We can see, that the LP filter distorts the signal less than the Median filter. Another issue is the power consumption, which can be on the level of 1W per channel. The power consumption is an important factor for experiment powered by solar panels.

### B. Notch filter

Notch filter is infinite impulse response (IIR) filter, which is currently used by AERA. Its power consumption is much lower than the power consumption of the Median filter, which is important for the experiment powered by solar panels. Notch filter is not adaptive and can reject only four arbitrary chosen frequencies from the signal. After programming in the FPGA, these frequencies cannot be changed dynamically, because of potential instability of the filter. These selected frequencies

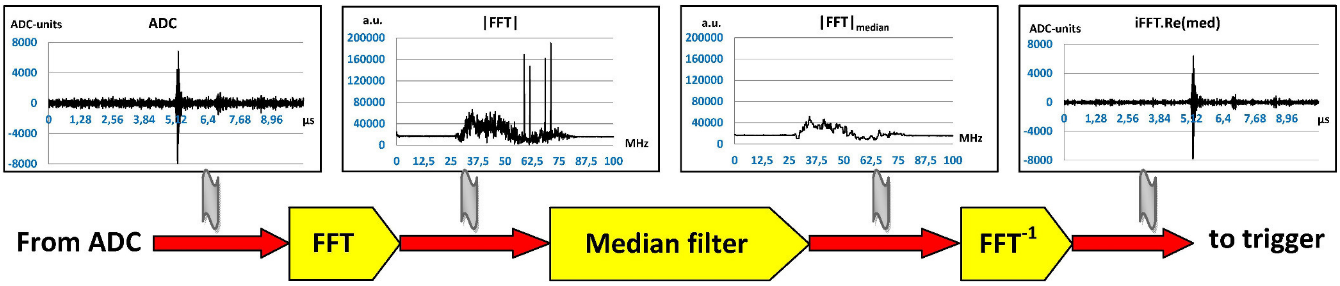


Fig. 15. Diagram showing the data flow in Median filter. Input signals are transfered to frequency domain by FFT and after the filtering by Median filter, cleaned signals are transfered to time domain again using inversed FFT.

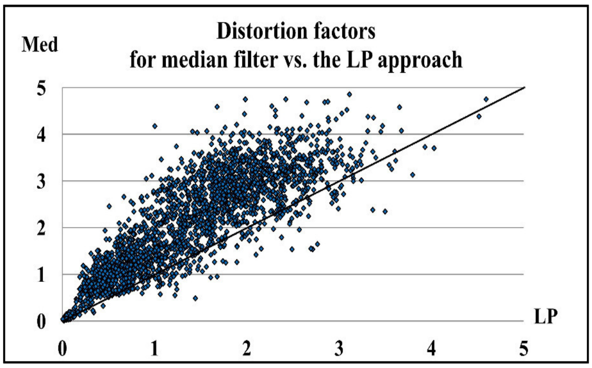


Fig. 16. Comparison of distortion factors for Median and LP filters. In most cases the LP filter distorts the signal less than the Median one.

were found by averaging data from entire array. Fig. 17 shows the comparison of the Notch and LP filters. The efficiency of the Notch filter is very high only near rejected frequencies. Suppression outside the rejected band is much lower. The current consumption of the IIR and the several variants of the LP filter is shown on the fig. 18. The variant with 32 stages has the power consumption compared to the IIR filter. Despites of their limitations, Median and Notch filters were implemented and have successfully worked in the AERA radio stations for several years.

## VII. HPS IMPROVEMENTS

The laboratory tests shows high efficiency of the LP filter, however noise on pampas is much more sophisticated. Besides of narrow-band mono-carriers tests show additional, non-stationary RFI. This type of contamination is weakly affected by the LP filter and strongly decreases filters efficiency. Currently used Notch filter also cannot suppress this kind of contamination, because it can only reject fixed frequency RFI. The main reason of weak suppression by LP filter is probably too long refreshment time of filter coefficients. In laboratory tests, coefficients were calculated by virtual processor NIOS®. This processor can work with maximal frequency 100 MHz, but for tests this frequency was lowered to 80 MHz (to have safety margin). Refreshment time of the filter coefficients in these conditions is 190 ms. The LP filter can suppress RFI, which are longer than the time of its refreshment (fig. 19

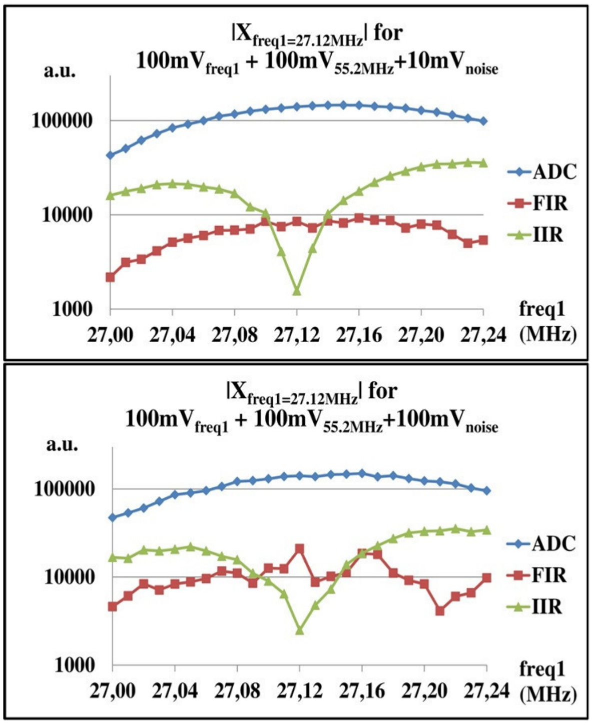


Fig. 17. Comparison of efficiency of LP FIR filter and IIR Notch filter for two variants of noise level. Suppression factor of the Notch filter is more efficient than LP filter only in narrow frequency band near the rejected frequency. Higher noise slightly decreases efficiency of the FIR filter.

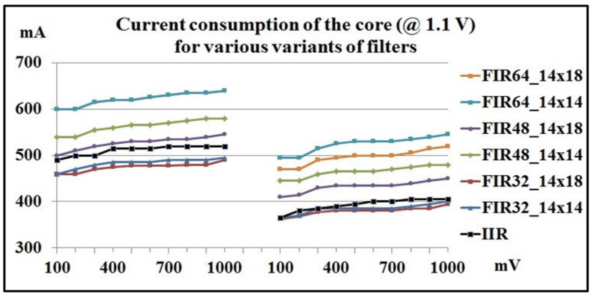


Fig. 18. The current consumption calculated for several variants of the LP filter and for currently used IIR Notch filter. Power consumption for LP filter with 32 stages is on the same level as for IIR filter.

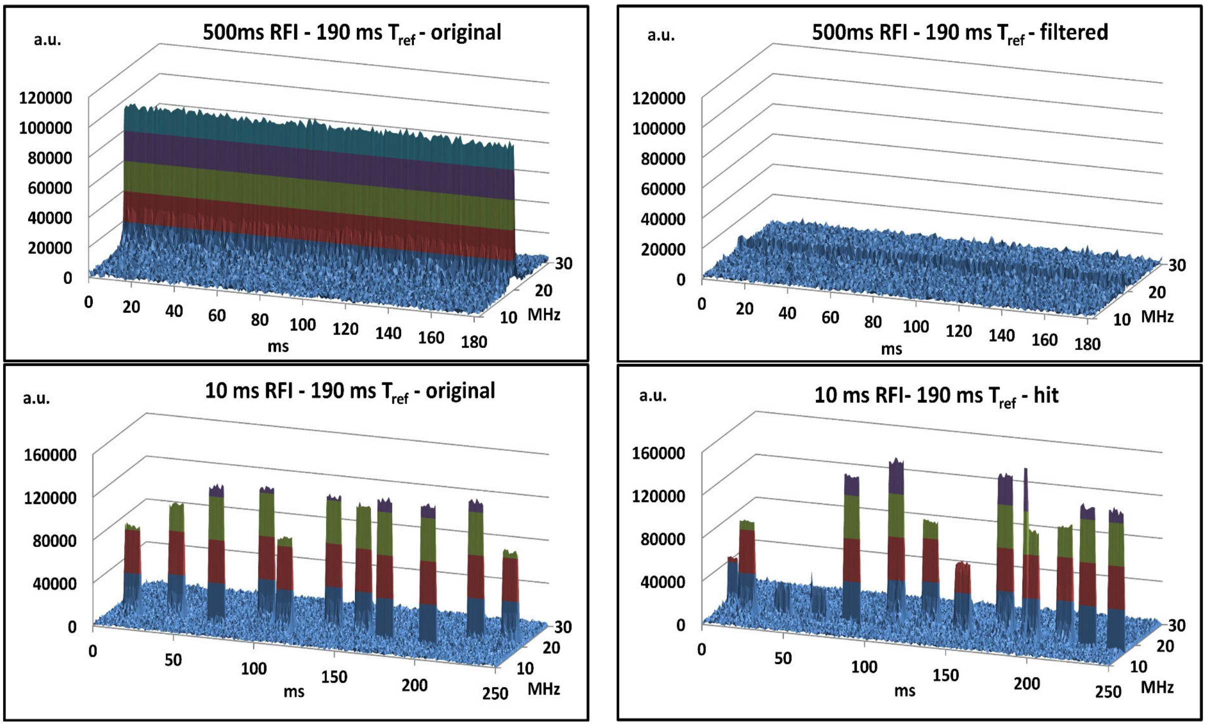
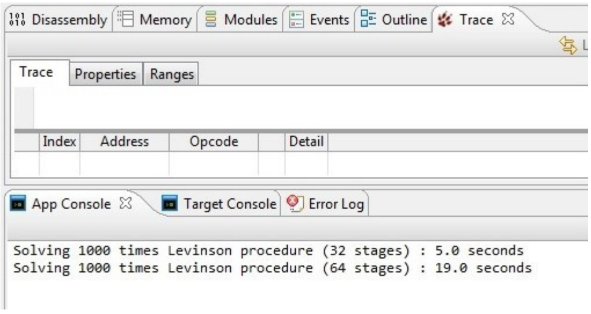


Fig. 19. Suppression achieved by LP filter with coefficients calculated in 190 ms by NIOS® for signals much longer (upper plots) and much shorter (bottom plots) than the refreshment time of the filter. In first case, the suppression is almost total. In second case suppression depends on similarity of data used for calculation and currently filtered data.

upper graphs). If the RFIs length is shorter than refreshment time of LP coefficients, the suppression factor depends on the similarity the signal on which the coefficients were calculated to signal, which is currently filtered (fig. 19 bottom graphs). We can define the suppression factor:

$$SF = \frac{\sum FFT_{original} - \sum FFT_{filtered}}{\sum FFT_{original}}. \tag{10}$$

To reduce the non-stationary RFI, coefficients of the LP filter have to be calculated more frequently. To achieve this we can use Cyclone V SoC HPS, which contains embedded ARM processor. This processor can work with maximal frequency 925 MHz. To have the safety margin, all laboratory tests were made using 800 MHz. Fig. 20 shows the calculation time of the coefficients for two variants of the LP filter. Refreshment time can be reduced from 190 ms to several milliseconds. Fig. 21 shows the simulated results of the filtering by this filter for three different lengths of RFI. The efficiency of the filter is very high for all tests. Suppression factors for case with 100 ms length RFI is SF=72% and for 50 ms RFI SF=59%. If the coefficients were calculated during the gap between the signals, the filter does not suppress the signal. It is seen on all graphs containing filtered data. In 50 ms and 20 ms RFI one of peaks is increased. The probable reason is a short gap between two peaks, so old coefficients were calculated using the data from the previous peak. Every peak is independent, so filter may not work properly in this case. If the refreshment time of LP coefficients is slightly lower than the length of the RFI (fig. 21 bottom graphs), suppression is highly dependent on the moment of starting the calculation of new coefficients. The best scenario is when the calculation starts just after new RFI shows. New set of coefficients starts working near the middle of the peak and can significantly suppress rest of the peak. If the calculation use data just before the new RFI, next calculation would be done in the middle of the RFI occur and starts working at the end or just after the RFI. The suppression factor for this case is much lower than in longer RFI cases, but it is still on acceptable level SF=24%.



```
Solving 1000 times Levinson procedure (32 stages) : 5.0 seconds
Solving 1000 times Levinson procedure (64 stages) : 19.0 seconds
```

Fig. 20. Calculation time for solving by ARM processor system of 32 and 64 linear equations for two variants of LP filter using Levinson algorithm. Calculation time can be shorten from 190 ms to several ms.

The communication between the FPGA and ARM processor is much faster than between the NIOS® and FPGA. The 128 bit bridge between the ARM and FPGA allows sending up

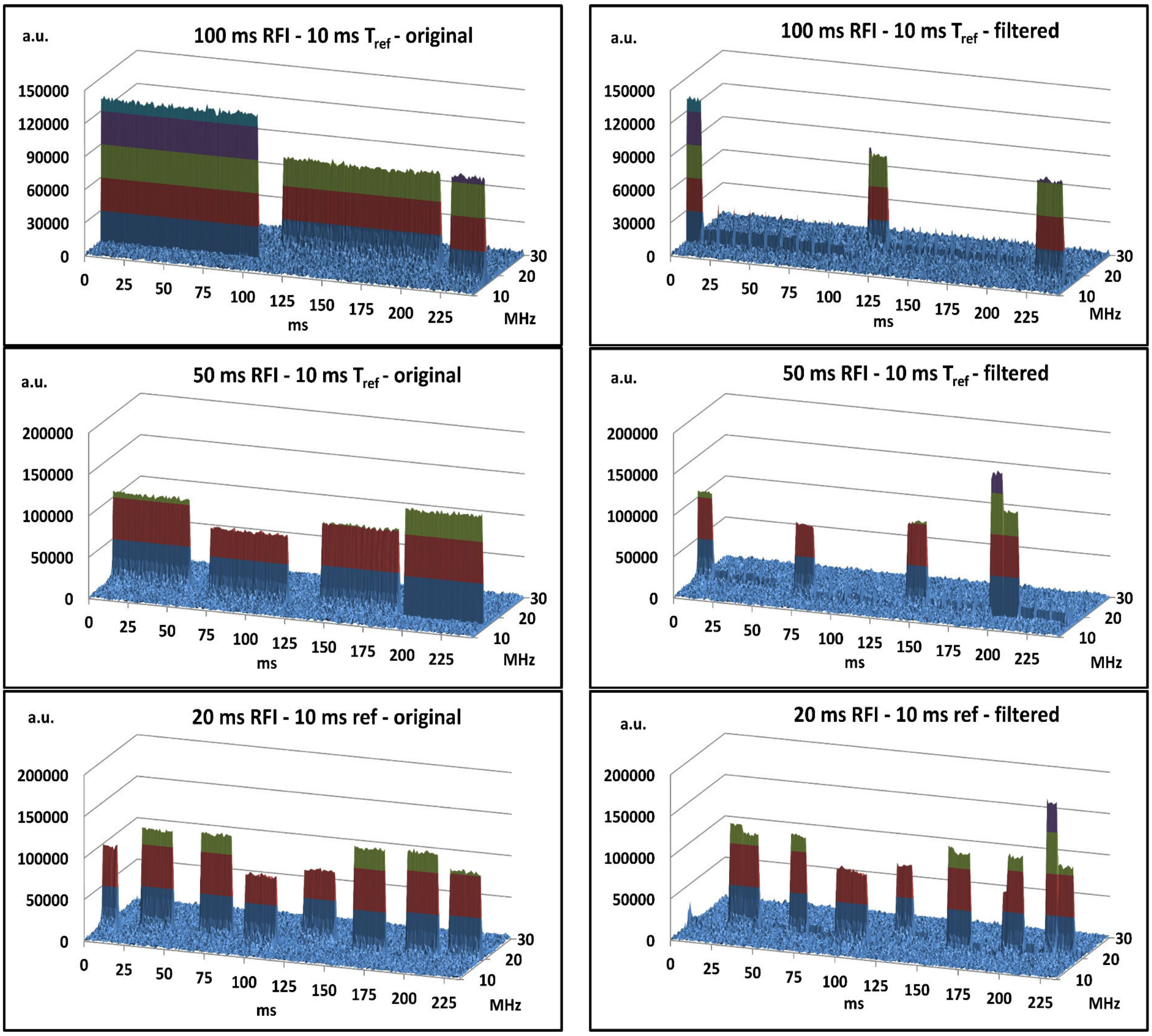


Fig. 21. Suppression achieved by LP filter with coefficients calculated in 10 ms by ARM processor for signals much longer (upper plots) and slightly longer (middle and bottom plots) than the refreshment time of the filter. In upper and middle graphs suppression is very high. The filter does not work only when its coefficients were calculated during the gap between peaks. In bottom graphs suppression strongly depends on the moment of coefficients calculation.

to 10 coefficients in one clock cycle. This means that 64 coefficients can be send in 7 clock cycles and 32 coefficients in 4 clock cycles. Virtual processor NIOS® allows to send only one coefficient per clock cycle.

## VIII. Conclusions

The Linear Predictor is a new kind of filter, which can be used in AERA radio stations. Fast adaptation to new environment conditions, high efficiency in suppression contaminations and reasonable power consumption show it can succesfully replace currently used Notch and Median filters. Additionally, the LP filter leaves almost unaffected signal used for triggering. Using the ARM processor allows to suppress very short non-stationary RFI, which are not suppressed by currently used filters. Moreover, the LP filter is very flexible. The shorter versions of the filter can be implemented in radio stations, where RFI is relativelly weak. Longer versions can be used in radio stations, where filtering is crucial.

## Acknowledgment

This work was supported by the National Science Centre (Poland) under NCN Grant No. 2013/08/M/ST9/00322

## References

[1] H. Falcke,W. D. Apel, A. F. Badea, et al., "Detection and imaging of atmospheric radio flashes from cosmic ray air showers", *Nature*, vol. **435**, pp. 313-316, May 2005.

[2] D. Ardouin, A. Bell'etoile, D. Charrier, et al., "Radioelectric field features of extensive air showers observed with CODALEMA" *Astropart. Phys.*, vol. **26**, pp. 341-350, Dec. 2006.

[3] J. Abraham et al., [Pierre Auger Collaboration], “Properties and Performance of the Prototype Instrument for the Pierre Auger Observatory”, *Nucl. Instr. Meth.*, ser. A, vol. 523, pp. 50-95, May 2004.

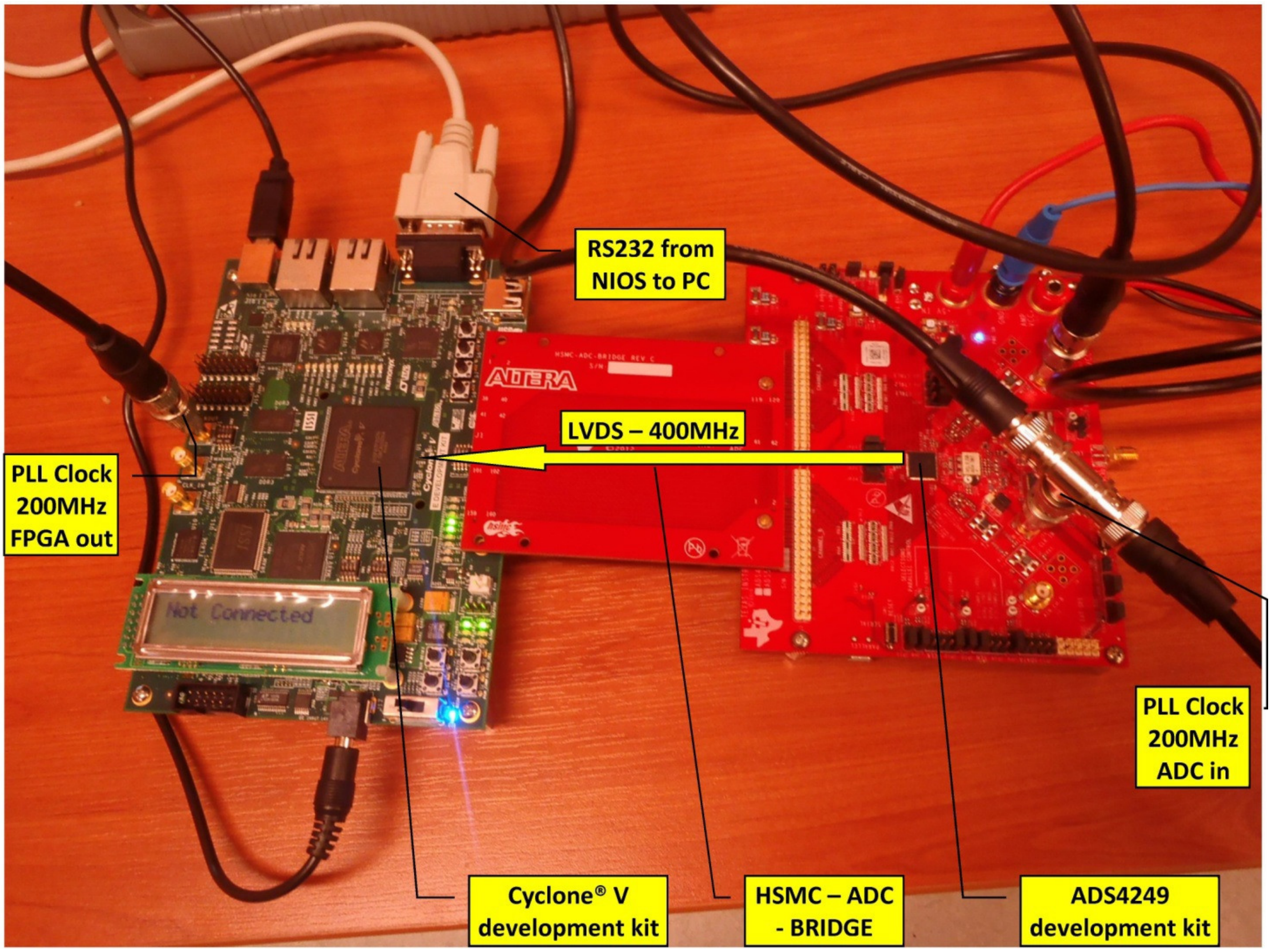


Fig. 22. Connections between developement kit with Cyclone V used for laboratory tests and Texas Instr. ADS4249EVM Evaluation Module. Data were transfered using LVDS technicque by Altera HSMC-ADC-BRIDGE.

[4] S. Fliescher for the Pierre Auger Collaboration, "Radio detection of cosmic ray induced air showers at the Pierre Auger Observatory", *Nucl. Instr. Meth.*, ser. A, vol. **662**, pp. 124-129, Jan. 2012.

[5] D. J. Fegan, "Detection of elusive radio and optical emission from cosmic-ray showers in the 1960s", *Nucl. Instr. Meth.*, ser. A, vol. **662**, pp. 2-11, Jan. 2012.

[6] G. A. Askaryan, *Journal of Exp. and Theoretical Phys.*, vol. **14**, pp. 441, 1962.

[7] G. A. Askaryan, "Coherent Radio Emission from Cosmic Showers in Air and in Dense Media", *Journal of Exp. and Theoretical Phys.*, vol. **21**, pp. 658-659, Jan. 1965.

[8] J. Makhoul, "Linear prediction: A tutorial review" *Proc. of the IEEE*, vol. **63**, no. 4, pp. 561-580, Apr. 1975.

[9] Z. Szadkowski, E.D. Fraenkel, A. M. van den Berg, "FPGA/NIOS Implementation of an Adaptive FIR Filter Using Linear Prediction to Reduce Narrow-Band RFI for Radio Detection of Cosmic Rays", *IEEE Trans. on Nucl. Science*, vol. **60**, pp. 3483-3490, Oct. 2013.

[10] Z. Szadkowski, E.D. Fraenkel, D. Głas, R. Legumina, "An optimization of the FPGA/NIOS adaptive FIR filter using linear prediction to reduce narrow band RFI for the next generation ground-based ultra-high energy cosmic-ray experiment", *Nucl. Instr. Meth.*, ser. A, vol. **732**, pp. 535-539, June 2013.

[11] Z. Szadkowski, Ad M. van den Berg, E.D. Fraenkel, D. Głas, J. Kelley, C. Timmermans, T. Wijnen, "Analysis of the efficiency of the filters suppressing the RFI being developed for the extension of AERA", *33nd International Cosmic Ray Conference - July 2013 - Rio de Janeiro, Brazil*.

[12] Z. Szadkowski, D. Głas, C. Timmermans, T. Wijnen for the Pierre Auger Collaboration, "First results from the FPGA/NIOS Adaptive FIR Filter Using Linear Prediction Implemented in the AERA Radio Stations to Reduce Narrow Band RFI for Radio Detection of Cosmic Rays", *IEEE Real Time Conference - May 2014 - Nara, Japan*.

[13] Z. Szadkowski, D. Głas for the Pierre Auger Collaboration, "Adaptive Linear Predictor FIR Filter based on the Cyclone V FPGA with HPS to Reduce Narrow Band RFI in AERA Radio Detection of Cosmic Rays", *Advancements in Nuclear Instrumentations Measuerment Methods and their Applications - April 2015 - Lisbon, Portugal*.

DOI: 10.15439/2015F29
ACSIS, Vol. 5

# Estimation of numerical reproducibility on CPU and GPU

Fabienne Jézéquel[1,2,3], Jean-Luc Lamotte[1,2], and Issam Saïd[1,2]

[1]Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France
[2]CNRS, UMR 7606, LIP6, F-75005, Paris, France
[3]Université Panthéon-Assas, 12 place du Panthéon, F-75231 Paris CEDEX 05, France
*{Fabienne.Jezequel,Jean-Luc.Lamotte,Issam.Said}@lip6.fr*

***Abstract*—Differences in simulation results may be observed from one architecture to another or even inside the same architecture. Such reproducibility failures are often due to different rounding errors generated by different orders in the sequence of arithmetic operations. Reproducibility problems are particularly noticeable on new computing architectures such as multicore processors or GPUs (Graphics Processing Units). DSA (Discrete Stochastic Arithmetic) enables one to estimate rounding error propagation in simulation programs. In this paper, it is shown that DSA can be used to estimate which digits in simulation results may be different from one environment to another because of rounding errors. A particular implementation of DSA, which enables numerical validation in hybrid CPU-GPU environments, is described. The estimation of numerical reproducibility using DSA is illustrated by a wave propagation code which can be affected by reproducibility problems when executed on different architectures.**

## I. Introduction

RESULTS of numerical simulations may be different from one architecture to another, or even inside the same architecture if they are computed using different compilers for instance. In sequential or parallel environments, different orders in the sequence of floating-point operations may lead to differences in rounding error propagation and therefore to reproducibility failures. It must be pointed out that the cause of differences in results may be difficult to identify: rounding errors or bug? Such differences are particularly noticeable with new computing architectures such as multicore processors, GPUs (Graphics Processing Units) and APUs (Accelerated Processing Units). In high performance numerical simulations, reproducibility problems have been identified in various domains: energy science [1], climate science [2], atomic or molecular dynamics [3], [4], fluid dynamics [5]. Various studies have been carried out on numerical reproducibility on different architectures. On the one hand, strategies have been proposed [2], [3], [4], [5] to improve numerical accuracy, using for instance accurate summations. Other works aim at forcing the reproducibility of results, either affected by the same rounding errors [6], [7] or correctly rounded [8], [9], [10], [11].

DSA (Discrete Stochastic Arithmetic) [12], [13] enables one to estimate rounding error propagation in simulation programs. This paper shows that DSA can be used, not to force a code to be reproducible, but to estimate the number of digits in the results which may be different from one execution to another because of rounding errors. The CADNA[1] library [14], [15], [16], which implements DSA, enables the numerical quality estimation of sequential programs in C or Fortran and of parallel programs using MPI for communication [17]. This paper describes a version of CADNA, briefly introduced in [18], that can be used in hybrid CPU-GPU environments to estimate rounding errors in CUDA programs. This paper is organized as follows. In Section 2, differences in results provided by a wave propagation code executed on several architectures - CPU, GPU and APU - are pointed out. Section 3 presents the principles of DSA. Section 3 also describes the CADNA library and presents the particularities of a CADNA version for CPU-GPU codes. Section 4 shows that the reproducibility problems observed in wave propagation results can be explained by rounding error propagation thanks to the CADNA library. Finally, concluding remarks are presented in Section 5.

## II. Reproducibility failures in a wave propagation code

We consider the three-dimensional acoustic wave equation

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} - \sum_{b\in\{x,y,z\}} \frac{\partial^2}{\partial b^2} u = 0,$$

where $u$ is the particle velocity, $c$ is the wave velocity, and $t$ is the time. This equation, used for instance in oil exploration [19], is solved with an explicit finite difference scheme of order 2 in time and $p$ in space (in our case $p = 8$). We denote by $u^n_{i,j,k}$ (respectively $f^n_{i,j,k}$) the wave (respectively source) field in $(i, j, k)$ coordinates and $n$th time step, $a_l$ $(l = -p/2, \ldots, p/2)$ the finite difference coefficients, $\Delta t$ the time step and $\Delta h$ the spatial step size. Two mathematically equivalent implementations of the finite difference scheme are proposed:

$$u^{n+1}_{i,j,k} = 2u^n_{i,j,k} - u^{n-1}_{i,j,k} + \frac{c^2\Delta t^2}{\Delta h^2} S^n_{i,j,k} + c^2\Delta t^2 f^n_{i,j,k}$$

[1] URL address: http://www.lip6.fr/cadna

978-83-60810-66-8/$25.00©2015, IEEE

where

$$S^n_{i,j,k} = \sum_{l=-p/2}^{p/2} a_l \left(u^n_{i+l,j,k} + u^n_{i,j+l,k} + u^n_{i,j,k+l}\right) \quad (1)$$

or

$$S^n_{i,j,k} = \sum_{l=-p/2}^{p/2} a_l u^n_{i+l,j,k} + \sum_{l=-p/2}^{p/2} a_l u^n_{i,j+l,k} + \sum_{l=-p/2}^{p/2} a_l u^n_{i,j,k+l}. \quad (2)$$

In order to satisfy the CFL (Courant-Friedrichs-Lewy) necessary stability condition [20], the time step is computed by taking into account the wave velocity $c$, the spatial step size $\Delta h$ and the spatial order $p$. Because these two implementations require the same number of arithmetic operations, they should lead to similar performance. However it would be interesting to determine whether they differ in the numerical quality of their results.

The code is executed for $64 \times 64 \times 64$ space steps and 1000 time iterations in IEEE-754 binary32 arithmetic with rounding to the nearest [21] and the following environments:

- AMD Opteron 6168 CPU with gcc 4.7.2 compiler;
- NVIDIA C2050 GPU (Graphics Processing Unit) with CUDA (Compute Unified Device Architecture) platform;
- NVIDIA K20c GPU with OpenCL (Open Computing Language);
- AMD Radeon HD 7970 GPU with OpenCL;
- AMD Trinity APU (Accelerated Processing Unit) with OpenCL.

Different kinds of reproducibility problems are observed. The results numerically vary

1) from one execution to another inside a GPU or an APU; these repeatability problems are due to differences in the execution order of the threads;
2) from one implementation of the finite difference scheme to another; the maximal relative difference between results is of the order of $10^{-1}$ to 1 depending on the architecture, and the mean value of the relative difference between results is of the order of $10^{-5}$ whatever the architecture;
3) from one architecture to another; again, the maximal relative difference between results is of the order of $10^{-1}$ to 1 and its mean value is $10^{-5}$.

Indeed if two sets of results computed in binary32 are compared, the results at the same space coordinates can have from 0 to 7 significant digits in common, and the average number of common significant digits is about 4. We recall that results computed using binary32 arithmetic precision can have at most 7 correct significant digits. To illustrate these reproducibility problems, Table I presents at three space coordinates $(i, j, k)$, $0 \leq i, j, k \leq 63$, the results obtained after 1000 times iterations using the processing units and languages previously mentioned. These results have different orders of magnitude. Both implementations of the finite difference scheme are considered. Considering the example points presented in Table I, any two results computed at the same point in the space domain have 3 to 6 common significant digits.

TABLE I
RESULTS COMPUTED AT THREE DIFFERENT POINTS IN THE SPACE DOMAIN

| | Point in the space domain | | |
|---|---|---|---|
| | $p_1$: (0, 19, 62) | $p_2$: (50, 12, 2) | $p_3$: (20, 1, 46) |
| AMD Opteron CPU with gcc | | | |
| scheme 1 | -1.110479 | 54.54238 | 614.1038 |
| scheme 2 | -1.110426 | 54.54199 | 614.1035 |
| NVIDIA C2050 GPU with CUDA | | | |
| scheme 1 | -1.110204 | 54.54224 | 614.1046 |
| scheme 2 | -1.109869 | 54.54244 | 614.1047 |
| NVIDIA K20c GPU with OpenCL | | | |
| scheme 1 | -1.109953 | 54.54218 | 614.1044 |
| scheme 2 | -1.111517 | 54.54185 | 614.1024 |
| AMD Radeon GPU with OpenCL | | | |
| scheme 1 | -1.109940 | 54.54317 | 614.1038 |
| scheme 2 | -1.110111 | 54.54170 | 614.1044 |
| AMD Trinity APU with OpenCL | | | |
| scheme 1 | -1.110023 | 54.54169 | 614.1062 |
| scheme 2 | -1.110113 | 54.54261 | 614.1049 |

## III. ESTIMATING ROUNDING ERRORS WITH DISCRETE STOCHASTIC ARITHMETIC (DSA)

### *A. Principles of DSA*

Based on a probabilistic approach, the CESTAC method [12] allows the estimation of rounding error propagation which occurs with floating-point arithmetic. When no overflow occurs, the exact result of any non exact floating-point arithmetic operation is bounded by two consecutive floating-point values $R^-$ and $R^+$. The basic idea of the method is to perform each arithmetic operation $N$ times, randomly rounding each time, with a probability of 0.5, to $R^-$ or $R^+$. The computer's deterministic arithmetic, therefore, is replaced by a stochastic arithmetic where each arithmetic operation is performed $N$ times before the next one is executed, thereby propagating the rounding error differently each time. The CESTAC method furnishes us with $N$ samples $R_1, \ldots, R_N$ of the computed result $R$. The value of the computed result, $\overline{R}$, is the mean value of $\{R_i\}_{1 \leqslant i \leqslant N}$ and the number of exact significant digits in $\overline{R}$, $C_{\overline{R}}$, is estimated as

$$C_{\overline{R}} = \log_{10}\left(\frac{\sqrt{N}\left|\overline{R}\right|}{\sigma \tau_\beta}\right)$$

$$\text{with } \overline{\mathrm{R}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{R_i} \text{ and } \sigma^2 = \frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\mathrm{R_i} - \overline{\mathrm{R}}\right)^2.$$

$\tau_\beta$ is the value of Student's distribution for $N - 1$ degrees of freedom and a probability level $1 - \beta$. In practice $N = 3$ and $\beta = 0.05$. Indeed, it has been shown [22], [23] that $N = 3$ is in some reasonable sense the optimal value. The estimation with $N = 3$ is more reliable than with $N = 2$ and increasing the size of the sample does not improve the quality of the estimation. The probability of overestimating the number of exact significant digits of at least 1 is 0.054% and the probability of underestimating the number of exact significant digits of at least 1 is 29%. By choosing $\beta = 0.05$, we prefer to guarantee a minimal number of exact significant digits with a high probability (99.946%), even if we are often pessimistic by 1 digit. The complete theory can be found in [12], [23].

The validity of $C_{\overline{R}}$ is compromised if the two operands in a multiplication or the divisor in a division are not significant [23]. It is essential, therefore, that these results with no significance are detected and reported, since their subsequent use may invalidate the method. The need for this dynamic control of multiplications and divisions has led to the concept of the computational zero. A computed result is a computational zero, denoted by @.0, if $\forall i, R_i = 0$ or $C_{\overline{R}} \leq 0$. This means that a computational zero is either a mathematical zero or a number without any significance, *i.e.* numerical noise.

To establish consistency between the arithmetic operators and the relational operators, discrete stochastic relations are defined as follows. Let $X = (X_1, ..., X_N)$ and $Y = (Y_1, ..., Y_N)$ be two results computed using the CESTAC method, we have from [24]

$$X = Y \text{ if and only if } X - Y = @.0;$$
$$X > Y \text{ if and only if } \overline{X} > \overline{Y} \text{ and X} - \text{Y} \neq @.0;$$
$$X \geq Y \text{ if and only if } \overline{X} \geq \overline{Y} \text{ or X} - \text{Y} = @.0.$$

Discrete Stochastic Arithmetic (DSA) [12], [13] is the combination of the CESTAC method, the concept of the computational zero, and the discrete stochastic relationships.

### *B. Numerical validation of sequential codes using DSA*

The CADNA software [14], [15], [16] is a library which implements DSA in any code written in C++ or in Fortran and allows to use new numerical types: the stochastic types. In practice, classic floating-point variables are replaced by the corresponding stochastic variables, which are composed of three floating-point values and an integer to store the accuracy. The library contains the definition of all arithmetic operations and order relations for the stochastic types. For instance, let us consider an arithmetic operation $\circ \in \{+, -, *, /\}$ between two stochastic variables A and B. This arithmetic operation is performed three times on the associated floating-point values $A_i \circ B_i$, the rounding mode being randomly set to rounding towards $+\infty$ or $-\infty$. The control of accuracy is performed only on variables of stochastic type. Only exact significant digits of a stochastic variable are printed or "@.0" for a computational zero. Because all operators are redefined for stochastic variables, the use of CADNA in a program requires only a few modifications: essentially changes in the declarations of variables and in input/output statements. The CADNA software has been successfully used for the numerical validation of real-life applications [17], [25], [26], [27], [28].

Attention has been paid to rounding mode setting in terms of performance, because the rounding mode must be frequently changed. On CPU the rounding mode is determined by two bits in the Control Word, a 16-bit register in the FPU (Floating-Point Unit). At the beginning of a program using CADNA, the rounding mode is arbitrarily set to $-\infty$. Then the rounding mode is randomly changed using the CADNA *rnd_switch* function that switches the rounding mode from $+\infty$ to $-\infty$, or from $-\infty$ to $+\infty$. To reduce the cost of rounding mode changes, the *rnd_switch* function is written in assembly language and is specific to the processor and the compiler chosen. The *rnd_switch* function changes in the FPU Control Word the two bits associated with the rounding mode. For performance reasons, a random number generator is not called at each arithmetic operation. A long random sequence is generated at the beginning of the program and stored in an array. Then successive array elements are used cyclically when random numbers are required.

CADNA can detect numerical instabilities which occur during the execution of the code. When a numerical instability is detected, dedicated CADNA counters are incremented. At the end of the run, the value of these counters together with appropriate warning messages are printed on standard output. These warnings are of two types.

1) Warnings related to the self-validation of CADNA. These include: unstable multiplication where the two operands are computational zeroes and unstable division where the divisor is a computational zero. These warnings indicate that the validity of $C_{\overline{R}}$ has been compromised and the CADNA results cannot be relied on.
2) Warnings concerning other numerical instabilities. These instabilities can occur in overloaded mathematical functions or in branching statements involving a computational zero. A numerical instability is also reported in the case of a cancellation, *i.e.* the subtraction of two very close values which generates a sudden loss of accuracy.

At the end of the run, each type of anomaly together with their occurrences are printed. If no anomaly has been detected the computed results are reliable and the accuracy of each has been correctly estimated up to a certain probability. Otherwise the messages need to be analysed, the source of the anomaly identified and, if necessary, the code changed. The user can specify the instabilities to be detected. One may choose, for instance, to activate only self-validation, to detect all types of instabilities or to deactivate the detection of instabilities.

### *C. Numerical validation of hybrid CPU-GPU codes using DSA*

An asynchronous implementation of the CESTAC method for the estimation of rounding errors in GPU codes written in CUDA is proposed in [29]. Unfortunately with such an implementation of the CESTAC method, the whole code is executed several times with the random rounding mode and no instability in arithmetic operations can be detected. We present here a version of CADNA which implements DSA for the numerical validation of hybrid CPU-GPU codes written in

CUDA. This version differs from the sequential version in two main respects: rounding mode change and instability detection.

On GPU an arithmetic operation can be performed with a specified rounding mode. For instance a multiplication with rounding towards $+\infty$ can be executed using the *fmul_ru* function and a multiplication with rounding towards $-\infty$ using the *fmul_rd* function. Therefore a stochastic operation on GPU implies three floating-point operations randomly rounded towards $+\infty$ or $-\infty$ using the appropriate arithmetic function. Unlike on CPU, random numbers are not stored in a global array on GPU, because it is incompatible with GPU programming paradigms. Each GPU thread being independent, it generates random numbers using its own seed and taking into account its own thread index. On GPU, random numbers are generated when they are required, *i.e.* during the stochastic operations. As a remark, because of the robustness of accuracy estimation by DSA [22], the quality of the random number generator is not a critical issue: only boolean values are required.

On CPU, numerical instabilities that occur during the execution are counted. Such a count is not performed on GPU because it would consume shared memory and require many atomic operations. On GPU an unsigned char is associated with each result to store the numerical instabilities that have affected it. Each bit of this char is associated with a type of instability. For instance, its last bit is set to 1 if the result has been affected by at least one unstable multiplication. Therefore in the CPU-GPU version of CADNA a stochastic variable is composed of three floating-point variables and four unsigned char: one for the accuracy, one for the instabilities and two for padding to respect memory alignment. Instability detection increases the execution time with CADNA. Cancellation detection is particularly costful because it requires to compare for all additions and subtractions the operands accuracy with the result accuracy. For performance reasons, the main instabilities can be detected with the GPU version of CADNA: the instabilities related to the self-validation of CADNA (unstable multiplications and unstable divisions) and the unstable branching statements.

## IV. Estimation of reproducibility in wave propagation results by means of DSA

The acoustic wave propagation code has been executed with the CADNA library on CPU and on GPU. Results presented in this section have been computed in the following environments:

- an AMD Opteron 6168 CPU with gcc 4.7.2 compiler;
- an NVIDIA C2050 GPU with CUDA 5.0 platform.

With implementations (1) and (2) of the finite difference scheme, the number of exact significant digits in the results computed with CADNA varies from 0 to 7. On CPU its mean value is 4.06 with both schemes; on GPU it is 3.43 with scheme (1) and 3.49 with scheme (2). These remarks are consistent with the observations described in Section II. Numerical instabilities occur during the execution: 272,394 losses of accuracy due to cancellations with scheme (1) and 285,186 with scheme (2). This kind of instability is detected by the CPU version of CADNA if the subtraction of two close floating-point numbers leads to a loss of accuracy of at least 4 digits.

Table II presents results obtained on CPU and on GPU at the same points in the space domain as in Table I. These results have been computed, on the one hand, using CADNA and, on the other hand, using IEEE floating-point arithmetic with rounding to the nearest. With CADNA, only the exact significant digits, *i.e.* the digits not affected by rounding errors, are printed. Results in the first four rows in Table II have been computed using binary32 arithmetic precision. Results in the last row have been obtained in binary64 on CPU with CADNA. Although CADNA prints 11 to 14 exact significant digits in these three results, only their first 10 digits are reported in Table II. The number of exact significant digits estimated by CADNA depends on the point considered in the space domain. As already mentioned in Section III, accuracy estimation by DSA is rather pessimistic than optimistic. Because of the probabilistic aspect of DSA, the number of exact significant digits estimated by CADNA may slightly differ on CPU and on GPU. In Table II one can notice that the digits provided by CADNA in binary32 are in common with those computed in binary64. Results reported in Table II have been computed using implementation (1) of the finite difference scheme. The same digits are provided by CADNA with the other implementation, except one less digit is given on GPU for point $p_1$.

TABLE II
Results computed at three different points in the space domain with and without CADNA using implementation (1) of the finite difference scheme

| | Point in the space domain | | |
|---|---|---|---|
| | $p_1$: (0, 19, 62) | $p_2$: (50, 12, 2) | $p_3$: (20, 1, 46) |
| IEEE CPU | -1.110479 | 54.54238 | 614.1038 |
| IEEE GPU | -1.110204 | 54.54224 | 614.1046 |
| CADNA CPU | -1.1 | 54.54 | 614.104 |
| CADNA GPU | -1.11 | 54.5 | 614.10 |
| Reference | -1.108603879 | 54.54034021 | 614.1041156 |

Figures 1 and 2 present, respectively on CPU and on GPU, the number of exact significant digits estimated by CADNA in the results computed with scheme (1) with respect to their absolute values. Similar results are observed with the other scheme. The highest results (in absolute value) are affected by low rounding errors and the highest rounding errors impact negligible results. Although the same trend can be observed on CPU and on GPU, there are differences between the two distributions due to the probabilistic aspect of DSA. Depending on the point in the space domain, the number of exact significant digits may be higher on CPU or on GPU. The average difference between results accuracy on CPU and on GPU is 0.6 digit. Furthermore because of differences in the execution order of the threads, the accuracy distribution may be slightly different from one execution to another on GPU.

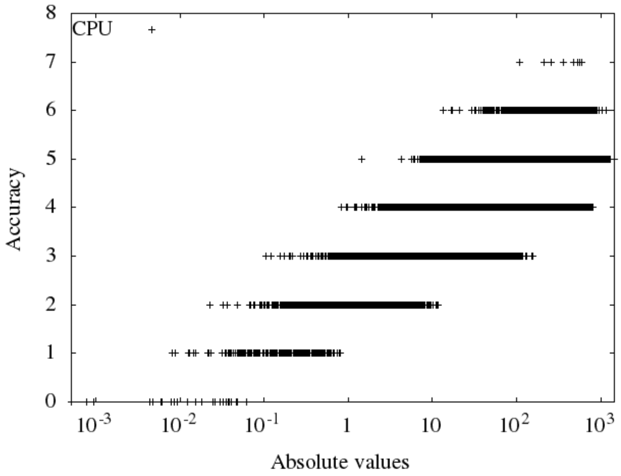


Fig. 1. Number of exact significant digits in the results computed on CPU with respect to their absolute values.

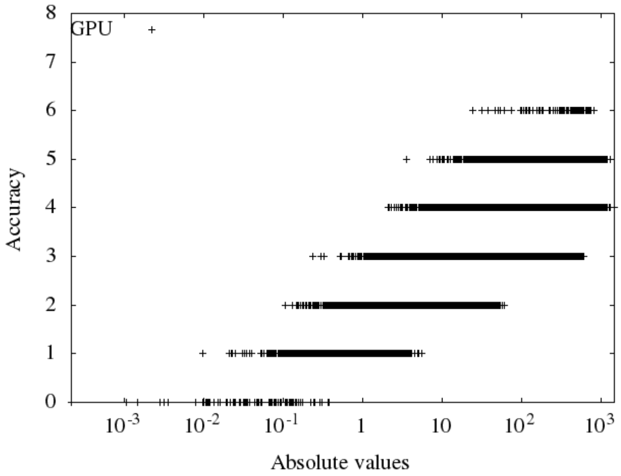


Fig. 2. Number of exact significant digits in the results computed on GPU with respect to their absolute values.

Table III presents execution times of the acoustic wave propagation code in the environments mentioned at the beginning of this section. Because the execution times measured with implementations (1) and (2) of the finite difference scheme are similar, only the performance of implementation (1) is reported in Table III. The code has been run in binary32 both on CPU and on GPU. CADNA has been used on CPU with several kinds of instability detection:

- the detection of all kinds of instabilities;
- no detection of instabilities. With this mode, which is not recommended, the execution time can be considered the minimum that can be obtained whatever instability detection chosen;
- the detection of unstable multiplications, unstable divisions and unstable branching statements. This mode, which enables the self-validation of CADNA, is also available on GPU.

One can notice that the cost of CADNA with instability detection in multiplications, divisions and branching statements is very close to its cost with no instability detection. Actually this code cannot generate such instabilities: in all multiplications at least one operand is a constant, all divisors are constants and it has no branching statement. The cost of CADNA with the detection of any kind of instability is 2.6 times higher. This is essentially due to the cancellation detection which is particularly expensive in terms of execution time. The cost of CADNA on GPU is about twice lower than on CPU with the same level of instability detection. This may be explained by the pipeline flush at each change of rounding mode on CPU which affects instruction level parallelism.

TABLE III
EXECUTION TIMES WITH AND WITHOUT CADNA ON CPU AND GPU

| CPU | | | |
|---|---|---|---|
| execution | instability detection | execution time (s) | ratio |
| IEEE | - | 110.8 | 1 |
| CADNA | all instabilities | 4349 | 39.3 |
| | no instability | 1655 | 14.9 |
| | mul., div., branching | 1663 | 15.0 |

| GPU | | | |
|---|---|---|---|
| execution | instability detection | execution time (s) | ratio |
| IEEE | - | 0.80 | 1 |
| CADNA | mul., div., branching | 5.73 | 7.2 |

## V. Conclusion

In this paper, we have shown that DSA can provide an estimation of the reproducibility of numerical programs. By estimating which digits are affected by rounding errors, DSA may explain why differences are observed in the results of a program executed in different environments. Therefore when the deployment of a code on a parallel architecture generates differences in the computed results, the presence of a bug can possibly be discarded. Based on a probabilistic approach, DSA can provide a trend of the distribution or the evolution of results accuracy. If results differences are due to different orders in the sequence of floating-point operations, a similar trend should be provided by DSA whatever the environment chosen. But a sequential implementation of DSA is not sufficient and efficient methods must be proposed for the numerical validation of large scale simulation programs. This paper has shown the feasibility of numerical validation with DSA for CPU-GPU programs. Furthermore DSA can be used for accuracy estimation in distributed memory environments [17]. It has recently been shown how to take advantage of SIMD units such as AVX (Advanced Vector eXtensions) in programs using DSA [30]. However, work must still be carried out to extend efficiently DSA to emerging computing architectures that are prone to numerical reproducibility failures.

## References

[1] O. Villa, D. Chavarría-Miranda, V. Gurumoorthi, A. Márquez, and S. Krishnamoorthy, “Effects of floating-point non-associativity on numerical computations on massively multithreaded systems,” in *Cray User Group Meeting (CUG 2009)*, Atlanta, Georgia, USA, May 2009, pp. 1–11.

[2] Y. He and C. Ding, "Using accurate arithmetics to improve numerical reproducibility and stability in parallel applications," *The Journal of Supercomputing*, vol. 18, no. 3, pp. 259–277, 2001.

[3] M. Cleveland, T. Brunner, N. Gentile, and J. Keasler, "Obtaining identical results with double precision global accuracy on different numbers of processors in parallel particle Monte Carlo simulations," *Journal of Computational Physics*, vol. 251, pp. 223–236, 2013.

[4] M. Taufer, O. Padron, P. Saponaro, and S. Patel, "Improving numerical reproducibility and stability in large-scale numerical simulations on GPUs," in *IEEE International Symposium on Parallel Distributed Processing (IPDPS)*, Atlanta, Georgia, USA, 2010, pp. 1–9.

[5] R. W. Robey, J. M. Robey, and R. Aulwes, "In search of numerical consistency in parallel programming," *Parallel Computing*, vol. 37, no. 4-5, pp. 217–229, Apr. 2011.

[6] J. Demmel and H. D. Nguyen, "Fast reproducible floating-point summation," in *21st IEEE Symposium on Computer Arithmetic (ARITH 21)*, Austin, Texas, USA, Apr. 2013, pp. 163–172.

[7] ——, "Parallel reproducible summation," in *IEEE Transactions on Computers, Special Section on Computer Arithmetic*, to appear.

[8] S. Collange, D. Defour, S. Graillat, and R. Iakymchuk, "A reproducible accurate summation algorithm for High-Performance Computing," in *SIAM Workshop on Exascale Applied Mathematics Challenges and Opportunities (EX14) held as part of the 2014 SIAM Annual Meeting*, Chicago, Illinois, USA, Jul. 2014.

[9] ——, "Reproducible and accurate matrix multiplication for High-Performance Computing," in *The 16th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics (SCAN'14)*, Würzburg, Germany, Sep. 2014.

[10] R. Iakymchuk, S. Collange, D. Defour, and S. Graillat, "Reproducible triangular solvers for high-performance computing," in *12th International Conference on Information Technology: New Generations (ITNG)*, Las Vegas, Nevada, USA, Apr. 2015. [Online]. Available: http://hal.archives-ouvertes.fr/hal-01116588v2

[11] S. Collange, D. Defour, S. Graillat, and R. Iakymchuk, "Full-speed deterministic bit-accurate parallel floating-point summation on multi- and many-core architectures," http://hal.archives-ouvertes.fr/hal-00949355v3/PDF/superaccumulator.pdf, Feb. 2015. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00949355v3

[12] J. Vignes, "A stochastic arithmetic for reliable scientific computation," *Mathematics and Computers in Simulation*, vol. 35, pp. 233–261, 1993.

[13] ——, "Discrete Stochastic Arithmetic for validating results of numerical software," *Numerical Algorithms*, vol. 37, no. 1–4, pp. 377–390, Dec. 2004.

[14] F. Jézéquel and J.-M. Chesneaux, "CADNA: a library for estimating round-off error propagation," *Computer Physics Communications*, vol. 178, no. 12, pp. 933–955, 2008.

[15] F. Jézéquel, J.-M. Chesneaux, and J.-L. Lamotte, "A new version of the CADNA library for estimating round-off error propagation in Fortran programs," *Computer Physics Communications*, vol. 181, no. 11, pp. 1927–1928, 2010.

[16] J.-L. Lamotte, J.-M. Chesneaux, and F. Jézéquel, "CADNA_C: A version of CADNA for use with C or C++ programs," *Computer Physics Communications*, vol. 181, no. 11, pp. 1925–1926, 2010.

[17] S. Montan and C. Denis, "Numerical verification of industrial numerical codes," in *ESAIM: Proc.*, vol. 35, Mar. 2012. doi: 10.1051/proc/201235006 pp. 107–113.

[18] F. Jézéquel and J.-L. Lamotte, "Numerical validation of Slater integrals computation on GPU," in *The 14th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics (SCAN'10)*, Lyon, France, Sep. 2010, pp. 78–79.

[19] H. Calandra, R. Dolbeau, P. Fortin, J.-L. Lamotte, and I. Said, "Forward seismic modeling on AMD Accelerated Processing Unit," in *Rice Oil & Gas HPC Workshop*, Houston, Texas, USA, Mar. 2013.

[20] B. Gustafsson, H.-O. Kreiss, and J. Oliger, *Time Dependent Problems and Difference Methods*, 2nd ed. Wiley, 2013.

[21] IEEE Computer Society, *IEEE Standard for Floating-Point Arithmetic*. IEEE Standard 754-2008, Aug. 2008. ISBN 978-0-7381-5752-8

[22] J.-M. Chesneaux and J. Vignes, "Sur la robustesse de la méthode CESTAC," *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, vol. 307, pp. 855–860, 1988.

[23] J.-M. Chesneaux, "L'arithmétique stochastique et le logiciel CADNA," Habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris, France, Nov. 1995.

[24] J.-M. Chesneaux and J. Vignes, "Les fondements de l'arithmétique stochastique," *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, vol. 315, pp. 1435–1440, 1992.

[25] F. Jézéquel, F. Rico, J.-M. Chesneaux, and M. Charikhi, "Reliable computation of a multiple integral involved in the neutron star theory," *Math. Comput. Simulation*, vol. 71, no. 1, pp. 44–61, 2006.

[26] N. Scott, F. Jézéquel, C. Denis, and J.-M. Chesneaux, "Numerical 'health check' for scientific codes: the CADNA approach," *Computer Physics Communications*, vol. 176, no. 8, pp. 507–521, Apr. 2007.

[27] F. Jézéquel, J.-L. Lamotte, and O. Chubach, "Parallelization of Discrete Stochastic Arithmetic on multicore architectures," in *10th International Conference on Information Technology: New Generations (ITNG), Las Vegas, Nevada (USA)*, Apr. 2013.

[28] J. Brajard, P. Li, F. Jézéquel, H.-S. Benavidès, and S. Thiria, "Numerical Validation of Data Assimilation Codes Generated by the YAO Software," in *SIAM Annual Meeting, San Diego, California (USA)*, Jul. 2013.

[29] W. Li, S. Simon, and S. Kiess, "On the numerical sensitivity of computer simulations on hybrid and parallel computing systems," in *International Conference on High Performance Computing and Simulation (HPCS)*, Istanbul, Turkey, Jul. 2011, pp. 510–516.

[30] P. Eberhart, J. Brajard, P. Fortin, and F. Jézéquel, "Towards high performance stochastic arithmetic," in *The 16th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics (SCAN'14)*, Würzburg, Germany, Sep. 2014.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 681–686

DOI: 10.15439/2015F34
ACSIS, Vol. 5

# TRACO: An Automatic Loop Nest Parallelizer for Numerical Applications

Marek Palkowski, Tomasz Klimek, Wlodzimierz Bielecki
West Pomeranian University of Technology in Szczecin
ul. Zolnierska 49, 71-210 Szczecin, Poland
Email: mpalkowski@wi.zut.edu.pl, tklimek@wi.zut.edu.pl, wbielecki@wi.zut.edu.pl

***Abstract*—We present the source-to-source TRACO compiler allowing for increasing program locality and parallelizing arbitrarily nested loop sequences in numerical applications. Algorithms for generation of tiled code and extracting synchronization-free slices composed of tiles are presented. Parallelism of arbitrary nested loops is obtained by creating a kernel of computations represented in the OpenMP standard to be executed independently on many CPUs. We consider benchmarks, typical from compute-intensive sequences of algebra operations or numerical computation from industry and engineering. The speed-up of programs generated by TRACO are discussed. Related compilers and techniques are considered. Future work is outlined.**

## I. Introduction

EFFICIENT parallel numerical algorithms for commonly occurring problems in scientific computing are more difficult to write than sequential ones. Developers must analyze their performance, granularity and scalability. Optimizing and parallelizing compiler research with empirical evaluation is significant for an efficient usage of widely available multi-core systems.

Because for many numerical kernels and solvers most computations are contained in program loop nests, automatic extraction of parallelism available in loop nests is extremely important for multi-core processing. However, there is a lack of automated and completed tools permitting for exposing parallelism in serial programs. The most advanced approach to improve program locality and parallelization is based on the Affine Transformation Framework (ATF). Unfortunately, this approach can fail to parallelize loop nests exposing storage-related dependences, and as consequence potential parallelism is left unexploited in some cases [1]. This paper presents an alternative approach to increasing program locality and parallelization which is implemented in the open source tool, TRACO [2], based on calculating the **TRA**nsitive **C**l**O**sure of dependence graphs. It currently parallelizes loop nests being written in the C language.

The purpose of this source-to-source compiler is to automatically convert existing serial numerical applications to parallel ones to be run on multicore systems and high performance computers. It produces parallel target code that is semantically identical with original source code.

## II. Background

The source-to-source TRACO compiler implements Iteration Space Slicing (ISS) techniques together with the free-scheduling, variable privatization and parallel reduction techniques. Output code, produced by TRACO, is compilable and contains OpenMP directives [3]. TRACO is available at the website http://traco.sourceforge.net.

ISS was introduced by Pugh and Rosser [4]. It takes dependence information as input to find all statement instances that must be executed to produce the correct values for the specified array elements. Dependences available in a loop nest are described by dependence relations with constraints presented by means of the Presburger arithmetic that is the first-order theory of the integers in the language $L$ having 0, 1 as constants, +,- as binary operations, and equality =, order $<$ and congruences $\equiv_n$ modulo all integers $n$ $\geq$1 as binary relations.

Coarse-grained code is presented with synchronization-free slices. TRACO uses the dependence analysis [5] proposed by Pugh and Wonnacott where dependences are represented by dependence relations. This analysis is implemented in Petit [6].

Standard operations on relations and sets are used, such as intersection ($\cap$), union ($\cup$), difference (-), domain (dom $R$), range (ran $R$), relation application ($S'$= $R(S)$: $e'\in S'$iff exists $e$ s.t. $e\to e'\in R,e\in S$), positive transitive closure of relation $R$, $R$+ = $\{[e]\to[e']$ : $e\to e'\in$ $R$ $\vee$ $\exists$ $e$ $''$, $e\to$e$''\in$ $R$ $\wedge$ $e''\to e'\in$ $R$+$\}$, transitive closure $R$* = $R$+ $\cup$ $I$. In detail, the description of these operations is presented in papers [5], [7].

The positive transitive closure for a given relation $R$, $R^+$, is defined as follows [7]
$R^+ = \{e \to e' :\ e \to e' \in R \vee \exists e'' s.t.\ e \to e'' \in R \wedge e'' \to e' \in R^+\}$.
It describes which vertices $e'$ in a dependence graph (represented by relation $R$) are connected directly or transitively with vertex $e$.

Transitive closure, $R^*$, is defined as follows [8]: $R^* = R^+ \cup I$, where $I$ the identity relation. It describes the same connections in a dependence graph (represented by $R$) that $R^+$ does plus connections of each vertex with itself.

To perform operations on sets and relations as well as calculating transitive closure, TRACO uses the ISL library [9].

978-83-60810-66-8/$25.00©2015, IEEE

## III. Iteration Space Slicing for Parallelism Extraction

ISS algorithms, presented in paper [1], allow us to generate parallel code representing synchronization-free slices. An (iteration-space) slice is defined as follows.

**Definition 1**. Given a dependence graph defined by a set of dependence relations, a slice *S* is a weakly connected component of this graph, i.e., a maximal sub-graph such that for each pair of vertices in the sub-graph there exists a forward or backward path.

**Definition 2**. An ultimate dependence source is a source that is not the destination of another dependence. Given a dependence relation *R,* describing all the dependences in a loop, set, $S_{UDS}$, including all ultimate dependence sources can be calculated as domain(*R*) - range(*R*).

**Definition 3**. The representative of a slice is its lexicographically minimal ultimate source.

An approach to extract synchronization-free slices takes two steps [1]. First, for each slice, a representative statement instance is defined (it is the lexicographically minimal statement instance from all the sources of a slice). Next, slices are reconstructed from their representatives and code scanning these slices is generated.

Given a dependence relation *R,* describing all the dependences in a loop, we calculate set of all ultimate dependence sources of slices , $S_{UDS}$, as follows

$$S_{UDS} = domain(R) - range(R). \tag{1}$$

In order to find elements of $S_{UDS}$ that are representatives of slices, we build a relation, $R_{USC}$, that describes all pairs of the ultimate dependence sources being transitively connected in a slice, as follows:

$$R_{USC} = \{[e] \to [e'] : e, e' \in S_{UDS}, e \prec e', (R^*(e) \cap R^*(e'))\}. \tag{2}$$

The condition ($e \prec e'$) in the constraints of relation $R_{USC}$ above means that *e* is lexicographically smaller than *e*′. Such a condition guarantees that the lexicographically smallest element from *e* and *e*′ will always appear in the input tuple of $R_{USC}$ (its representative source), i.e., it can never appear in the output tuple. The intersection (*R**(*e*)∩*R**(*e*′)) in the constraints of $R_{USC}$ guarantees that vertices *e* and *e*′ are transitively connected, i.e., they are the sources of the same slice.

Next, set, $S_{repr}$, containing representatives of each slice is found as $S_{repr}$ = $S_{UDS}$ - range($R_{USC}$). Each element *e* of set $S_{repr}$ is the lexicographically minimal statement instance of a synchronization-free slice. If *e* is the representative of a slice with multiple sources, then the remaining sources of this slice can be found applying relation ($R_{USC}$)* to *e*, i.e., ($R_{USC}$)*(*e*). If a slice has the only source, then ($R_{USC}$)*(*e*)=*e*. The elements of a slice represented with *e* can be found applying relation *R** to the set of sources of this slice:

$$S_{slice} = R^*((R_{USC})^*(e)). \tag{3}$$

To generate code, we insert in the first positions of the tuple of set $S_{slice}$ the elements of the tuple of set $S_{repr}$ (together with corresponding constraints) and apply the CLooG library [10] to so extended set $S_{slice}$ .

To parallelize loop nests which expose a single synchronization-free slice, time partitioning can be applied. The algorithm, presented in our paper [11], allows us to generate time partitions and corresponding fine-grained parallel code on the basis of the free schedule; all statement instances of a time partition can be executed in parallel, while partitions are enumerated sequentially. The free schedule function is defined as follows.

**Definition 4** [12]. The *free schedule* is the function that assigns discrete time of execution to each loop statement instance as soon as its operands are available, that is, it is mapping $\sigma$:*LD*→ $\mathbb{Z}$ such that

$$\sigma(p) = \begin{cases} 0 \; if \; there \; is \; no \; p_1 \in LD \; s.t. \; p_1 \to p \\ 1 + max(\sigma(p_1), \sigma(p_2), ..., \sigma(p_n)); \\ \qquad\qquad p, p_1, p_2, ..., p_n \in LD; \\ p_1 \to p, p_2 \to p, ..., p_n \to p, \end{cases}$$

where $p, p_1, p_2, ..., p_n$ are loop statement instances, *LD* is the loop domain, $p_1 \to p, p_2 \to p, ..., p_n \to p$ mean that the pairs $p_1$ and $p$, $p_2$ and $p$, ...,$p_n$ and $p$ are dependent, $p$ represents the dependence destination, while $p_1, p_2, ..., p_n$ represent the sources of dependences, *n* is the number of operands of statement instance *p* (the number of dependences whose destination is statement instance *p*). The free schedule is the fastest legal schedule [12]. In paper [11] we presented fine-grained parallelism extraction based on the power *k* of relation *R*.

The idea of the algorithm is the following [11]. Given relations $R_1$, $R_2$, ..., $R_m$, representing all dependences in a loop nest, we first calculate $R = \bigcup_{i=1}^{m} R_i$ and then $R^k$, where $R^k = \underbrace{R \circ R \circ ...R}_{k}$, "∘" is the composition operation. Techniques of calculating the power *k* of relation *R* are presented in the following publications [8], [9], [14] and they are out of the scope of this paper. Let us only note that given transitive closure $R^+$, we can easily convert it to the power *k* of *R*, $R^k$, and vice versa, for details see [9].

Given set *UDS* comprising all loop nest statement instances that are ready to execution at time *k*=0, each vertex, belonging the set $S_k$ = $R^k(UDS)$ - $R^+ \circ R^k(UDS)$, is connected in the dependence graph, defined by relation *R*, with some vertex(ices) represented by set *UDS* with a path of length *k* . Hence at time *k*, all the statement instances belonging to the set $S_k$ can be scheduled for execution and it is guaranteed that *k* is as few as possible.

## IV. Loop Tiling

Tiling is a very important iteration reordering transformation for both improving data locality and extracting loop nest parallelism. TRACO allows users generate parallel tiled code by means of algorithms based on the transitive closure of a dependence graph [13].

First, we form set *TILE(**II**, **B**)* including iterations belonging to a parametric tile as follows

*TILE*(***II**, **B***) = {[***I***] | ***B***II* +***LB*** $\leq$ ***I*** $\leq$ min( ***B****(***II*** +**1**) + ***LB*** -**1**, ***UB***) AND ***II*** $\geq$ 0}, where vectors ***LB*** and ***UB*** include the lower and upper loop index bounds of an original loop nest, respectively; diagonal matrix ***B*** defines the size of a rectangular original tile; elements of vectors ***I*** and ***II*** represent the original loop nest indices and the identifiers of tiles, respectively; **1** is the vector whose all elements have value 1.

*TILE*(***II**, **B***) represents a tile of the rectangular shape with a fixed size defined by the user.

Sets *TILE_LT* and *TILE_GT* are the unions of all the tiles whose identifiers are lexicographically less and greater than that of *TILE*(***II***, ***B***), respectively.

*TILE_LT* ={[***I***] | exists ***II***$'$ s. t. ***II***$'$ $\prec$ ***II*** AND ***II***, ***II***$'$ in *II_SET* AND ***I*** in *TILE*(***II***$'$, ***B***)},

*TILE_GT* ={[***I***] | exists ***II***$'$ s. t. ***II***$'$ $\succ$ ***II*** AND ***II***, ***II***$'$ in *II_SET* AND ***I*** in *TILE*(***II***$'$, ***B***)}.

Set *TILE_ITR = TILE - R*$^+$( *TILE_GT* ) does not include any invalid dependence target.

Set *TVLD_LT* = ( *R*$^+$(*TILE_ITR*) $\cap$ *TILE_LT*) - *R*$^+$(*TILE_GT*) includes all the iterations that i) belong to the tiles whose identifiers are lexicographically less than that of set *TILE_ITR*, ii) are the targets of the dependences whose sources are contained in set *TILE_ITR*, and iii) are not any target of a dependence whose source belong to set *TILE_GT*.

Set *TILE_VLD = TILE_ITR* $\cup$ *TVLD_LT* represents target tiles.

To generate code, we form set *TILE_VLD_EXT* by means of inserting i) into the first positions of the tuple of set *TILE_VLD* indices $ii_1, ii_2, ..., ii_d$; ii) into the constraints of set *TILE_VLD* the constraints defining tile identifiers ***II*** $\geq$ 0 and ***B***II*+***LB*** $\leq$ ***UB***.

The resulting code can be produced by means of applying any code generator to scan elements of set *TILE_VLD_EXT* in the lexicographic order, for example, CLooG [10]. TRACO implements an extended approach allowing for tiling imperfectly nested loops also.

All the presented algorithms implemented in the TRACO compiler are based on the transitive closure of a dependence relation representing all dependences in a loop nest. Loop nest tiling and iteration space slicing can be combined in order to increase program locality and the grain size of parallel code. For this purpose, we form relation *R_TILE* that describes dependences among all tiles but ignores dependences within each tile as follows.

*R_TILE*:={[***II***]->[***JJ***]: exist ***I***, ***J*** s.t. (***II**, **I***) in *TILE_VLD_EXT*(***II***) *AND* (***JJ**, **J***) in *TILE_VLD_EXT*$_i$(***JJ***) *AND* ***J*** in *R*(***I***)},

where ***II**, **JJ*** are the vectors representing tile identifiers. Such a relation can be used to extract slices comprising tiles or free-scheduling in the same way as it is described in Section III except from instead of relation $R$ relation *R_TILE* has to be used.

## V. TRACO USAGE FOR THE HYDRO-FRAGMENT CODE

In this section, we present the usage of the TRACO compiler to optimize the code of the hydrodynamics fragment of the first kernel of the Livermoore Loops suite (*k1*) [15].

```
for ( l=1 ; l<=loop ; l++ ) {
 for ( k=0 ; k<n ; k++ ) {
  x[k] = q + y[k]*( r*z[k+10] + t*z[k+11] );
 }
}
```

The following relation describes dependences in this program.

*R* := [n,loop] -> { [l,k,13] -> [l',k',13] :( k' = k and 1 <= l < l' <= loop and 0 <= k < n and l < loop and 1 <= loop and 2 <= l' ) },

where here and further on "13"states for the loop nest statement identifier, defined by the line number of this statement in the source code.

First of all, we generate parallel synchronization-free code. For this purpose, we calculate the transitive closure of relation *R*, $R^*$:

$R^*$ := [n, loop] -> { [l, k, 13] -> [l', k, 13] : l >= 1 and k >= 0 and k <= -1 + n and l' >= 1 + l and l' <= loop; [l, k, v] -> [l, k, v] }.

Next, we expose slice representatives. Because $R_{UCS}$ = $\oslash$, we have

S$_{REPR}$ =*UDS*:= {[n, loop] -> [0, i1, 13] : loop >= 17 and i1 >= 0 and 16i1 <= -1 + n }.

Now, we form set, $S_{slice}$, representing slices:

$S_{slice}$ := [n, loop, c1] -> ] [i0, c1, 13] : c1 >= 0 and c1 <= -1 + n and i0 <= loop and loop >= 2 and i0 >= 1 },

and insert in the first positions of the tuple of set $S_{slice}$ the elements of the tuple of set $S_{REPR}$(together with corresponding constraints) and apply to so extended set $S_{slice}$ CLooG to get the following parallel code.

```
if (n + loop >= 3 && loop >= 1)
#pragma omp parallel for
  for (c1 = 0; c1 < n; c1 += 1)
     if (c1 >= 0 && loop >= 2 && n >= c1 + 1)
        for (c0 = 1; c0 <= loop; c0 += 1)
        x[c1]=q+y[c1]*(r*z[c1+10]+t*z[c1+11]);
```

Below, we demonstrate how tiled code can be generated for the same example. First, we define a rectangular parametric tile of the size 16x16 as follows.

*TILE*:= [ll, kk, n, loop] -> { [l, k, 13] : l >= 1 + 16ll and l >= 1 and l <= loop and l <= 16 + 16ll and k >= 16kk and k >= 0 and k <= -1 + n and k <= 15 + 16kk and ll >= 0 and kk >= 0 and loop >= 1 and n >= 1 }.

Next, we calculate the following sets:

*TILE_LT*:= [ll, kk, n, loop] -> { [l, k, 13] : l >= 1 + 16ll and l >= 1 and l <= loop and l <= 16 + 16ll and k >= 0 and k <= -1 + 16kk and n >= 1 + 16kk and ll >= 0 and loop >= 16ll and kk >= 0 },

*TILE_GT*:= [ll, kk, n, loop] -> { [l, k, 13] : l >= 1 + 16ll and l >= 1 and l <= loop and l <= 16 + 16ll and k >= 16 + 16kk and k <= -1 + n and ll >= 0 and loop >= 16ll and kk >= 0 and n >= 1 + 16kk , *TILE_ITR*:= [n, loop, ll, kk] -> [l, k, 13] : ll >= 0 and kk >= 0 and l >= 1 + 16ll and l <= 16 + 16ll and l <= loop and k >= 16kk and k <= 15 + 16kk and k <= -1 + n },

*TVLD_LT* = $\oslash$, *TILE_VLD* = *TILE_ITR*,

*TILE_VLD_EXT*:= [n, loop] -> { [i0, i1, i2, i3, 13] : i0 >= 0 and i1 >= 0 and i2 >= 1 + 16i0 and i2 <= 16 + 16i0 and i2 <= loop and i3 >= 16i1 and i3 <= 15 + 16i1 and i3 <= -1 + n }.

Applying CLooG to set *TILE_VLD_EXT,* we generated the following tiled code (without parallelism).

```
for(c0=0; c0 <= (loop-1)/16; c0 += 1)
 for(c1=0; c1 <= (n-1)/16; c1 += 1)
  for(c2=16*c0+1; c2<=min(loop,16*c0+16); c2++)
   for(c3=16*c1; c3<=min(n-1,16*c1+15); c3++)
    x[c3]=q+y[c3]*(r*z[c3+10]+t*z[c3+11]);
```

To generate parallel synchronization-free tiled code, we calculate tile representatives (the lexicographically minimal tiles of slices) comprised in set *TILE_SOUR* and a relation *R_TILE* that describes dependences among all tiles but ignores dependences within each tile as follows. Next, the transitive closure of that relation, $R\_TILE^*$, is calculated, and finally set, *SLICE*, representing slices comprising tiles are extracted. The corresponding sets and relations for the loop nest above are as follows.

*TILE_SOUR*:= [n, loop] -> { [ i1, i3, 13] : loop >= 17 and i1 >= 0 and i3 >= 16i1 and i3 <= 15 + 16i1 and i3 <= -1 + n }, where "13"is the statement identifier,

*R_TILE:*= [n, loop] -> { [i0, i1, 13] -> [o0, i1, 13] : i0 >= 0 and i1 >= 0 and 16o0 <= -1 + loop and o0 >= 1 + i0 and 16i1 <= -1 + n and 16i0 <= -2 + loop },

$R\_TILE^*$:= [n, loop] -> { [i0, i1, 13] -> [o0, i1, 13] : i0 >= 0 and i1 >= 0 and 16o0 <= -1 + loop and o0 >= 1 + i0 and 16i1 <= -1 + n and 16i0 <= -2 + loop },

*SLICE:*= [n, loop, c1] -> { [i0, i1, i2, i3, 13] : loop >= 17 and i2 <= 16 + 16i0 and i2 <= loop and i3 >= 16c1 and i3 >= 0 and i3 <= 15 + 16c1 and i3 <= -1 + n and i3 <= 15 + 16i1 and i3 >= 16i1 and i2 >= 1 and i2 >= 1 + 16i0 }.

Applying CLooG to set *SLICE* and preprocessing the code returned by CLooG, we get the following OpenMP C parallel code, where line 1 represents the OpenMP *parallel for* directives pointing out that the *for* loop in line 2 can be executed in parallel; line 2 and line 4 include the *for* loops enumerating tile identifiers whereas line 5 and line 6 present *the for* loops scanning statement instances within a tile whose identifier is defined by the indices of the loops in lines 2 and 4.

```
1.#pragma omp parallel for
2.for (c1 = 0; c1 <= floord(n - 1, 16); c1++)
3. if (loop >= 17)
4.  for (c0 = 0; c0 <= (loop-1)/16; c0 += 1)
5.   for(c2=16*c0+1;c2<=min(16*c0+16,loop);c2++)
6.    for(c3=16*c1; c3<=min(n-1,16*c1+15); c3++)
7.      x[c3]=q+y[c3]*(r*z[c3+10]+t*z[c3+11]);
```

## VI. Related work

Well-known automatic parallelization of numerical algorithms is based on the polyhedral model, which provides an abstraction to perform high-level transformations such as loop-nest optimization and parallelization on affine loop nests. The polyhedral source to source tools: Pluto [16], POCC and PTile [17] transform C programs to expose parallelism and improve data locality simultaneously. The core transformation framework mainly works by finding affine transformations for efficient loop nest tiling and fusion, but not limited to those.

TABLE I
Numerical programs

| Benchmark | Description |
|---|---|
| advect3d | Advection Kernel for Weather Modeling |
| correl | Correlation Computation |
| dct | Discrete Cosinus Transform |
| doitgen | Multi resolution Analysis Kernel |
| dsyr2k | Symmetric rank-2k operations |
| gemver | Vector Multiplication and Matrix Addition |
| k6 | General Linear Recurrence Equations |
| tce | Tensor Contraction Expressions |

The polyhedral model approach includes the following three steps: i) program analysis aimed at translating high level codes to their polyhedral representation and to provide data dependence analysis based on this representation, ii) program transformation with the aim of improving program locality and/or parallelization, iii) code generation [18], [19], [20], [21], [22]. All above three steps are available in the approach presented in this paper. But there exists the following difference in step ii): in the polyhedral model "*a (sequence of) program transformation(s) is represented by a set of affine functions, one for each statement*" [23] while the presented approach does not find and use any affine function. It applies the transitive closure of a program dependence graph to specific subspaces of the original loop nest iteration space. At this point of view the program transformation step is rather within the Iteration Space Slicing Framework introduced by Pugh and Rosser [4]: "*Iteration Space Slicing takes dependence information as input to find all statement instances from a given loop nest which must be executed to produce correct values for the specified array elements* ". The key step in Iteration Space Slicing is calculating the transitive closure of a loop nest dependence graph.

To extract affine transformations, the polyhedral model assumes that first "time-partition constraints" are to be formed, then a solution to them has to be found. The "time-partition constraints" [18], [19], [24] represent the condition that if one iteration is depend upon the other, then the first must be assigned to a time that is no earlier than that of the second; if they are assigned to the same time, then the first has to be executed after the second. If there exist more than one linearly independent solutions to the time-partition constraints formed for a loop nest, then it is possible to derive affine transformations allowing for loop nest parallelization and program locality improvement. Otherwise, the polyhedral model fails to expose parallelism and improve program locality.

Some limitations of affine transformation are considered in paper [1]. The main drawback of the affine transformation framework is that there does not always exist two or more

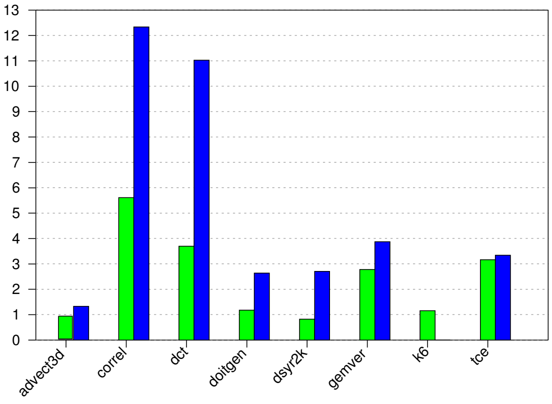


Fig. 1. Speed-up1 for tiled codes (the green bars) and speed-up2 for parallel tiled codes (the blue bars) of the studied numerical programs.

independent solutions to the time-partition constraints. For example, the studied numerical algorithm *k6* (General Linear Recurrence Equations) cannot be tiled or parallelized by PLUTO or other polyhedral tools because there exists the only solution to corresponding time-partition constraints. To extract parallelism, TRACO needs only the transitive closure of a dependence graph, so it lacks the drawbacks inherent for affine transformations.

## VII. Numerical programs

In this section, we present numerical programs chosen for carrying out experiments (Table I). We consider applications of numerical algorithms in industry and engineering: *advect3d*, *tce*, and *dct*; typical programs from compute-intensive sequences of algebra operations: *correl*, *doitgen*, *dsyr2k*, *gemver*, and *k6*. Source codes of these numerical programs are available at the TRACO website [2].

*advect3d* is the Runga-Kutta advection core from the NCOMMAS code for mesoscale weather modeling [25]. *tce* is a sequence of four nested loops, occurring in Tensor Contraction Expressions that appear in computational quantum chemistry problems [26]. Discrete cosinus transform (*dct*) is important to numerous applications in science and engineering, from lossy compression of audio and images to spectral methods for the numerical solution of partial differential equations.

The reminding benchmarks are linear algebra programs: *gemver* is a composition of BLAS operations used for householder bidiagonalization, *doitgen* is an in-place 3D-2D matrix product, and symmetric rank-2k operations *dsyr2k*. General linear recurrence equations *k6* is a kernel from the Livermore loops [15]. *correl* creates a correlation matrix and is used also in data mining [27].

The computations of *k6*, *dsyr2k*, *advect3d*, *tce*, and *gemver* are represented by perfectly nested loops, while *doitgen*, *dct*, and *correl* are represented by imperfectly nested loops.

## VIII. Experiments

We have applied TRACO to each of the benchmarks, presented in Table 1, to generate serial tiled and parallel synchronization-free tiled code. For the *k6* program, TRACO is able to generate only serial tiled code (for this benchmark, PLUTO fails to generate tiled and any parallel code).

To run programs, we have used a computer with Intel i5-4670 3.40 GHz processors (Haswell, 2013), 6MB cache and 8GB RAM. Table II presents execution time for original (*t_untiled*), serial tiled (*t_tiled*) and parallel tiled (*t_par_tiled*) applications (for 4 CPUs) for the different problem size and tile (block) size as well as the speed up of tiled code, where *speed-up1=t_untiled/t_tiled*, *speed-up2=t_untiled/t_par_tiled*, *speed-up3=t_tiled/t_par_tiled.*

Analysing the data in Table II, we may conclude that serial tiled code for each program, except from *advect3d* and *dsyr2k*, demonstrates positive speed-up (see speed-up1), for the *correl* program, speed-up is equal about 5,5. This fact can be explained by increasing tiled code locality in comparison with that of untiled code.

Serial tiled code does not increase the locality of the original *advect3d* and *dsyr2k* programs, so the corresponding serial tiled programs are not faster than the original ones.

All parallel synchronization-free tiled programs expose positive speed-up (speed-up2 and speed-up3). Two parallel programs, *dct* and *corcol,* demonstrate super linear speed-up2 ( for four CPUs, the speed-up of those programs is about 12 and 11, respectively). The reason for super-linear speed-up is that the working set of a problem is greater than the cache size when executed sequentially, but can fit nicely in each available cache when executed in parallel.

The best program speed-ups (speed-up2) are presented in a graphical way in Figure 1. All parallel tiled programs, generated by TRACO, are faster than serial tiled code – see speed-up3.

## IX. Conclusion

In this paper, we presented applying the transitive closure of dependence graphs, implemented in TRACO, for automatic producing both serial and parallel tiled code for chosen numerical applications. Loop nest computations are divided into multiple slices which are mapped to processors as threads. TRACO allows users to achieve significant speed-up of parallel numerical algorithms on shared memory machines with multi-core processors. The effectiveness of applying TRACO is comparable or better than that of other well-known optimizing compilers.

In the future, we plan to implement in TRACO techniques allowing for tiled code scalability and tile shapes different from the rectangular one, first of all parallelepiped tiles that will allow us to increase parallelism degree of parallel tiled code.

## References

[1] A. Beletska, W. Bielecki, A. Cohen, M. Palkowski, K. Siedlecki, “Coarse-grained loop parallelization: Iteration space slicing vs affine transformations”. *Parallel Computing*, vol. 37, pp. 479–497, 2011.

[2] The TRACO Compiler, http://traco.sourceforge.net, 2015.

[3] OpenMP Specification, version 3.1, http://www.openmp.org, 2014.

TABLE II
EXECUTION TIMES OF ORIGINAL, TILED AND PARALLEL TILED CODE, SPEED-UP FOR THE STUDIED NUMERICAL PROGRAMS.

| loop | size | block | t_untiled | t_tiled | t_par_tiled | speed-up1 | speed-up2 | speed-up3 |
|---|---|---|---|---|---|---|---|---|
| advect3d | 200 | 16 | 0.144 | 0.185 | 0.112 | 0.778 | 1.286 | 1.652 |
| | | 32 | | 0.164 | 0.098 | 0.878 | 1.469 | 1.673 |
| | 300 | 16 | 0.421 | 0.496 | 0.356 | 0.849 | 1.183 | 1.393 |
| | | 32 | | 0.445 | 0.317 | 0.946 | 1.328 | 1.403 |
| correl | 1000 | 16 | 0.730 | 0.383 | 0.222 | 1.906 | 3.288 | 1.725 |
| | | 32 | | 0.364 | 0.164 | 2.005 | 4.451 | 2.220 |
| | 1200 | 16 | 3.553 | 0.728 | 0.320 | 4.880 | 11.103 | 2.275 |
| | | 32 | | 0.633 | 0.288 | 5.613 | 12.337 | 2.220 |
| dct | 512 | 16 | 0.277 | 0.205 | 0.116 | 1.351 | 2.388 | 1.776 |
| | | 32 | | 0.174 | 0.060 | 1.592 | 4.617 | 2.900 |
| | 1024 | 16 | 4.884 | 1.520 | 0.655 | 3.213 | 7.456 | 2.320 |
| | | 32 | | 1.321 | 0.443 | 3.697 | 11.025 | 2.981 |
| doitgen | 250 | 32 | 3.369 | 2.689 | 2.420 | 1.253 | 1.392 | 1.111 |
| | | 64 | | 2.868 | 1.278 | 1.175 | 2.636 | 2.244 |
| | 350 | 32 | 12.806 | 10.797 | 8.050 | 1.186 | 1.591 | 1.341 |
| | | 64 | | 10.777 | 5.604 | 1.188 | 2.285 | 1.923 |
| dysr2k | 1024 | 16 | 2.037 | 2.484 | 0.753 | 0.820 | 2.705 | 3.298 |
| | | 32 | | 1.977 | 1.347 | 1.030 | 1.512 | 1.467 |
| | 1536 | 16 | 6.702 | 6.341 | 4.529 | 1.057 | 1.480 | 1.400 |
| | | 32 | | 7.523 | 2.933 | 0.891 | 2.285 | 2.565 |
| gemver | 6000 | 200 | 0.645 | 0.254 | 0.190 | 2.539 | 3.395 | 1.336 |
| | | 400 | | 0.260 | 0.224 | 2.481 | 2.879 | 1.160 |
| | 10000 | 200 | 2.038 | 0.734 | 0.526 | 2.777 | 3.875 | 1.395 |
| | | 400 | | 1.041 | 0.689 | 1.958 | 2.958 | 1.510 |
| k6 | 1500 | 16 | 12.456 | 10.778 | - | 1.156 | - | - |
| | | 64 | | 11.166 | - | 1.116 | - | - |
| | 2000 | 16 | 35.723 | 34.189 | - | 1.045 | - | - |
| | | 64 | | 34.21 | - | 1.044 | - | - |
| tce | 200 | 32 | 0.293 | 0.226 | 0.189 | 1.296 | 1.550 | 1.195 |
| | | 64 | | 0.256 | 0.166 | 1.145 | 1.765 | 1.542 |
| | 300 | 32 | 16.779 | 8.518 | 4.944 | 1.970 | 3.394 | 1.722 |
| | | 64 | | 5.304 | 5.019 | 3.163 | 3.343 | 1.056 |

[4] W. Pugh, E. Rosser, "Iteration space slicing and its application to communication optimization". *In International Conference on Supercomputing*, pp. 22–228, 1997.

[5] W. Pugh, D. Wonnacott, "An exact method for analysis of value-based array data dependences". In Sixth Annual Workshop on Programming Languages and Compilers for Parallel Computing, Springer-Verlag, 1993.

[6] W. Kelly et. al., "New User Interface for Petit and Other Extensions", User Guide, 1996.

[7] W. Kelly et. al., "The omega library interface guide". Technical report, College Park, MD, USA, 1995.

[8] W. Kelly, W. Pugh, E. Rosser, T. Shpeisman, "Transitive closure of infinite graphs and its applications", *Int. J. Parallel Programming*, vol. 24 (6), pp. 579–598, 1996.

[9] S. Verdoolaege, "Integer Set Library - Manual", isl.gforge.inria.fr/manual.pdf, 2015.

[10] C. Bastoul, "Code Generation in the Polyhedral Model Is Easier Than You Think, PACT'13 IEEE International Conference on Parallel Architecture and Compilation Techniques", Juan-les-Pins, France, pp. 7–16, 2004.

[11] W. Bielecki, M. Palkowski, T. Klimek, "Free scheduling for statement instances of parameterized arbitrarily nested affine loops", *Parallel Computing*, vol. 38 (9), pp. 518–532, 2012.

[12] A. Darte, Y. Robert, F. Vivien, "Scheduling and Automatic Parallelization", Birkhauser, 2000.

[13] W. Bielecki, M. Palkowski, "Perfectly nested loop tiling transformations based on the transitive closure of the program dependence graph", Soft Computing in Computer and Information Science Advances in Intelligent Systems and Computing, vol. 342, pp. 309-320, 2015.

[14] W. Bielecki, T. Klimek, M. Palkowski, A. Beletska, "An Iterative Algorithm of Computing the Transitive Closure of a Union of Parameterized Affine Integer Tuple Relations", COCOA 2010: Fourth International Conference on Combinatorial Optimization and Applications, Lecture Notes in Computer Science, vol. 6508, pp. 104–113, 2010.

[15] T. Peters, "Livermore Loops coded in C", Kendall Square Res. Corp., http://www.netlib.org/benchmark/livermorec, 1992.

[16] U. Bondhugula, et al., "A practical automatic polyhedral parallelizer and locality optimizer", SIGPLAN Not., vol. 43 (6), pp. 101–113, urlhttp://pluto-compiler.sourceforge.net, 2008.

[17] PoCC the Polyhedral Compiler Collection, pocc.sourceforge.net, 2014.

[18] P. Feautrier, "Some efficient solutions to the affine scheduling problem: I. One-dimensional time", *Int. J. Parallel Program.*, Kluwer Academic Publishers, vol. 21, pp. 313–348, 1992.

[19] P. Feautrier, "Some efficient solutions to the affine scheduling problem. Part II. Multidimensional time", *International Journal of Parallel Programming*, vol. 21, pp. 389–420, 1992.

[20] J. Ramanujam, P. Sadayappan, "Tiling Multidimensional Iteration Spaces for Multicomputers", *Journal of Parallel and Distributed Computing*, Volume 16, Issue 2, pp. 108–120, 1992.

[21] A. W. Lim, M. S. Lam, "Communication-free parallelization via affine transformations 24th ACM Symp. on Principles of Programming Languages, Springer-Verlag, pp. 392–106, 1994.

[22] U. Bondhugula, et. al., "Automatic Transformations for Communication-Minimized Parallelization and Locality Optimization in the Polyhedral Model Compiler Constructure", In Proceedings of the CC'08/ETAPS'08, Springer, pp. 132–146, 2008.

[23] M. W. Benabderrahmane, L. N. Pouchet, A. Cohen, C. Bastoul, "The polyhedral model is more widely applicable than you think". Proceedings of the 19th joint European conference on Theory and Practice of Software, International Conference on Compiler Construction, Springer-Verlag, pp. 283–303, 2010.

[24] A. Lim, G. I. Cheong, M. S. Lam, "An Affine Partitioning Algorithm to Maximize Parallelism and Minimize Communication", In Proceedings of the 13th ACM SIGARCH International Conference on Supercomputing, ACM Press, pp. 228–23, 1999.

[25] L. J. Wicker and R. B. Wilhelmson, "Simulation and analysis of tornado development and decay within a three-dimensional supercell thunderstorm". *J. Atmos. Sci.*, vol. 52, pp. 2675–2703, 1995.

[26] The Tensor Contraction Engine ,http://www.csc.lsu.edu/~gb/TCE/, 2014.

[27] The Polyhedral Benchmark suite, *http://www.cse.ohio-state.edu/ pouchet/software/polybench/*, 2014.

DOI: 10.15439/2015F16
ACSIS, Vol. 5

# Kaprekar's transformations. Part I – theoretical discussion

Edyta Hetmaniok, Mariusz Pleszczyński, Ireneusz Sobstyl, Roman Wituła
Institute of Mathematics
Silesian University of Technology
Kaszubska 23, 44-100 Gliwice, Poland
Email: {edyta.hetmaniok,mariusz.pleszczynski,roman.witula}@polsl.pl

***Abstract*—The paper is devoted to discussion of the minimal cycles of the so called Kaprekar's transformations and some of its generalizations. The considered transformations are the self-maps of the sets of natural numbers possessing $n$ digits in their decimal expansions. In the paper there are introduced several new characteristics of such maps, among others, the ones connected with the Sharkovsky's theorem and with the Erdős-Szekeres theorem concerning the monotonic subsequences. Because of the size the study is divided into two parts. Part I includes the considerations of strictly theoretical nature resulting from the definition of Kaprekar's transformations. We find here all the minimal orbits of Kaprekar's transformations $T_n$, for $n = 3, ..., 7$. Moreover, we define many different generalizations of the Kaprekar's transformations and we discuss their minimal orbits for the selected cases. In Part II (ibidem), which is a continuation of the current paper, the theoretical discussion will be supported by the numerical observations. For example, we notice there that each fixed point, familiar to us, of any Kaprekar's transformation generates an infinite sequence of fixed points of the other Kaprekar's transformations. The observed facts concern also several generalizations of the Kaprekar's transformations defined in Part I.**

## I. Introduction

SUBJECT concerning the form, description and coexistence of orbits of the given map $F\colon X \to X$ became a chart-topping object of research after popularization of the Sharkovsky's theorem ([1], [8], [9], [10], [23], [27], [28]). We shall recall it to the Readers.

Let $\mathbb{N}$ denote the set of all positive integers. The following ordering of elements of $\mathbb{N}$ is called the Sharkovsky's ordering of $\mathbb{N}$:

$$\begin{gathered}3, 5, 7, 9, \dots, 2\cdot 3, 2\cdot 5, 2\cdot 7, 2\cdot 9, \dots,\\ 2^2\cdot 3, 2^2\cdot 5, 2^2\cdot 7, 2^2\cdot 9, \dots \\ 2^k\cdot 3, 2^k\cdot 5, 2^k\cdot 7, 2^k\cdot 9, \dots, 2^4, 2^3, 2^2, 2, 1.\end{gathered}\tag{1}$$

**Sharkovsky's theorem**. *The following facts hold:*

(a) *If $f\colon [0,1] \to [0,1]$ is a continuous map then there exists $n = n(f) \in \mathbb{N} \cup \{2^\infty\} \cup \{0\}$ such that the set $Per(f)$ of periods of all periodic orbits of $f$ is equal to the set of all $m \in \mathbb{N}$ located on the right side of $n$ in the Sharkovsky's order (if $n = 2^\infty$ then, by definition, $Per(f) = \{2^k\colon k = 0, 1, 2, \dots\}$, whereas, if $n = 0$ then $Per(f) := \mathbb{N}$).*

(b) *If $n \in \mathbb{N} \cup \{2^\infty\} \cup \{0\}$ then there exists a continuous map $f\colon [0,1] \to [0,1]$ such that the set $Per(f)$ is equal to the set of all $m \in \mathbb{N}$ located on the right side of $n$ in the Sharkovsky's order and for two selected cases, $n = 2^\infty$ and $n = 0$, the set $Per(f)$ is equal to the one defined above.*

In the subject-matter referring to the Sharkovsky's theorem we know a lot at the moment and many facts have been also till now discovered, like for example the description of periodic orbits of triod (see [2]), the generalizations of Sharkovsky's theorem for hereditarily decomposable chainable continua (see [22], [25], [26]) and the new order for periodic orbits of interval maps (see [5] and references therein). Another important fact (which we intend to discuss in this study as well) concerns not only the periods of a given map but also the so called orbit type. It was at first defined by S. Baldwin in [3] for maps of an interval (see also [24] and references therein) and next extended by others (for example in [4] for the maps of a circle and in [21] for the groups and the groups of graphs). We will use here the following definition [1]. If $f\colon X \to X$, where $X \subset \mathbb{R}$ has $n$-elements (minimal) orbit $\{x_0, f(x_0), \dots, f^{n-1}(x_0)\}$, where $f^k$ denotes the $k-$times composition of $f$, then this orbit induces a cyclic permutation of order $n$, called the orbit type. More precisely, if the points of this orbit are indexed in increasing order $x_1 < x_2 < \dots < x_n$, then the respective orbit type $p$ is defined by $p(k) = j$ whenever $f(x_k) = x_j$. In other words, if $x_0 = x_{k_1}$, $f(x_0) = x_{k_2}, \dots$, $f^{(n-1)}(x_0) = x_{k_n}$, then the orbit type $p$ is equal to $(k_1, k_2, \dots, k_n)$. We note that there exists $(n-1)!$ orbit types of order $n$.

We say that the orbit type guarantees a period-3 point if any continuous function with an orbit of that type possesses a three-element orbit. Eric Lundberg proved in paper [19] that

$$\lim_{n\to\infty} \frac{\gamma_n}{(n-1)!} = 1,$$

where $\gamma_n$ denotes the number of orbit types of order $n$ that guarantees a period-3 point.

Let us emphasize that almost all the above results cannot be transformed so obviously onto many equally interesting cases of maps, even so numerically attractive like the self-maps of the finite sets.

A reason for creating this paper was the information, surprising for the Authors, about the existence of the so called Kaprekar constant [16], [17], which appeared to be,

978-83-60810-66-8/$25.00©2015, IEEE

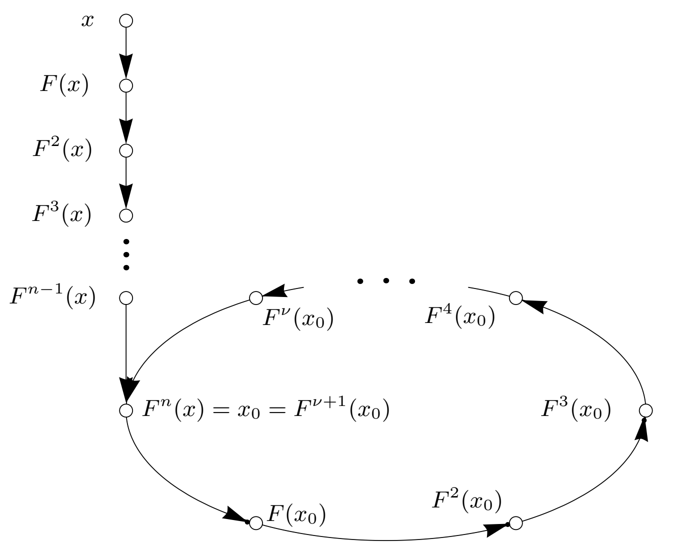


Fig. 1. Graphical illustration of a finite set $X$ and a map $F\colon X \to X$, where $X = \{F^k(x) : k \in \mathbb{N}\}$ for some $x \in X$, possessing one nontrivial and proper orbit

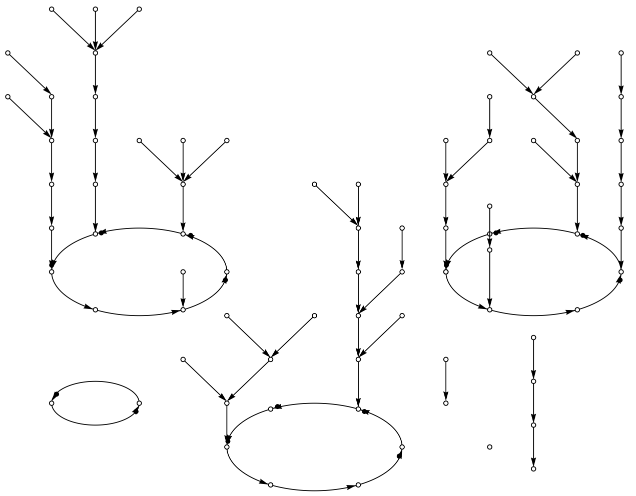


Fig. 2. Graphical illustration of any map $F : X \to X$ operating, where $X$ is a finite set of all indicated circle-points

no more no less, a single element of a single orbit of some map (we will describe this map in Section 2) onto the finite set of all natural numbers with four-digit decimal expansion. Let us notice in this moment that if $F\colon X \to X$ and $X$ is a finite nonempty set then for every $x \in X$ there exists $n \in \mathbb{N}_0$ such that $n$-th $F$-iteration of $x$, i.e. the element $F^n(x)$, belongs to some minimal orbit of $F$. This means, by definition, that certain subset of $X$ is of the form $\{x_0 = F^{\nu+1}(x_0), F(x_0), F^2(x_0), \ldots, F^\nu(x_0)\}$, where $\nu \in \mathbb{N}_0$. The above facts are illustrated in Figures 1 and 2.

Let us note that in general case there is no connection between values $n$ and $\nu$ (more precisely, for any $n, \nu \in \mathbb{N}$, for the set $X$ composed of elements – circles like in Fig.1, we construct a map described in Fig.1 proving that there is no relation between $n$ and $\nu$). However, we should remember that in the case of some specific maps (and even for the families of maps) the relation between $n$ and $\nu$ may appear!

In case when $F$ is a bijection on $X$, that is permutation on $X$, then every element of set $X$ belongs to some $F$-orbit ($F$-orbit is created by elements of each cycle of permutation $F$). Certainly, if $F$ is not a bijection on $X$ then the situation is also easy to describe, at least from the theoretical point of view, namely the set

$$\mathbb{X} := \bigcap_{k=0}^{card\, X} F^k(X)$$

is a set-theoretical union of all orbits of the map $F$, and moreover, $F$ restricted to $\mathbb{X}$ is a bijection on $\mathbb{X}$. Set $\mathbb{X}$ is the largest fixed subset of map $F$, it means if $Y \subset X$ and $F(Y) = Y$, then $Y \subset \mathbb{X}$. Henceforward we will call such set as the maxinvariant subset of $F$. The only problem in this situation is the actual form of set $\mathbb{X}$? (In Figure 2 the set $\mathbb{X}$ is equal to the union of final single points and all points located on the indicated ellipses.) Of course equally essential, although much more difficult in practice, is the description of all orbits of map $F$.

In this paper, as the input set $X$ we will take the families containing numbers 0, $10^{k-1} - 1$ and the natural numbers possessing $k$ digits in their decimal expansion, that is

$$X = X(k) = \{0\} \cup \{n \in \mathbb{N}\colon 10^{k-1} - 1 \leq n < 10^k\}$$

for each $k \in \mathbb{N}$. This additional "condition" will enable to reduce determination of the orbits of the so called Kaprekar's transformations $T_k\colon X(k) \to X(k)$ – described in the next section – to solution of some diophantine equations. Although we have learnt about orbits of many maps $T_k$, this knowledge did not help us unfortunately to answer the basic question: how many orbits do these maps possess in dependence on the value of parameter $k$ for any $k \in \mathbb{N}$? In both parts of our study we are able to answer this question only for values $k \leq 20$.

In Part II of our considerations we will present many various remarks, facts and conjectures which arose basically by observing the numerical results concerning the description of the orbits of maps $T_n$ for $n \leq 20$. We will prove, among others, that the fixed points of these maps generate the infinite sequences of the fixed points of maps $T_{a\,n+b}$, $n \in \mathbb{N}$, for some natural numbers $a$ and $b$.

Additionally, we have noticed that many from among the maps investigated by us (including the generalizations of the Kaprekar's transformation – we define them in last section – however, with regard to this paper length, we will present the appropriate considerations in a separate paper) preserve the strong Sharkovsky's order (the Sharkovsky's order, respectively). It should be understood in the following way.

**Definition 1.** *Map $T\colon X \to X$, where $X$ is a finite set, preserves the strong Sharkovsky's order if the elements of the set of cardinalities of all orbits of this map can be ordered* *in the sequence $k_1, k_2, \dots, k_n$, being the sequence of natural numbers, successive in the sense of order (1).*

**Definition 2.** *Map $T\colon X \to X$, where $X$ is a finite set, preserves the Sharkovsky's order if the elements of the set of cardinalities of all orbits of this map can be ordered in the sequences $k_1^{(r)}, k_2^{(r)}, \dots, k_n^{(r)}$, $r = 1, 2, \dots, s$, successive in the sense of order (1), and the different values of superscript $r$ correspond with the different "numbers of levels" of description (1). More precisely, the first level of description (1) is formed by the numbers*

$$3, 5, 7, 9, 11, \dots,$$

*the second level of description (1) is made by the numbers*

$$2 \cdot 3, 2 \cdot 5, 2 \cdot 7, 2 \cdot 9, 2 \cdot 11, \dots,$$

*the third level of description (1) is created by the numbers*

$$4 \cdot 3, 4 \cdot 5, 4 \cdot 7, 4 \cdot 9, 4 \cdot 11, \dots, \quad \textit{and so on},$$

*and finally "the last level" of description (1) is formed by the numbers*

$$\dots, 2^5, 2^4, 2^3, 2^2, 2, 1.$$

Reason of these definitions is also worth to recall. So, as it is easy to prove, for any one-to-one sequence $k_1, k_2, \dots, k_n$ of natural numbers there exist the sets $X_i$, $i = 1, 2, \dots, n$, (pairwise disjoint and such that $cardX_i = k_i$) and the map $T\colon \bigcup_{i=1}^{n} X_i \to \bigcup_{i=1}^{n} X_i$, for which the sets $X_i$ are the only minimal orbits.

Moreover, we have investigated the minimal cycles of the discussed here maps with regard to the Erdős-Szekeres theorem, as well as to the maximal length of monotonic intervals of the given cycle (see [30]) and, at last, by paying the special attention to the relatively new but extremely dynamic theory of "pattern avoiding permutations" (see [6], [20]).

Let us recall here at least few essential definitions and facts. Let $\mathbf{a} = \{a_i\}_{i=1}^{n}$ be a one-to-one sequence of real numbers. Each subsequence $\mathbf{b}$ of $\mathbf{a}$ having the form $\{a_l, a_{l+1}, ..., a_{l+r}\}$ for some $l, r \in \mathbb{N}_0$, $1 \le l \le l{+}r \le n$, will be called an interval of $\mathbf{a}$. A subsequence $\mathbf{b}$ of $\mathbf{a}$ is said to be a monotonic interval of $\mathbf{a}$ whenever $\mathbf{b}$ is an interval of $\mathbf{a}$ and, simultaneously, $\mathbf{a}$ is a monotonic sequence. Moreover, we will denote by $l(\mathbf{a}) := n$ the number of elements of $\mathbf{a}$ called as the length of $\mathbf{a}$, by $d(\mathbf{a})$ – the maximal number from among the numbers denoting the lengths of all decreasing subsequences of $\mathbf{a}$ and finally by $i(\mathbf{a})$ – the maximal number from among the numbers denoting the lengths of all increasing subsequences of $\mathbf{a}$.

**Erdős-Szekeres' theorem.** *Let us suppose that $\mathbf{a}$ is a finite one-to-one sequence of real numbers. Then we have*

$$d(\mathbf{a})\, i(\mathbf{a}) \ge l(\mathbf{a}).$$

The above theorem comes from the joint paper by Erdős and Szekeres concerning the Ramseys problem [12]. Next, Wituła et al. in [30] have discussed whether the given one-to-one sequence $\mathbf{a}$ of all numbers $1, 2, ..., n$ (which means that $\mathbf{a}$ can be identified with the respective permutation on set $\{1, 2, ..., n\}$) contains a monotonic interval $\mathbf{b}$ of length 3. The following fact is, among others, proven there.

**Theorem 1.** *Let $\mathbf{a} = \{a_i\}_{i=1}^{3n}$ be a permutation on $\{1, 2, ..., 3n\}$ and let $n \ge 4$. If $i(\mathbf{a}) = n$, $d(\mathbf{a}) = 3$, $a_k = 3n$ and $a_l = 1$ for some $k < l$, then $\mathbf{a}$ contains a monotonic interval $\mathbf{b}$ of length 3.*

In the next section of this paper we will present the definition of Kaprekar's transformations $T_n$ and we will formulate the conditions describing the elements of minimal orbits of $T_n$ for $4 \le n \le 7$. In fact, it will be only the necessary conditions, yet they will "reduce" enough the sets of natural numbers containing the maxinvariant subset of the respective Kaprekar's transformation, so that the final calculations will be possible to make even by hand.

## II. Kaprekar's transformations

In this section we discuss the Kaprekar's transformations

$$\begin{aligned} T_n\colon \{0\} \cup \{\alpha\colon 10^{n-1} - 1 \le \alpha < 10^n\} \to \\ \to \{0\} \cup \{\alpha\colon 10^{n-1} - 1 \le \alpha < 10^n\} \end{aligned}$$

for every $n \in \mathbb{N}$, defined in the following way. We set $T_n(0) = 0$ and let $\alpha \in \mathbb{N}$ be any $n$-digit number, the decimal expansion of which is composed of digits $0 \le a_1 \le a_2 \le \ldots \le a_n \le 9$. We take

$$\begin{aligned} T_n(\alpha) &:= \sum_{k=1}^{n} (a_k - a_{n-k+1}) 10^{k-1} = \\ &= a_n a_{n-1} \ldots a_1 - a_1 a_2 \ldots a_n. \end{aligned}$$

The orbits of operator $T_n$ will be called as the $T_n$-orbits for every $n \in \mathbb{N}$. Moreover, we will call the $k$-fold composition of operator $T_n$, for any $k, n \in \mathbb{N}$, as the $T_n$-composition. Next, the fixed points of operator $T_n$, where $n \in \mathbb{N}$, will be called as the Kaprekar's constants of $n$-th order.

Let us note that Hindu mathematician Dattathreya Ramachandra Kaprekar has started in 1949 in paper [16] the discussion on the, called now, Kaprekar's transformations $T_n$. The classical Kaprekar's constant, that is number 6174, was also announced in this paper. But only in paper [17] Kaprekar proved that after applying operator $T_4$ at most 7-times every four-digit number in base 10 leads to the same result, that is $6174 = T_4(6174)$.

Properties of operator $T_5$, acting on the five-digit integers in bases $r < 13$, were investigated by Charles W. Trigg [29], the mathematician well-known mostly for his great involvement in the issues of recreational mathematics. Next, Klaus E. Eldridge and Seok Sagong in their paper [11] from 1988 described the convergence of $\{T_3^n(x)\}_{n=1}^{\infty}$ for all three-digit numbers $x$ for any base $r \in \mathbb{N}$, $r \ge 2$. They obtained, among others, the following result.

**Theorem 2.**

a) *$T_3^n(x)$ is convergent (in usual sense) to nontrivial constant (also called the Kaprekar's constant) if and only if*

*$r$ is even. The respective Kaprekar's constant is equal to the 3-digit number $\left(\frac{r-2}{2}, r-1, \frac{r}{2}\right)$ in base $r$.*

b) *If $r$ is odd then $T_3$ possesses (except the trivial orbit) only one two-element orbit consisting of numbers $\left(\frac{r-3}{2}, r-1, \frac{r+1}{2}\right)$ and $\left(\frac{r-1}{2}, r-1, \frac{r-1}{2}\right)$ in base $r$.*

Papers [7], [15], [18] are also devoted to the discussion on Kaprekar's transformations.

We will present now the descriptions of elements of orbits of maps $T_n$ for values of $n$ equal in turn 5,6,7 and 4. These facts are partly new and originally presented.

**Theorem 3.** *Every orbit of operator $T_5$ must contain exclusively the numbers of the form $ABA \times 99$, where $0 \leq B \leq A \leq 9$.*

*Proof:* For five-digit number $n$ composed only of digits $0 \leq e \leq d \leq c \leq b \leq a \leq 9$ we have

$$\begin{aligned} T_5(n) &= (a-e)(10^4-1)+(b-d)(10^3-10) = \\ &= 99 \times ((a-e) \times 101 + (b-d) \times 10) \\ &= 99 \times ABA, \end{aligned}$$

where $0 \leq B := b-d \leq A := a-e \leq 9$. ■

**Corollary 1.** *The orbits of operator $T_5$ can be sought just and only from among the $T_5$-iteration of the following 54 numbers*

$$\begin{aligned} &101 \times 99,\ 111 \times 99,\ 202 \times 99,\ 212 \times 99,\ 222 \times 99, \\ &303 \times 99,\ 313 \times 99,\ 323 \times 99,\ 333 \times 99, \\ &\vdots \\ &909 \times 99,\ 919 \times 99,\ 929 \times 99, \ldots, 999 \times 99. \end{aligned}$$

*Moreover, each $T_5$-orbit must be the subset of the above set of numbers.*

**Remark 1.** *We have $T_5(99 \times 111) = T_5(99 \times 999)$.*

**Theorem 4.** *Each orbit of operator $T_7$ must contain only the numbers of the form*

$$AB(A+C)BA \times 99, \tag{2}$$

*where $0 \leq C \leq B \leq A \leq (A+C) \leq 9$, or*

$$A(B+1)(A+C-10)BA \times 99, \tag{3}$$

*where $1 \leq C \leq B \leq A \leq 9 < (A+C)$ and $B \leq 8$.*

*Proof:* Let $n$ be the seven-digit number composed of the following seven digits

$$0 \leq g \leq f \leq e \leq d \leq c \leq b \leq a \leq 9.$$

Then we have

$$\begin{aligned} &T_7(n) = (a-g)(10^6-1)+(b-f)(10^5-10)+ \\ &+(c-e)(10^4-10^2) = 99 \times ((a-g) \times 10101+ \\ &+(b-f) \times 1010 + (c-e) \times 100) = \\ &= \begin{cases} AB(A+C)BA \times 99, & \text{if } A+C \leq 9, \\ A(B+1)(A+C-10)BA \times 99, & \text{if } A+C > 9 \\ & \text{and } B \leq 8, \end{cases} \end{aligned}$$

where $A := a-g$, $B := b-f$, $C := c-e$. It is obvious that we have $0 \leq C \leq B \leq A \leq 9$. ■

**Corollary 2.** *Orbits of operator $T_7$ can be sought only from among the $T_7$-compositions on the following numbers (we give first the numbers defined by formula (2)):*

$$\begin{aligned} &10101 \times 99,\ 11111 \times 99,\ 11211 \times 99, \\ &20202 \times 99,\ 21212 \times 99,\ 21312 \times 99, \\ &22222 \times 99,\ 22322 \times 99,\ 22422 \times 99, \\ &30303 \times 99, \ldots, \\ &\vdots \\ &90909 \times 99, \ldots, 98989 \times 99,\ 999999 \times 99, \end{aligned}$$

*and (it is about 163 numbers described by formula (3)):*

$$\begin{aligned} &99089 \times 99,\ 99189 \times 99, \ldots, 99889 \times 99, \\ &98079 \times 99,\ 98179 \times 99, \ldots, 98679 \times 99, \\ &\vdots \\ &93029 \times 99,\ 93129 \times 99, \\ &92019 \times 99, \\ &89088 \times 99,\ 89188 \times 99, \ldots, 89688 \times 99, \\ &88078 \times 99,\ 88178 \times 99, \ldots, 88578 \times 99, \\ &\vdots \\ &84038 \times 99,\ 84138 \times 99, \\ &83028 \times 99, \\ &\vdots \\ &67066 \times 99,\ 67166 \times 99,\ 67266 \times 99, \\ &66056 \times 99,\ 66156 \times 99, \\ &65046 \times 99, \\ &56055 \times 99. \end{aligned}$$

**Theorem 5.** *Each orbit of operator $T_6$ contains only the numbers described by the following seven formulae*

$$9 \times A(A+B)(A+B+C)(A+B)A, \tag{4}$$

*where $0 \leq C \leq B \leq A \leq A+B+C \leq 9$, or*

$$9 \times A(A+B+1)(A+B+C-10)(A+B)A, \tag{5}$$

*where $0 \leq C \leq B \leq A \leq A+B \leq 8$ and $10 \leq A+B+C < 20$, or*

$$9 \times (A+1)0(A+B+C-10)9A, \tag{6}$$

*where $1 \leq C \leq B \leq A \leq 9$ and $A+B=9$, or*

$$9 \times (A+1)(A+B-9)(A+B+C-9)(A+B-10)A, \tag{7}$$

*where $0 \leq C \leq B \leq A \leq 9$ and $10 \leq A+B \leq A+B+C \leq 18$, or*

$$9 \times (A+1)(A+B-8)(A+B+C-19)(A+B-10)A, \tag{8}$$

*where* $0 \le C \le B \le A \le 9$ *and* $A+B+C \ge 19$ *(we note that then* $A+B \ge 10$*) and* $A+B \le 17$*, or*

$$9 \times 110(C-1)89, \tag{9}$$

*where* $C \ge 1$*, or*

$$9 \times 109989. \tag{10}$$

*Proof:* In order to get the presented formulae let us assume that $n$ is the natural six-digits number composed of the digits $0 \le a_6 \le a_5 \ldots \le a_1 \le 9$. Then we obtain

$$T_6(n) = 9((a_1-a_6)(10^4+10^3+10^2+10+1)+ \\ +(a_2-a_5)(10^3+10^2+10)+(a_3-a_4)10^2).$$

By taking $A := a_1 - a_6$, $B := a_2 - a_5$, $C := a_3 - a_4$ we find

$$T_6(n) = A10^4+(A+B)10^3+(A+B+C)10^2+(A+B)10+A,$$

where $0 \le C \le B \le A \le 9$. The only thing which left is to analyze the value of sums $A+B+C$ and $A+B$ which gives the thesis of theorem. ■

**Remark 2.** *Although we have as many as seven different formulae describing potential numbers belonging to the orbits of operator* $T_6$*, their description can be directly generated in easy way. However, we will omit here this description.*

**Remark 3.** *It was numerically proved by the Authors that operator* $T_6$ *possesses three fixed points (the Kaprekar's constants of sixth order):*

$$0,\ 549945,\ 631764$$

*and one 7-element orbit (we give it in a table in Part II of this paper). The information on an existing 7-element orbit is omitted in the table presented in the Polish version of Wikipedia (http://pl.wikipedia.org/wiki/Stala_Kaprekara).*

**Theorem 6.** *Orbits of operator* $T_4$ *contain only the numbers described by formulae*

$$9 \times A(A+B)A, \tag{11}$$

*where* $0 \le B \le A \le A+B \le 9$*, or*

$$9 \times (A+1)(A+B-10)A, \tag{12}$$

*where* $1 \le B \le A \le 9$ *and* $A+B \ge 10$*.*

*Proof:* Let $n \in \mathbb{N}$ be the four-digit number composed of digits $0 \le d \le c \le b \le a \le 9$. Then we have

$$\begin{aligned} T_4(n) &= (a-d)(10^3-1)+(b-c)(10^2-10) = \\ &= 9 \times \big((a-d)(10^2+10+1)+(b-c)10\big) = \\ &= \begin{cases} 9 \times A(A+B)A, & \text{if } A+B \le 9, \\ 9 \times (A+1)(A+B-10)A, & \text{if } A+B > 9, \end{cases} \end{aligned}$$

where $A := a-d$, $B := b-c$. Certainly we have $0 \le B \le A \le 9$. ■

**Remark 4.** *Formulae (11) and (12) describe the following 45 numbers*

$$\begin{array}{l} 111 \times 9,\ 121 \times 9, \\ 222 \times 9,\ 232 \times 9,\ 242 \times 9, \\ 333 \times 9,\ 343 \times 9,\ 353 \times 9,\ 363 \times 9, \\ 444 \times 9,\ 454 \times 9, \ldots, 484 \times 9, \\ 555 \times 9,\ 565 \times 9, \ldots, 595 \times 9, \\ 605 \times 9, \\ 666 \times 9,\ 676 \times 9,\ 686 \times 9^*,\ 696 \times 9, \\ 706 \times 9,\ 716 \times 9,\ 726 \times 9, \\ 777 \times 9,\ 787 \times 9,\ 797 \times 9, \\ 807 \times 9,\ 817 \times 9, \ldots, 847 \times 9, \\ 888 \times 9,\ 898 \times 9,\ 908 \times 9, \ldots, 968 \times 9,\ 999 \times 9. \end{array}$$

*where by* $^*$ *we have distinguished the Kaprekar's constant. Directly calculating (even by hand – if we are extremely dogged) we can verify that* $T_4$ *possesses only one orbit*

$$\{686 \times 9 = 6174\}.$$

*Let us recall, that this fixed point of* $T_4$*, i.e. number 6174, is called the Kaprekar's constant (of fourth order).*

## III. Final remarks

Authors of this paper, apart from the discussed here Kaprekar's transformations, have also defined and investigated the minimal orbits (cycles, respectively) of few generalizations of these transformations, like for example

— the symmetric Kaprekar's transformation

Let $a_1a_2\ldots a_n$ be the decimal expansion of number $a \in \mathbb{N}$, $10^{n-1} \le a < 10^n$. Then the $n$-th symmetric Kaprekar's transformation $M$ is defined as

$$M(a_1a_2\ldots a_n) = \sum_{k=1}^{n} |c_k - b_k| 10^{k-1}$$

where $(b_1, b_2, \ldots, b_n)$ and $(c_1, c_2, \ldots, c_n)$ are the sequences, nondecreasing and nonincreasing, respectively, composed of the digits $a_1, a_2, ..., a_n$. We include to the set of $n$-digit numbers also the number zero. Orbits of operators $M$ for the odd values $n \le 19$, although "quite easy" to calculate even by hand, surprise yet with their final form. We will present here only few quantitative pieces of information.

So, if $n = 2k+1$, $1 \le k \le 5$, then $M$ possesses only the fixed points and $k$-element orbits, for $n = 13$ operator $M$ possesses two fixed points, 0 and 65432101...6, four 2-element cycles, eleven 3-element cycles and 827 cycles of length 6 (sic). For $n = 15$ the operator $M$ possesses 44 fixed points, 342 different 2-elements orbits and 2678 different 4-elements orbits. For $n = 17$ the operator $M$ possesses only 6 fixed points, 32 different 2-element orbits and 6060 different 4-element orbits. Finally, for $n = 2^k$ the operator $M$ possesses only trivial orbit $= \{0\}$ for every $k \in \mathbb{N}$.

— nonoptimal Kaprekar's transformations
One of the examples of this transformation, called by us the $Q$-Kaprekar's transformation, is defined as

$$Q_n(A) := (a_n - a_2)10^{n-1} + (a_{n-1} - a_1)10^{n-2} + \\ + \sum_{k=1}^{n-2}(a_k - a_{n-k+1})10^{k-1},$$

where $0 \leq a_1 \leq a_2 \leq \ldots \leq a_n \leq 9$ are the all digits of decimal expansion of number $A$. We note that, in contrast to the Kaprekar's transformation $T_4$, the transformation $Q_4$ possesses two 2-element orbits: $\{2187, 6543\}$ and $\{3285, 5274\}$ and the trivial fixed point. Next, $Q_5$ possesses the trivial fixed point and the 2-element orbit $\{52974, 54963\}$ (in contrast, transformation $T_5$ has four different orbits). Transformations $Q_6$ and $T_6$ have both three fixed points and, respectively, the 8-element orbit and the 7-element orbit. Transformations $Q_7$ and $T_7$ possess both the trivial fixed point and one 8-element orbit (but of different orbit types).

— general Kaprekar's transformations
We take that the natural number $A$, $10^{n-1} \leq A < 10^n$, possesses the following decimal expansion $A = d_1 d_2 \ldots d_n$. Let $a_1 := \max\{d_1, d_2, \ldots, d_n\}$, $a_2 := \max\{d_2, d_3, \ldots, d_n\}$ and in general $a_k := \max\{d_k, d_{k+1}, \ldots, d_n\}$, for $k = 1, 2, \ldots, n$. The announced general Kaprekar's transformations are defined by relations

$$d_{\sigma,\pi}(A) := \sum_{k=1}^{n} |d_{\sigma(k)} - d_{\pi(k)}| 10^{n-k},$$

$$d_{\sigma,\pi}^{weak}(A) := \left| \sum_{k=1}^{n} (d_{\sigma(k)} - d_{\pi(k)}) 10^{n-k} \right|,$$

and

$$D_{f,g}(A) := \sum_{k=1}^{n} |d_{f(k)} - d_{g(k)}| 10^{n-k},$$

$$D_{f,g}^{weak}(A) := \left| \sum_{k=1}^{n} (d_{f(k)} - d_{g(k)}) 10^{n-k} \right|,$$

$$R_f(A) := \sum_{k=1}^{n} |a_k - a_{f(k)}| 10^{n-k},$$

$$R_f^{weak}(A) := \left| \sum_{k=1}^{n} (a_k - a_{f(k)}) 10^{n-k} \right|,$$

for any permutations $\sigma$, $\pi$ on set $\{1, 2, \ldots, n\}$ and for any functions $f, g\colon \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$.

## References

[1] L. Alseda, J. Llibre, M. Misiurewicz, *Combinatorial Dynamics and Entropy in Dimension One*, World Scientific Publ. Co.; 2000 (second ed.), http://dx.doi.org/10.1142/4205

[2] L. Alseda, J. Llibre and M. Misiurewicz, "Periodic orbits of maps of $Y$", *Trans. Amer. Math. Soc.*, vol. 313 No 2, 1989, pp. 475–538, http://dx.doi.org/10.2307/2001417

[3] S. Baldwin, "Generalizations of a theorem of Sarkovskii on orbits of continuous maps of the real line", *Discrete Math.*, vol. 67, 1987, pp. 111–127.

[4] L. Block, E.M. Coven, L. Jonker and M. Misiurewicz, "Primary cycles on the circle", *Trans. Amer. Math. Soc.*, vol. 311 No 1, 1989, pp. 323–335, http://dx.doi.org/10.1090/S0002-9947-1989-0974779-4

[5] A. Blokh and M. Misiurewicz, "New order for periodic orbits of interval maps", *Ergodic Th. & Dynam. Soc.*, vol. 17, 1997, pp. 565–574, http://dx.doi.org/10.1017/S0143385797084927

[6] M. Bóna, *A Walk Through Combinatorics*, World Scientific, Singapore; 2011, http://dx.doi.org/10.1142/8027

[7] M. Brown and J. Slifker, "A periodic sequence – problem 6439", *Amer. Math. Monthly*, vol. 92 No 3, 1985, p. 218, http://dx.doi.org/10.2307/2687867

[8] K. Ciesielski and Z. Pogoda, "On ordering the natural numbers, or, the Sharkovski Theorem", *Amer. Math. Monthly*, vol. 115 No 2, 2008, pp. 159–165.

[9] B.S. Du, "A simple proof of Sharkovsky's theorem", *Amer. Math. Monthly,* vol. 111, 2004, pp. 595–599, http://dx.doi.org/10.2307/4145161

[10] B.S. Du, "A simple proof of Sharkovsky's theorem revisited", *Amer. Math. Monthly,* vol. 114, 2007, pp. 152–155, http://dx.doi.org/10.2307/27642145

[11] K.E. Eldridge and S. Sagong, "The determination of Kaprekar convergence and loop convergence of all three-digit numbers", *Amer. Math. Monthly,* vol. 95 No 2, 1988, pp. 105–112, http://dx.doi.org/10.2307/2323062

[12] P. Erdős, G. Szekeres, "A combinatorial problem in geometry", *Compositio Math.,* 1935, pp. 463–470.

[13] R.K. Guy, "Conway's RATS and other reversals", *Amer. Math. Monthly,* vol. 96 No 5, 1989, pp. 425–428, http://dx.doi.org/10.2307/2325149

[14] H. Hasse and G.D. Prichett, "The determination of all four-digit Kaprekars constants", *J. Reine Angew. Math.,* vol. 299/300, 1978, pp. 113–124.

[15] J.H. Jordan, "Self producing sequences of digits", *Amer. Math. Monthly,* vol. 71 No 1, 1964, pp. 61–64, http://dx.doi.org/10.2307/2311308

[16] D.R. Kaprekar, "Another solitarie game", *Scripta Mathematica,* vol. 15, 1949, pp. 244–245.

[17] D.R. Kaprekar, "An interesting property of the number 6174", *Scripta Mathematica,* vol. 21, 1955, p. 304.

[18] R.M. Krause, N. Miller and C.W. Trigg, "Kaprekar's constant", *Amer. Math. Monthly,* vol. 78 No 2, 1971, pp. 197–198, http://dx.doi.org/10.2307/2317638

[19] E. Lundberg, "Almost all orbit types imply period–3", *Topology Appl.,* vol. 154, 2007, pp. 2741–2744, http://dx.doi.org/10.1016/j.topol.2007.05.009

[20] T. Mansour, *Combinatorics of Set Partitions*, CRC Press, Boca Raton; 2013.

[21] J. Mihelić, L. Furst and U. Cibej, "Exploratory equivalence in graphs: Definition and algorithms", *Proc. FedCSIS,* ACSIS, vol. 2, 2014, pp. 447ij456, http://dx.doi.org/10.15439/2014F352

[22] P. Minc and W.R.R. Transue, "Sarkovskii's theorem for hereditarily decomposable chainable continua", *Trans. Amer. Math. Soc.,* vol. 315, 1989, pp. 173–188, http://dx.doi.org/10.2307/2001378

[23] M. Misiurewicz, "Remarks on Sharkovsky's Theorem", *Amer. Math. Monthly,* vol. 104 No 9, 1997, pp. 846–847, http://dx.doi.org/10.2307/2975290

[24] J. Mulvey, "A geometric algorithm to decide the forcing relation on cycles", *Real Analysis Exchange,* vol. 23 No 2, 1997–1998, pp. 709–717.

[25] D.J. Ryden, "The Sarkovskii order for periodic continua", *Topology Appl.,* vol. 154 No 11, 2007, pp. 2253–2264, http://dx.doi.org/10.1016/j.topol.2007.03.001

[26] D.J. Ryden, "The Sarkovskii order for periodic continua II", *Topology Appl.,* vol. 155 No 2, 2007, pp. 92–104, http://dx.doi.org/10.1016/j.topol.2007.08.009

[27] A.N. Sharkovskii, "Coexistence of cycles of a continuous map of the line into itself", *Ukrain. Math. J.,* vol. 16, 1964, pp. 61–71 (in Russian).

[28] A.N. Sharkovskii, "Coexistence of cycles of a continuous map of the line into itself", *Internat. J. Bifurcation Chaos,* vol. 5 No 5, 1995, pp. 1263–1273 (English translation of [27]), http://dx.doi.org/10.1142/S0218127495000934

[29] C.W. Trigg, "Kaprekar's routine with five-digit integers", *Math. Magazine,* vol. 45 No 3, 1972, pp. 121–129, http://dx.doi.org/10.2307/2687867

[30] R. Wituła, D. Słota and R. Seweryn, "On Erdős theorem for monotonic subsequences", *Demonstratio Math.*, vol. 40 No 2, 2007, pp. 239–259.

DOI: 10.15439/2015F208
ACSIS, Vol. 5

# Trigger Based on the Artificial Neural Network implemented in the Cyclone V FPGA for a Detection of Neutrino-Origin Showers in the Pierre Auger surface detector

Zbigniew Szadkowski
Dariusz Głas
University of Łódź
Department of Physics and Applied Informatics
Faculty of High-Energy Astrophysics
E-mail : zszadkow@kfd2.phys.uni.lodz.pl
E-mail : dglas@uni.lodz.pl

Krzysztof Pytel
University of Łódź
Department of Physics and Applied Informatics
90-236 Łódź, Pomorska 149
Faculty of Informatics
Email : kpytel@uni.lodz.pl

***Abstract*—Observations of ultra-high energy neutrinos became a priority in experimental astroparticle physics. Up to now, the Pierre Auger Observatory did not find any candidate on a neutrino event. This imposes competitive limits to the diffuse flux of ultra-high energy neutrinos in the EeV range and above.**

**The prototype Front-End boards for Auger-Beyond-2015 with Cyclone® V E can test the neural network algorithm in real pampas conditions in 2015. Showers for muon and tau neutrino initiating particles on various altitudes, angles and energies were simulated in CORSIKA and OffLine platforms giving pattern of ADC traces in Auger water Cherenkov detectors. The 3-layer 12-10-1 neural network was taught in MATLAB by simulated ADC traces according the Levenberg - Marquardt algorithm. New sophisticated trigger implemented in Cyclone® V E FPGAs with large amount of DSP blocks, embedded memory running with 120 - 160 MHz sampling may support to discover neutrino events in the Pierre Auger Observatory.**

## I. Introduction

OBSERVATION of ultrahigh energy cosmic rays (UHECR) of energy 1−100 EeV ($10^{18}$ − $10^{20}$ eV) has stimulated much experimental as well as theoretical activity in the field of astro-particle Physics [1]. Although many mysteries remain to be solved, such as the origin of the UHECRs, their production mechanism and composition, we know that it is very difficult to produce these energetic particles without associated fluxes of ultrahigh energy neutrinos (UHE$\nu$s) [2]. In the "bottom-up" scenarios, protons and nuclei are accelerated in astrophysical shocks, while pions are produced by cosmic ray interactions with matter or radiation at the source [3]. In the "top-down" models, protons and neutrons are produced from quark and gluon fragmentation with a production of much more pions than nucleons [4]. Furthermore, protons and nuclei also produce pions due to the Greisen-Zatsepin-Kuzmin (GZK) cutoff [5]. The ultrahigh energy cosmic rays (UHECR) flux above $\sim 5\times 10^{19}$ eV is significantly suppressed according to expectations based on the UHECRs interaction of with the cosmic microwave background (CMB) radiation [6]. For primary protons, the photo-pion production is responsible for the GZK effect, thus UHE$\nu$s are produced from decayed charged pions. However, their fluxes are doubtful [4] and if the primaries are heavy nuclei, the UHE$\nu$s should be strongly suppressed [7].

Neutrinos can show directly sources of their production due to no deflection by magnetic fields. Unlike photons they travel inviolate from the sources carrying an impression of the production model. UHE$\nu$s can be detected with arrays of detectors at ground level that are currently being used to measure extensive showers produced by cosmic rays [8]. The main challenge is an extraction from the background, induced by regular cosmic rays, showers initiated by neutrinos. Due to a very small neutrino cross-section for interactions, higher probability of a detection is at high zenith angles [9] due to a bigger atmosphere slant depth provides thicker target for neutrino interactions. Inclined showers starting a development deeply in the atmosphere can be a signature of neutrino events.

## II. Triggers

Each water Cherenkov detector of the surface array has a 10 $m^2$ water surface area and 1.2 m water depth, with three 9-inch photomultiplier tubes (PMTs) looking through optical coupling material into the water volume, which is contained in a Tyvek reflective liner. Each PMT provides signals, which are digitized by 40 MHz 10-bit Analog to Digital Converters (ADCs).

Triggers are generated for signals above amplitude thresholds. The trigger for the surface detector array is hierarchical. Two levels of trigger (called T1 and T2) are formed at each detector. T2 triggers are combined with those from other detectors and examined for spatial and temporal correlations, leading to an array trigger (T3). The T3 trigger initiates data acquisition and storage. Two independent trigger modes are

978-83-60810-66-8/$25.00©2015, IEEE

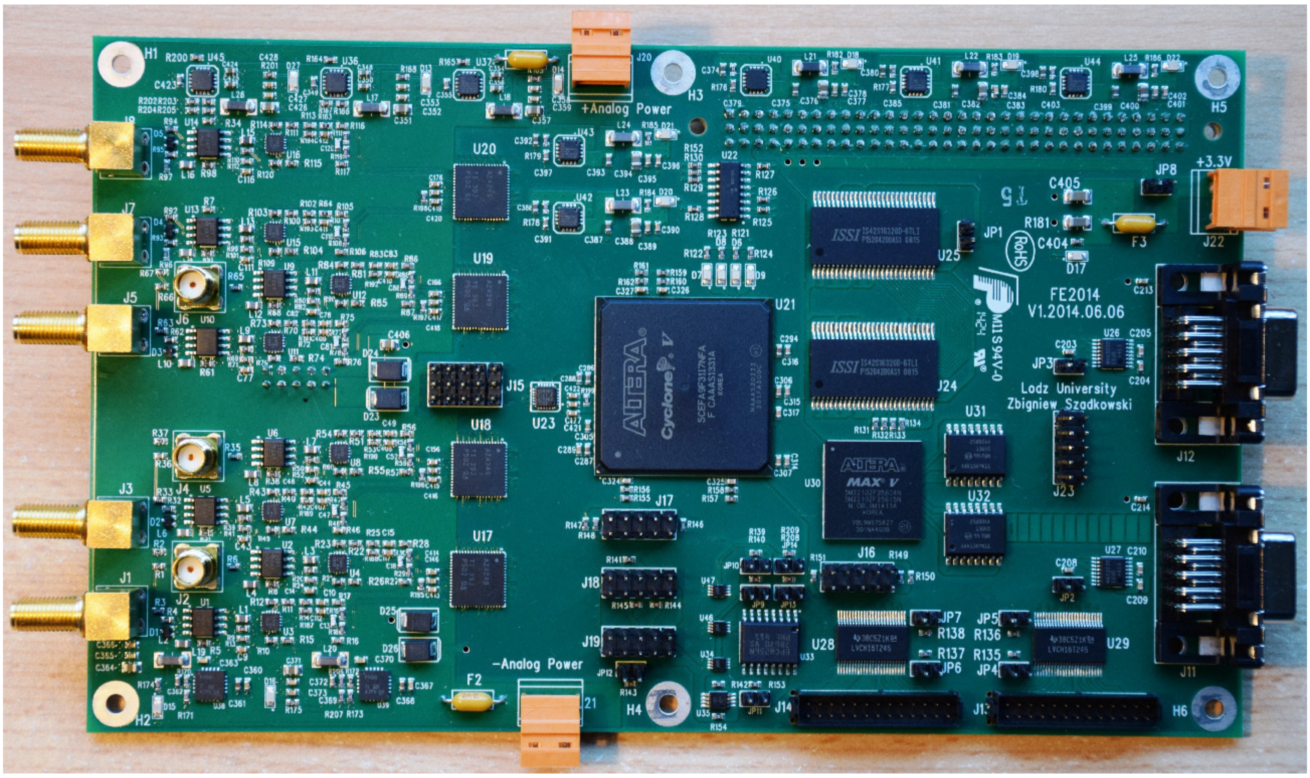


Fig. 1. The Front-End Board developed for the Auger-Beyond-2015 upgrade design used also for the ANN algorithm tests.

implemented as T1. The 1st T1 mode is a simple threshold trigger (TH) which requires the coincidence of the three PMTs each above 1.75 $I_{VEM}^{peak}$. The Vertical Equivalent Muon (VEM) is a unit used for a calibration corresponding to ∼50 ADC-units for 40 MHz sampling frequency. This trigger is used to select large signals that are rather concentrated in time. The 2nd mode is designated the "Time-over-Threshold" trigger (ToT) and at least 13 time-bins in 120 ADC bins of a sliding window of 3 $\mu$s are required to be above a threshold of 0.2 $I_{VEM}^{peak}$ in coincidence in 2 out of 3 PMTs. This trigger is intended to select sequences of small signals spread in time. From the point of view of short pulses characteristic for very inclined "old" showers as well as for relatively short signals for inclined "young showers the ToT trigger is useless.

FPGAs currently used in the Pierre Auger surface detectors analyze signal amplitudes in a time domain. Much higher efficiency can provide spectral triggers based i.e. on Discrete Cosine Transform [10] [11] [12]. Triggers based on an artificial neuron networks is an alternative approach focusing on specific patterns of ADC traces registered in Pierre Auger water Cherenkov detectors [13].

## III. ADC TRACES ANALYSIS

With the SD of the Pierre Auger Observatory we can detect and identify UHE neutrinos in the EeV range and above [14]. Due to much larger cross-section than neutrinos, the 1st interaction for protons, heavier nuclei and even photons usually appears shortly after entering the atmosphere. However, neutrinos can generate showers initiated deeply into the atmosphere. Vertical showers initiated by protons or heavy nuclei have a considerable amount of electromagnetic component at the ground ("young" shower front). However, at high zenith angles ($\theta \geq 75$) (thicker than about three vertical atmospheres), UHECRs interacting high in the atmosphere generate shower fronts dominated by muons at ground ("old" shower front), which generate narrow signals (short ADC traces) spreading over typically tens of nano-seconds in practically all the stations of the event. These traces can be recognized with 16-point DCT algorithm as well as with 16-point input AAN.

Calculations show [15] that "young" showers are spread in time over hundreds of nano-seconds. For the "old" showers practically only the muonic components survives. It gives a short bump in the SD. The "young" showers contains also some electromagnetic component, which enlarge in time an ADC traces. However, muonic components of "young" showers is ahead of the electromagnetic one and gives an early bump. The rising edge of the bump is not so sharp as for the "old" ones, but the signal next is also relatively fast attenuated, till the electromagnetic component starts to give its own contribution. The ANN approach can focus on the early bump, to select traces potentially generated by neutrinos.

On the other hand, independent simulations of showers in CORSIKA [16] an next calculation of water Cherenkov detectors (WCDs) response in OffLine package [17] showed

TABLE I
DISTANCES FROM THE PLACE OF THE FIRST INTERACTION TO DETECTOR FOR PROTONS AND MUON NEUTRINOS IN DEPENDENCE OF ZENITH ANGLE. ALL THE DISTANCES ARE IN $g/cm^2$. BECAUSE OF THE GEOMETRY NOT ALL THE DISTANCES ARE AVAILABLE FROM EVERY ANGLE.

| | | 500 | 1000 | 2000 | 3000 | 4000 | 5000 | 10000 |
|---|---|---|---|---|---|---|---|---|
| 80° | p | | | + | + | + | | |
| 80° | $\nu$ | + | + | + | + | + | | |
| 85° | p | | | + | + | + | + | |
| 85° | $\nu$ | + | + | + | + | + | + | |
| 89° | p | | | + | + | + | + | + |
| 89° | $\nu$ | + | + | + | + | + | + | + |

that for neutrino showers (initiated either by $\nu_\mu$ or $\nu_\tau$) for relatively big zenith angle (i.e. 70°) and low altitude (9 km) (to be treated as "young" showers before a maximum of development) give relatively short ADC traces and they can be analyzed also by 16-point pattern engines [13].

## IV. CORSIKA AND OFFLINE CALCULATIONS

The main motivation of an ANN implementation as a shower trigger is a fact that up to now the entire array did not registered any neutrino-induced event. The probably reasons are: a) a very low stream of neutrinos and b) amplitudes of ADC-traces are small and probably are below of threshold of standard 3-fold coincidence trigger. The main idea was to use ANN approach as a pattern recognition technique.

The input data for the ANN are simulated traces of protons and muon neutrinos, which hit the atmosphere in high zenith angles - 80°, 85° and 89°, respectively. The chosen energies of primary particles are $3 \times 10^8$, $10^9$, $3 \times 10^9$ and $10^{10}$ GeV, respectively. The distances from the place of the first interaction to the detector used for simulations are dependent of the angle and the type of particle (Table I). We decided not to simulate protons which are very close to detector, because the probability that proton will not interact on long way to detector is very low. Additionally, traces produced by this kind of interactions may include also electromagnetic part of the shower. These traces would look completely different than the rest and may significantly decrease the efficiency of the ANN.

There are 120 different categories. These categories are used as input by the CORSIKA simulation platform. The CORSIKA program simulates the cosmic ray shower initiated by the specific particle. The result of this simulation is distribution of the position and energies of the particles on the level of the detector. Simulations are relatively fast. All of the 120 categories had been simulated in a week. The simulated cosmic ray showers are the input for the OffLine package, which provides a response of the water Cherenkov detector and generates the ADC traces (signal waveforms). These simulations are very time consuming. As a result we got simulated traces from the photo-multipliers, as if they were triggered by standard T1 trigger. We have proven that 16-point input is sufficient for the ANN pattern recognition [13]. The next step was to find in the 16-point trace, which corresponds to triggered events. To clearly see the beginning of the event we decided that two first points should be on the pedestal level. Afterwards we subtracted the pedestal level from all used data. These extracted points we could finally use for training and testing our neural network.

For a training procedure, we decided to use half of the data available for the testing procedure. We arrange data, to have proton and neutrino traces alternately. The proton traces were treated as negative signals ("0" for the ANN) and the neutrino traces were treated as positive ones ("1" for the ANN). This step allowed the ANN to teach faster and get less errors while training. The testing procedure consists on assigning specific value to the trace. This value depends on the coefficients of the trained ANN. If the value is greater than the threshold, the trace is treated as a neutrino trace, otherwise it is treated as a proton trace. The efficiency of the neutrino recognition with specific threshold level can be defined as number of neutrino traces recognized correctly divided by the number of all neutrino traces. The proton mistakes level is defined as number of proton traces treated as neutrino traces divided by the number of all proton traces.

The testing procedure was divided into two stages. First, we wanted to find out if we could use the data from the specific category to distinguish muon neutrinos and protons for all the angles or all the energies. Simulated data contains only three different angles: 80°, 85° and 89°, respectively, but we do not expect the zenith angle of the particle would be exactly like them. If the ANN trained on the specific category with angle 85° can distinguish neutrinos and protons also for 89° and 80° with acceptable efficiency, we assumed it could distinguish protons and neutrinos also for full angle range: 80° - 89°. The same is for energies. The ANN had been also trained by the data of the specific energy and then it was tested on the other values of the chosen parameter. The second step of the testing procedure consisted on training the ANN by the randomly taken data from all categories. In this case we used 30% of the whole data for training the ANN. The result from this step should be better than in previous one, because the data used for teaching was taken from the wider set of data.

The efficiency of the ANN strongly depends on the data used for training. The positive and negative signals should be as different as possible, to increase the distinction of proton and neutrino traces. Our first results (Fig. 9 continuous line) shows, that ANN does not work properly. There was no separation between protons and neutrinos. When we looked at the data we used for teaching the ANN, we found that some of the neutrino and proton traces looked very similar to each other. Moreover the simulated traces produced by neutrino shower with various distances to the detector, but with the same energy and angle was diametrically different. The same effect we observed in the other angles and energies (Fig. 2) and in the traces produced by protons (Fig. 7). This effect is directly connected with the electromagnetic (EM) part of the shower. If the distance to the detector is short, the EM part of the shower gives second component in the traces, additional to standard muonic one. At high zenith angles the proton showers should not have the EM component, because it should disappear after 2000 - 3000 $g/cm^2$. Old

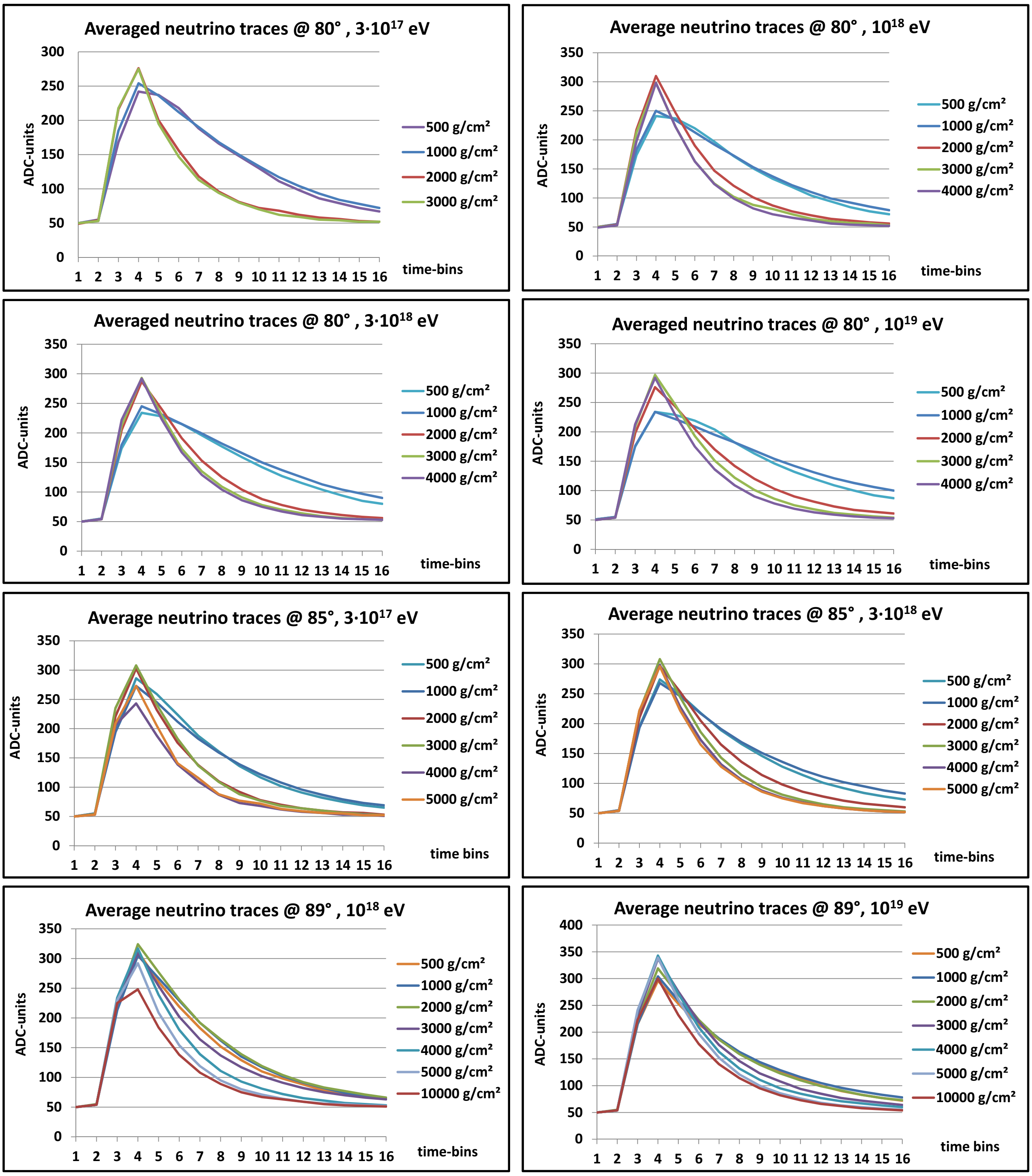
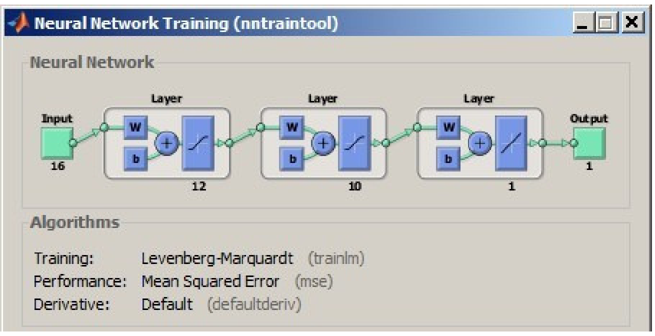


Fig. 2. Plots contain averaged neutrino signal waveforms (ADC traces) for various angles, energies and initialization points. The exponents of the traces at distances 500 and 1000 $g/cm^2$ are different than on the rest of distances. This effect does not depends on the energy and slightly depends on the angle.

neutrino showers looked like old proton neutrino showers, so we decided to separate the data and focus on recognizing only the young neutrino showers, where the EM component was still visible. We also decided to remove proton showers with visible EM component, because traces they generated looked similar to traces generated by young neutrino showers.

Moreover, for this showers the probability to occur on this angles was low. The data we decided to keep were all neutrino traces with distances 500 and 1000 $g/cm^2$ and all proton traces with two maximal distances for each angle. Fig. 5 shows the average traces for data at 80° after the separation. Neutrino and proton traces have completely different shapes and it would be easier to recognize the neutrino traces when ANN is learned and tested on this data. Fig. 6 shows the histogram of the average exponents of rejected and accepted proton and neutrino categories. We can observe that the exponents of the rejected neutrino categories corresponds with exponents of the accepted proton categories. This was probably the main reason of low distinction of protons and neutrinos by ANN. Additionally, average exponents of the accepted proton and neutrino categories are separated.

## V. MATLAB ANALYSIS



Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Fig. 3. An example of the ANN structure used for an optimization in Neutral Network Training of the MATLAB toolbox.

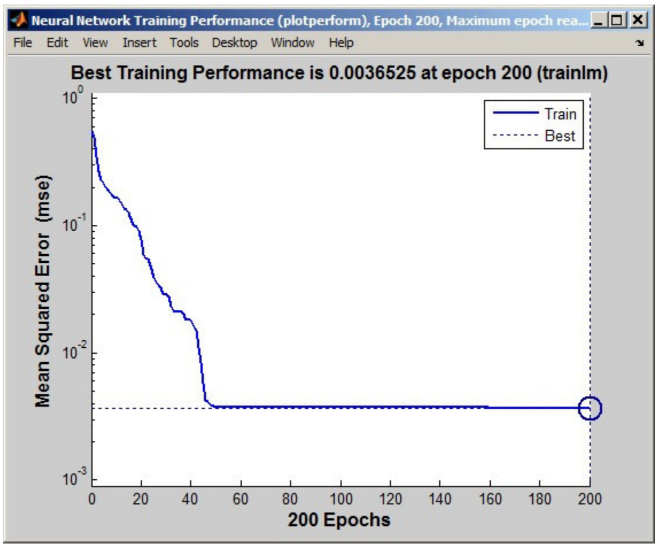

Neural Network Training Performance (plotperform), Epoch 200, Maximum epoch rea...
File Edit View Insert Tools Desktop Window Help
Best Training Performance is 0.0036525 at epoch 200 (trainlm)
Train
Best
Mean Squared Error (mse)
$10^0$
$10^{-1}$
$10^{-2}$
$10^{-3}$
0
20
40
60
80
100
120
140
160
180
200
200 Epochs

Fig. 4. Training performance for the 3-layer network 12-10-1

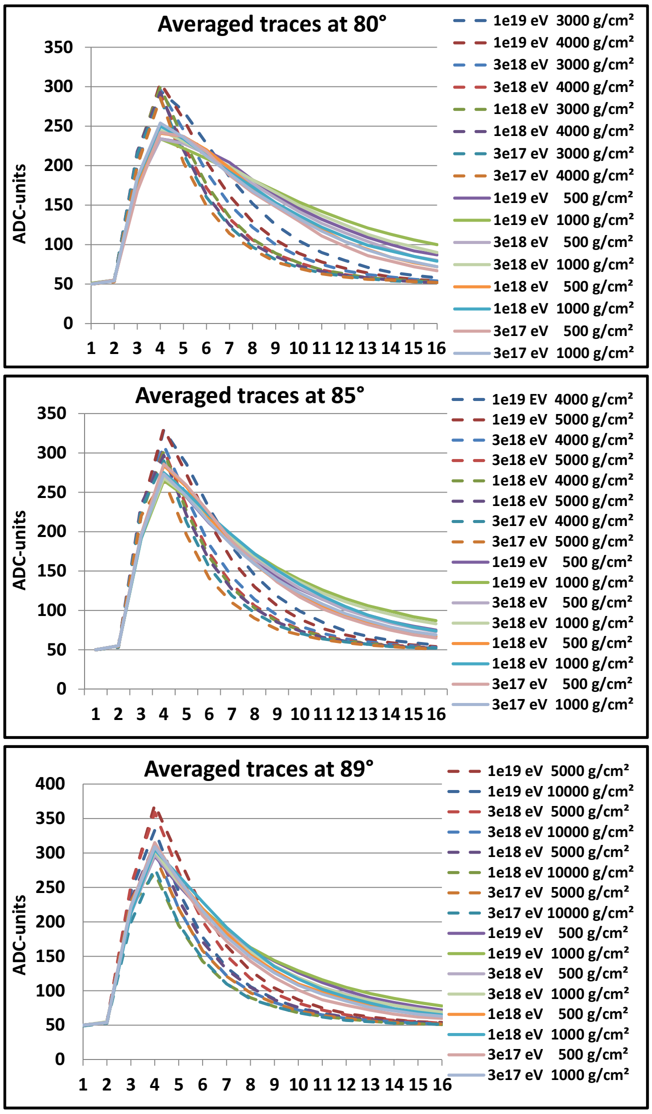


Fig. 5. The plot showing the differences between the traces produced by old proton showers (dashed line) and young neutrino showers (continuous line).

On Fig. 8a we can see how the angular efficiency behaves. In the case, when the ANN was tested on all categories, the ANN was learned using the specific category with parameters: angle 85°, energy $3\times10^{18}$ eV, distance to detector 2000 $g/cm^2$ both for protons and neutrinos. The separated case used 80°, $3\times10^{18}$ eV and 500 $g/cm^2$ as positive signals (neutrinos) and 80°, $3\times10^{18}$ eV and 4000 $g/cm^2$ as negative signals (protons). For the case, which use all categories for tests, the efficiency seems to be independent of the angle. The case, which use only traces produced by young neutrino and old proton showers ("separated" case) shows the efficiency changes slightly, when we change the angle. On Fig. 8b we can see that in both cases the efficiency of the ANN is independent of the energy (in the used energy range). Fig. 9 shows the comparison between both cases. The ANN tested on non-separated traces has problems with distinguishing the proton and neutrinos on every level of the threshold. The effectiveness of finding neutrino traces and the level of proton mistakes differs only slightly. The second ANN, tested on the separated traces, can recognize protons and neutrinos with acceptable efficiency. The proton mistakes level is much lower than efficiency of finding the neutrino traces, moreover the efficiency of finding neutrino traces is higher than in the previous case. Finally, Fig. 10 shows the difference of the non-separated and separated cases, when the data taken for training the ANN was taken randomly from all traces, appropriate for specific case. We can see that, the distinguishing the proton and neutrino traces increased in both cases, but the recognition of the neutrinos and protons is much better when we teach the ANN with the data from separated traces. In this case the neutrino efficiency to proton mistakes level ratio is greater than two, when in the other case this ratio is on the level of 1.3. This ratio is very important, because it shows if the ANN works or not. The ratio on the level of one means the ANN does not work, because the neutrino to proton traces quotient is the same as in the whole data. The data triggered as neutrino would be processed off-line in next steps. Lower level of proton mistakes and greater efficiency of recognizing neutrino traces means that amount of data to process off-line would be smaller, but neutrino traces would appear more frequently in it.

## VI. FPGA Implementation

A 16-input neuron for 14-bit data and 14-bit coefficients is shown on Fig. 11. An neuron output drives a neural transfer function - a tansig, which calculates a layer output from its net input. It can be implemented as ROM in embedded FPGA memory. We selected 14-bit input, 14-bit-output tansig implementation in RAM: a 2-port function with blocked writing left port to keep a reasonable compromise between a calculation accuracy and a memory size. The same array of coefficients is used for two independent neuron transfer functions.

The 12-10-1 network (Fig. 3) offers the best performance (Fig. 4) with a minimal resource occupancy, however, it requires 23 neurons. Due to the limited amount of DSP blocks, we could use this network for a single PMT only. The Quartus compiler allows a compilation with arbitrarily selected implementations of the multipliers: either in the DSP blocks or in logic elements only. An implementation of the multipliers in the Adaptive Logic Modules (ALMs) is much more resource-consuming (1247 ALMs instead of 107 ALMs + 8 DSP blocks). However, such a selection allows the implementation of a more complicated network, which provides a similar registered performance (keeps approximately the same speed). The 3-channel 12-10-1 network needs 36 neurons (1st layer implemented in the DSP blocks) + 33 neurons implemented in 41151 ALMs (36.5% of 5CEFA9F31I7).

## VII. Conclusion

We run CORSIKA simulation for proton, iron, $\nu_\mu$ and $\nu_\tau$ primaries. Output data collected on 1450 m (the level of the Pierre Auger Observatory) was an input for OffLine providing the ADCs response in the WCDs. Obtained ADC traces were thus used to teach the 12-10-1 neural network implemented already in the 5CEFA7F31I7 FPGA on the Cyclone® V development kit. This FPGA is smaller version of the chip

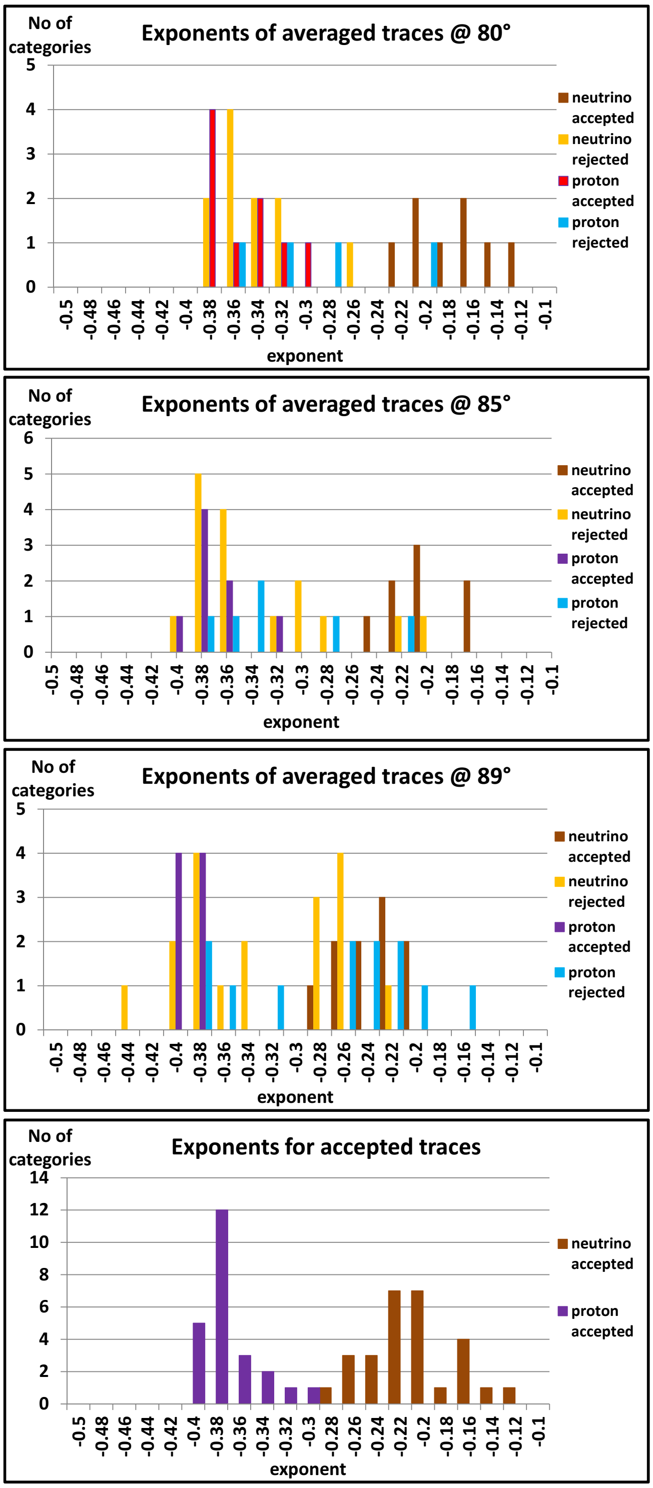


Fig. 6. A histogram of exponents of rejected and accepted traces for protons and neutrinos. The accepted neutrino traces exponents are separated from the accepted proton traces exponents. The rejected neutrino traces looks very similar like traces made by old proton showers. This may cause problems with ANN training and may increase the level of the proton traces recognized as neutrino traces by ANN.

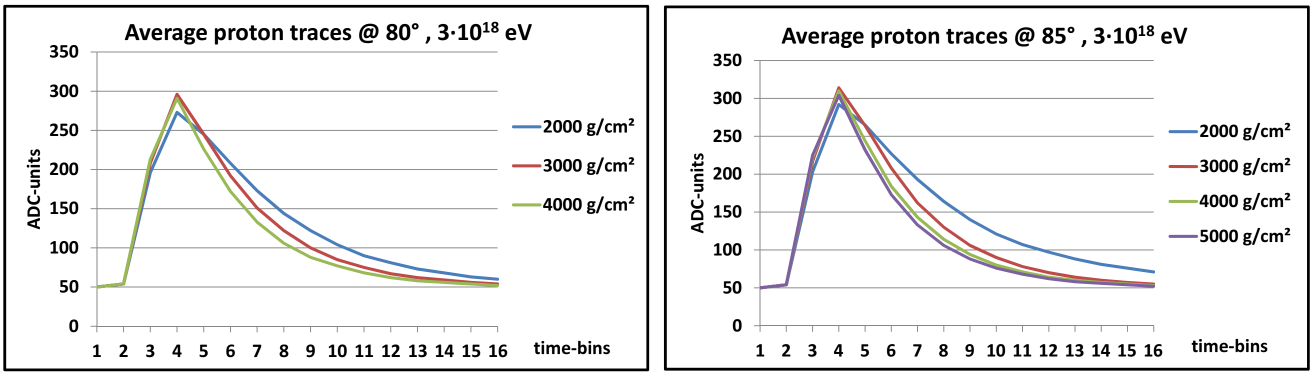


Fig. 7. Plots showing average traces for protons for different distances. As in Fig. 2 the shortest distances have different exponents than the rest.

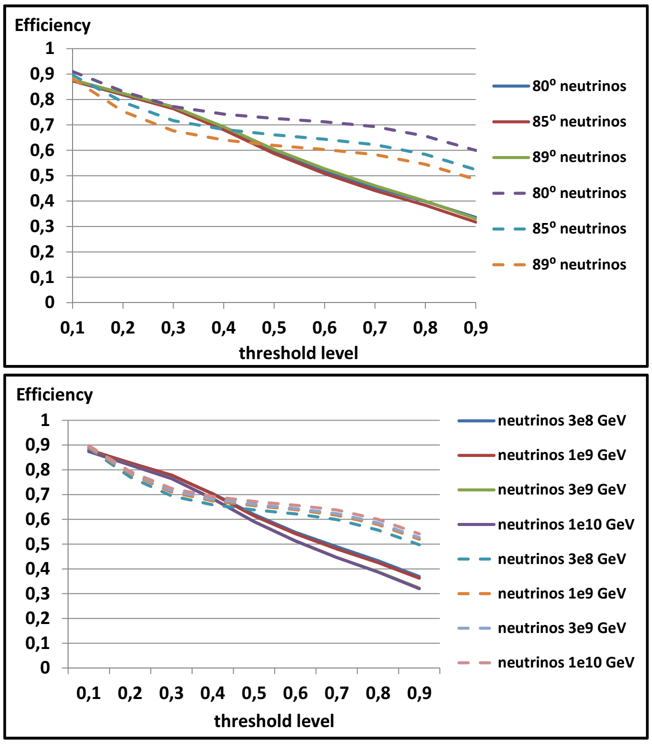


Fig. 8. The angular dependency of a neutrino efficiency as a function of a threshold level for angles (top) and energies (bottom). The continuous lines denote the data for testing were gathered from all traces. The dashed lines represent the case with separated traces.

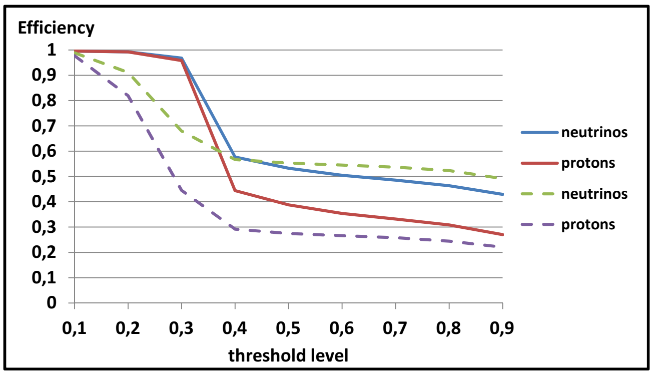


Fig. 9. The efficiency of the neutrinos recognition and proton mistakes level in function of threshold level for the non-separated (continuous lines) and separated traces (dashed lines) used for testing the ANN. The data taken for training are the same as in the Fig 8.

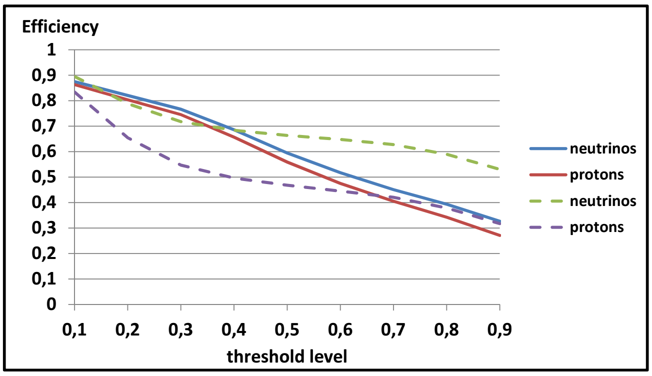


Fig. 10. The efficiency of the neutrino traces recognition and proton mistakes level in function of threshold level for the non-separated (continuous line) and separated traces (dashed line) used for testing the ANN. The traces used for training the ANN were randomly taken from all categories of non-separated or separated traces respectively.

being designed for the Front-End Board for Auger-Beyond-2015 task.

Preliminary results show that the ANN algorithm can detect neutrino events currently neglected by the standard Auger triggers.

The recognition efficiency of the neutrino traces by the ANN algorithm strongly depends on the differences between the data used for the ANN training. If we teach the ANN with the data containing only traces produced by young neutrino and old proton cosmic air showers we can reach the acceptable level of recognition. Moreover, we can distinguish protons and neutrinos, which means the ANN works on very promising level.

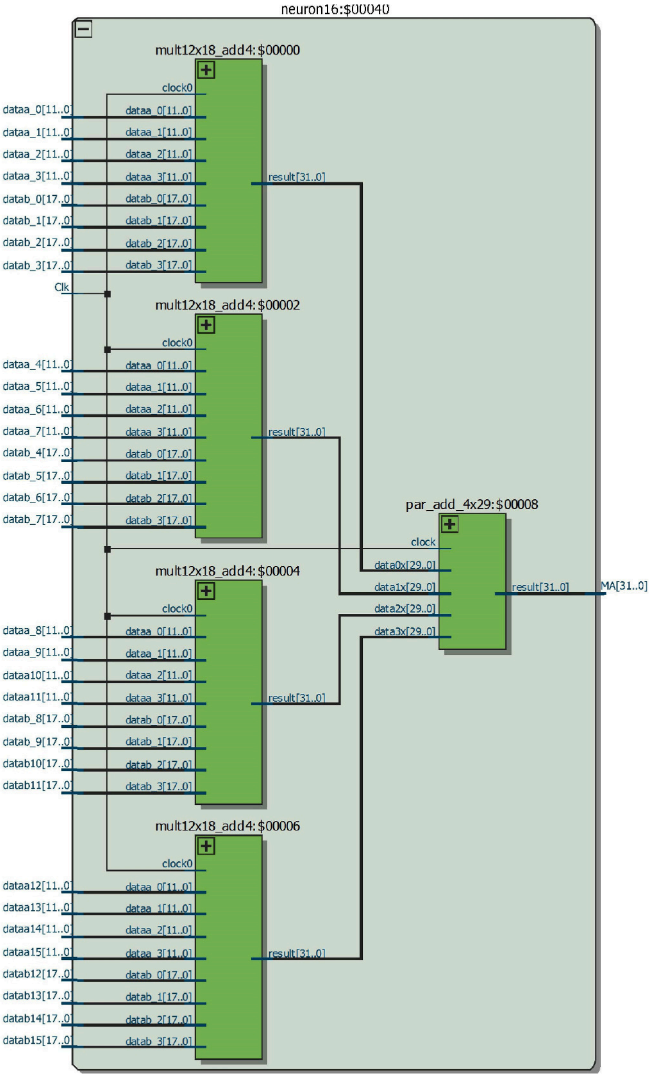


Fig. 11. An internal structure of an FPGA neuron.

## Acknowledgment

This work is supported by the National Science Centre (Poland) under NCN Grant No. 2013/08/M/ST9/00322. Authors thank the Pierre Auger Collaboration for an opportunity of using the CORSIKA and OffLine simulation packages.

## References

[1] M. Nagano and A. A. Watson, "Observations and implications of the ultrahigh-energy cosmic rays", *Rev. of Modern Phys.*, vol. 72, no. 3, pp. 689-732, 2000.
DOI: 10.1103/RevModPhys.72.689

[2] F. Halzen and D. Hooper, "High-energy neutrino astronomy: the cosmic ray connection", *Rep. on Progress in Phys.*, vol. 65, no. 7, p. 1025, 2002.
DOI: 10.1088/0034-4885/65/7/201

[3] J. K. Becker, "High-energy neutrinos in the context of multi-messenger astrophysics", *Phys. Reports*, vol. 458, no. 4-5, pp. 173-246, 2008.
DOI: 10.1016/j.physrep.2007.10.006

[4] P. Bhattacharjee and G. Sigl, "Origin and propagation of extremely high-energy cosmic rays", *Phys. Reports*, vol. 327, no. 3-4, pp. 109-247, 2000.
DOI: 10.1016/S0370-1573(99)00101-5

[5] K. Greisen, "End to the cosmic-ray spectrum?", *Phys. Rev. Lett.*, vol. 16, pp. 748-750, 1966.
DOI: 10.1103/PhysRevLett.16.748
G. T. Zatsepin and V. A. Kuzmin, "Upper limit of the spectrum of cosmic rays", *JETP Letters*, vol. 4, p. 78, 1966.
WOS:A19668298400011

[6] [Hi-Res Fly's Eye Collaboration], "First Observation of the Greisen-Zatsepin-Kuzmin Suppression", *Phys. Rev. Lett.*, vol. 100, no. 10, article 101101, 5 pages, 2008.
DOI: 10.1103/PhysRevLett.100.101101

[7] K. Kotera, D. Allard, and A. V. Olinto, "Cosmogenic neutrinos: parameter space and detectability from PeV to ZeV", *JCAP*, vol. 2010, no. 10, article 013, 2010.
DOI: 10.1088/1475-7516/2010/10/013

[8] E. Zas, "Neutrino detection with inclined air showers", *New Journal of Phys.*, vol. 7, p. 130, 2005.
DOI: 10.1088/1367-2630/7/1/130

[9] V. S. Berezinsky, A. Yu. Smirnov, "Cosmic neutrinos of ultrahigh energies and detection possibility", *Astrophys. and Space Sci.*, vol. 32, no. 2, pp. 461-482, 1975.
DOI: 10.1007/BF00643157

[10] Z. Szadkowski, "A spectral $1^{st}$ level FPGA trigger for detection of very inclined showers based on a 16-point Discrete Cosine Transform for the Pierre Auger Observatory", *Nucl. Instr. Meth.*, ser. A, vol. **606**, pp. 330-343, July 2009.
DOI: 10.1016/j.nima.2009.03.255

[11] Z. Szadkowski, "Trigger Board for the Auger Surface Detector with 100 MHz Sampling and Discrete Cosine Transform", *IEEE Trans. on Nucl. Science*, vol. 58, pp. 1692-1700, Aug. 2011
DOI: 10.1109/TNS.2011.2115252

[12] Z. Szadkowski, "Optimization of the Detection of Very Inclined Showers Using a Spectral DCT Trigger in Arrays of Surface Detectors", *IEEE Trans. on Nucl. Science*, vol. 60, pp. 3647-3653, Oct. 2013
DOI: 10.1109/TNS.2013.2280639

[13] Z. Szadkowski, K. Pytel, *Artificial Neural Network as a FPGA Trigger for a Detection of Very Inclined Air Showers*, IEEE Trans. on Nucl. Science, vol. 63, Issue 3, pp. 1002-1009, June 2015.
DOI: 10.1109/TNS.2015.2421412

[14] [Pierre Auger Collaboration], "Upper limit on the diffuse flux of ultrahigh energy tau neutrinos from the Pierre Auger Observatory", *Phys. Rev. Lett.*, vol. 100, no. 21, article 211101, 2008.
DOI: 10.1103/PhysRevLett.100.211101

[15] [Pierre Auger Collaboration], "Ultrahigh Energy Neutrinos at the Pierre Auger Observatory", *Adv. in High Energy Phys.* vol. 2013, Article ID 708680, 18 pages
DOI: 10.1155/2013/708680

[16] CORSIKA an Air Shower Simulation Program [Online]. Available: https://web.ikp.kit.edu/corsika/

[17] S. Argiro et al., "The offline software framework of the Pierre Auger Observatory", *Nucl. Instrum. and Meth.* ser. A, vol. 580, Issue. 3, pp. 1485-1496, Oct. 2007.
DOI: 10.1016/j.nima.2007.07.010

DOI: 10.15439/2015F159
ACSIS, Vol. 5

# DCT trigger for a detection of very inclined showers in the Pierre Auger surface detector Engineering Array in the new High-Resolution Front-End based on the Cyclone V FPGA

Zbigniew Szadkowski
University of Łódź
Department of Physics and Applied Informatics
Faculty of High-Energy Astrophysics
90-236 Łódź, Pomorska 149
e-mail : zszadkow@kfd2.phys.uni.lodz.pl

***Abstract*—The paper presents the first results from the trigger based on the Discrete Cosine Transform (DCT) operating in the new Front-End Boards with Cyclone V FPGA deployed in 8 test surface detectors in the Pierre Auger Engineering Array.**

**The patterns of the Analog-to-Digial Converter (ADC) traces generated by very inclined showers were obtained from the Auger database and from the CORSIKA simulation package supported next by OffLine reconstruction Auger platform which gives a predicted digitized signal profiles. Simulations for many variants of the initial angle of shower, initialization depth in the atmosphere, type of particle and its initial energy gave a boundary of the DCT coefficients used next for the on-line pattern recognition in the FPGA.**

**Preliminary results have proven a right approach. We registered several showers triggered by the DCT for 120 MSps and 160 MSps.**

## I. Introduction

THE AIM of the Pierre Auger Observatory [1] is measuring of cosmic rays at the ultra-high energy with sufficient statistics and resolution. 1680 water Cherenkov surface detectors (SD) are distributed over an area of 3000 km$^2$ for measuring the charged particles associated with extensive air showers (EAS) and 24 telescopes with 30 × 30 degrees of field of view and 12 m$^2$ mirror area each to observe the fluorescent light produced by charged particles in the EAS during operation on clear moonless nights. The simultaneous observation of EAS by the ground array and the fluorescent light known as a "hybrid" event [2] [3] improves the resolution of the reconstruction considerably and, thanks to the calorimetric nature of the fluorescent light emitted, provides energy measurements virtually independent of hadronic interaction models. Each SD station is equipped in three 9-inch photomultiplier tubes (PMTs) reading out the Cherenkov light from the 12 $m^3$ of purified water contained in each tank. Two independent trigger modes are implemented in the SD to detect, in a complementary way, the electromagnetic and muonic components of an air-shower [4]. The first mode is a simple threshold trigger, which requires the coincidence of the three PMTs each above 1.75 $I^{peak}_{VEM}$. This trigger is used to select large signals that are not necessarily spread in time. The second mode takes into account that, for other than very inclined showers or signals from more vertical showers very close to the shower axis, the arrival of particles and photons at the detector is dispersed in time, at least 13 bins in 120 time bins of a sliding window of 3 $\mu$s are required to be above a threshold of 0.2 $I^{peak}_{VEM}$ in coincidence in two of any 3 PMTs.

Very inclined showers generated by hadrons and starting their development early in the atmosphere produce a relatively thin muon pancake (∼1m thickness) on a detection level. Ultra-relativistic charged particles trespassing the water in a surface detector generate the Cherenkov light detected next in photo-multipliers (PMT). A direct light gives a peak with a very short rise time and fast exponential attenuation. The DCT trigger allows recognition of ADC traces with specific shapes The standard trigger requires 3-fold coincidences in a single time bin. The present sampling frequency in the surface detectors is 40 MHz. The new Front-End Board developed for the Auger-Beyond-2015 surface detector upgrade allows a sampling up to 250 MSps (120 MSps and 160 MSps were used in tests).

Neutrinos can generate showers starting their development deeply in the atmosphere, known as "young" [5] [6]. They contain a significant amount of an electromagnetic component, usually preceded by a muon bump. Simulations show [7] that it is often fully separated from the electromagnetic fraction and the 16-point DCT algorithm can also be used..

A probability of 3-fold coincidences of direct light corresponding to a standard Auger trigger is relatively low. Much more probable are 2-fold coincidences of a direct light. The 3rd PMT is next hit by reflected light, but with some delay. By fast sampling (120-160 MSps) this delay gives signal in the next time bins. The standard T1 trigger ceases giving a sufficient rate for horizontal and very inclined showers. The rate drops down below an acceptable level. We had to modify the T1 trigger to get approximately standard trigger rate.

978-83-60810-66-8/$25.00©2015, IEEE

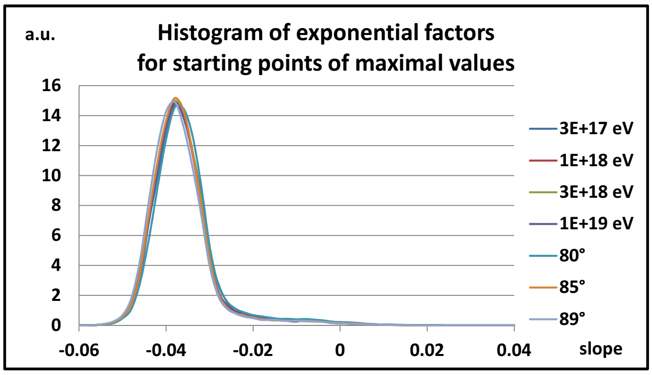


Fig. 1. Histogram of exponential factors for traces starting from the maximal values.

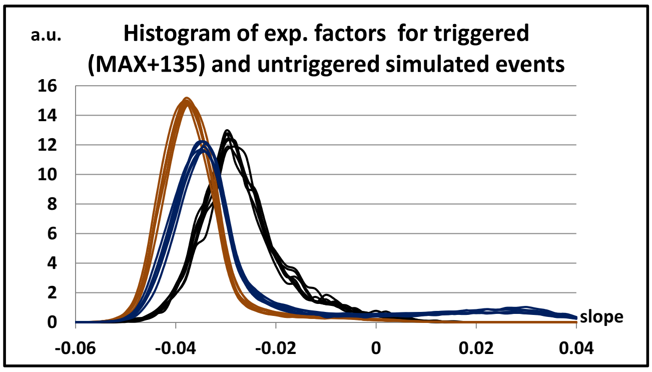


Fig. 2. Histogram of exponential factors for triggered traces starting from the maximal values (as in Fig. 1 - red), from the standard T1 threshold = 1.75 VEM (135 ADC-units)(blue) and for non-triggered (black) events.

Cherenkov light generated by very inclined showers crossing the Auger tank can reach the PMT directly without reflections on Tyvec® liners (a special material with 95% of reflectivity). Especially for "old" showers the muonic front is very flat. This together corresponds to very short direct light pulse falling on the PMT and in consequence very short rise time of the PMT response. For vertical or weakly inclined showers, where the geometry does not allow reaching the Cherenkov light directly on the PMT, the light pulse is collected from many reflections on the tank walls. Additionally, showers developing for not so high slant depth are relatively thick. These give a signal from a PMT as spread in time and relatively slow increasing. A very short rise time together with a relatively fast attenuated tail could be a signature of very inclined showers. We observe numerous very inclined showers crossing the full array but which "fire" only few surface detectors. For that showers much more tanks should have been hit. Muonic front produces PMT signals either too low for 3-fold coincidences or desynchronized in time. This may be a reason of "gaps" in the array of activated tanks.

Two-fold coincidences of DCT coefficients allow triggering signals currently being ignored due to either too high amplitude threshold or due to their de-synchronization in time causing a tank geometry. Three DCT engines implemented into Cyclone® V E 5CEFA9F31I7N FPGA used ~60% of DSP blocks generate the spectral trigger, when in at least 2 channels 8 DCT coefficients simultaneously are inside the acceptance lane. Additional veto signal (analyzing the amplitude) controls a trigger rate to avoid a saturation of a transmission channel. Both lab and long-term field measurements on the test tank confirm a high efficiency of the recognition of expected patterns of ADC traces.

Auger data (40 MSps) show that hadron induced showers with dominant muon component (investigated in a zenith angle of $70-90°$) give an early peak with a typical rise time mostly from 1 to 2 time bins and exponentially attenuated tail. Very inclined showers with well defined shape can be detected by a pattern recognition technique, in particular by a spectral trigger based on the Discrete Cosine Transform [8] [9] [10] [11]. The DCT trigger cannot be implemented into currently used Front-End Boards with ACEX® [12] and Cyclone® [13] FPGAs due to a lack of DSP blocks. New Front-End Board with Cyclone® V E FPGA successfully passed tests on the field collecting Auger data even with 160 MHz sampling [9]. The DCT trigger was tested with 120 and 160 MHz. Scaled to the 1st harmonics DCT coefficients are independent of the amplitude. They are determined by the shape only. This causes that the DCT trigger could also be accidentally generated by very low noise pulses with the required characteristics. In order to cut this type of the spurious triggers, the spectral trigger was also supported by the "veto" amplitude threshold of ~20 ADC-counts. Pulses above this threshold are analyzed next with the DCT engines.

Due to noise and other factors which can distort the shape of the PMT pulse, strict conditions imposed on all coefficients are too strong. The DCT trigger has been generated if a subset of DCT coefficients (similar to the Occupancy in the ToT) has been simultaneously "fired". Simulations performed for various artificial level of additional noise (0 - 6 ADC-counts) showed that even a significant noise contribution does not violate an efficiency of the DCT trigger. A little bit higher noise in Cyclone® V E Front-End in comparison to a noise level in the previous FEB generations with 40 MHz sampling should not be a significant factor. The new 160 MHz sampling Cyclone® V E Front-End Boards are not equipped in the anti-aliasing filters. This keeps a flexibility for various possible sampling frequencies, however, registered shape is not smoothed and provides even very short one time-bin peaks (6.25 ns) corresponding to reflections from Tyvec walls. Such a very precisely reconstruction of timing in the ADC traces violates in some sense a fundamental assumption on a smooth exponential attenuation of the tail. We had to introduce some corrections to take into account also this effect. The CORSIKA and OffLine simulations do not show that the reflections may significantly reduce an efficiency of this approach. However, the Auger simulation/reconstruction packages assume 60 MHz cut-off Nyquist filter, which in real Cyclone® V E Front-End has not been introduced. Thus, the simulations (available only for the "official" 120 MHz sampling frequency) give rather hints for higher 160 MSps sampling and real DCT

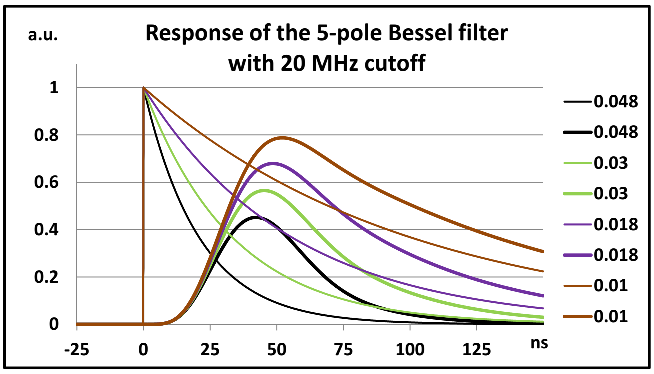


Fig. 3. Pulses as a response on a unity jump with an exponential attenuation with the factors $\alpha$ = 0.048, 0.03, 0.018 and 0.01, respectively.

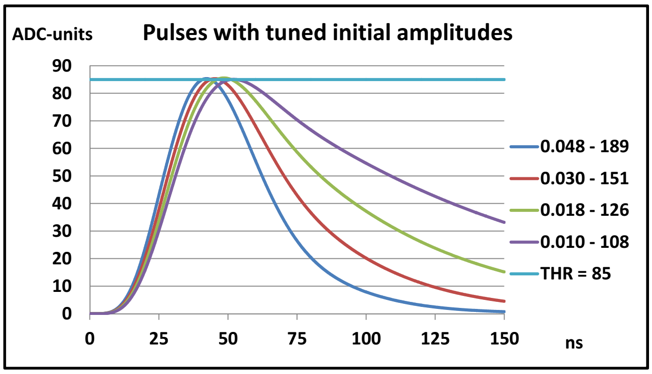


Fig. 4. Pulses filtered by the 5-pole Bessel filter with 20 MHz cutoff with amplitudes: 189, 151, 126 and 108 ADC-units, respectively for 0.048, 0.03, 0.18 and 0.01 exponential factors.

trigger implementation required several step by step optimizations.

## II. CORSIKA and OffLine simulations

We simulated the surface detector response on proton induced showers for energy $3 * 10^{17}$ eV, $10^{18}$ eV, $3 * 10^{18}$ eV and $10^{19}$ eV and for angles 80°, 85° and 89°. The initial points were selected for 2, 3, 4, 5 and 10 $kg/cm^2$, if the angle allowed on this geometrical configuration. ADC traces were collected as triggered (3-fold coincidences in a single time bin) and non-triggered (the standard T1 trigger condition not obeyed, traces registered in neighboring detectors). The standard T1 threshold 1.75 VEM was set in simulations.

Data from standard 40 MHz sampling simulation was used for an optimization of the condition of the DCT trigger for 120 and 160 MSps. Data for 16-point DCT trigger for 160 and 120 MSps corresponds to 100 and 133 ns of ADC trace, respectively. These intervals correspond to 4 and 5 time bins for the standard 40 MSps sampling. As expected proton induced showers generate ADC traces with rapid jump from a pedestal level with relative fast exponentially attenuated tails. We investigated exponential factors of traces in the interval of 125 ns (5 time-bins) selecting parts of traces starting from the maximum value (MAX for triggered (Fig. 1) and non-triggered) and from the standard threshold level (135 ADC-units for triggered only). According to Fig. 2 we see that ranges of exponential factors depend on amplitudes of signals. Traces investigated from the maximum values (variant A)(red curves on Fig. 2) shows faster attenuation than investigating from the T1 threshold level (variant B). Non-triggered traces (with amplitudes lower than T1 threshold level at least in one channel) show much slower attenuation (variant C). For we get the following ranges for attenuation factors $\alpha$ in the function ADC = exp(-$\alpha$*t), where t in nanoseconds.

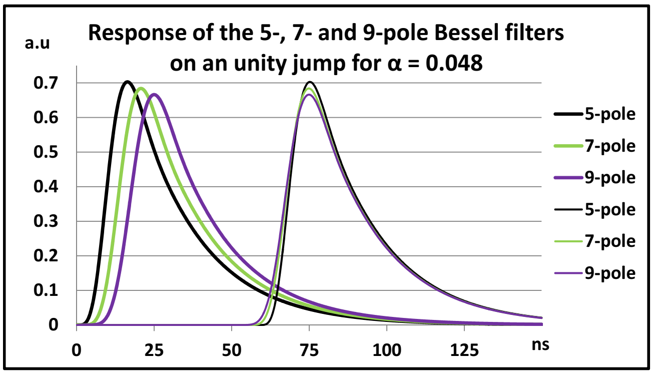


Fig. 5. Response on an unity jump of Bessel filters with 5,7,and 9 poles, respectively. The left curves show response on a jump for 0 ns. The right curves are adjusted to get maximum in the same time bin.

- variant A : 0.048 - 0.030 : ∼89-90% of traces
- variant B : 0.048 - 0.010 : ∼83-86% of traces
- variant C : 0.042 - 0.014 : ∼89-91% of traces

The variant B gives the worst estimation, however, it is not unexpected, as traces may still rise after crossing the T1 threshold. This variant is presented to show the dependence of the exponential factor on the amplitude of signal.

## III. 5-pole Bessel filter with 20 MHz cutoff

According to the Nyquist theorem digitized ADC data should be filtered by the anti-aliasing filter before their processing. For the 40 MSps sampling the cutoff of the filter is 20 MHz. The ACEX® and Cyclone® FEBs were equipped in 5-pole Bessel filters. The Bessel filter provides a maximally flat group/phase delay (maximally linear phase response), which preserves the wave shape of filtered signals in the passband. Higher order of the filter gives steeper frequency characteristic for stop-band, however it requires more sophisticated filter circuit with relatively low values of capacitors. The parasitic PCB capacity became no negligible factor. The transfer function H(s) for the 5-pole Bessel filter with 20 MHz cutoff is as follows:

$$H(s) = \frac{K}{(s-p)(s-z_1)(s-z_1^*)(s-z_2)(s-z_2^*)} \qquad (1)$$

where $K = 3.1336e + 40$, $p = -0.11642e9$, $z_1 = -0.07421e9 + 0.114e9 * j$, $z_2 = -0.10701e9 + 0.005563e9 * j$.

Let us calculate the response of the Bessel filter on the unity jump with the exponential attenuation tails with $\alpha$ = 0.048, 0.03, 0.018 and 0.01, respectively. The Laplace transform for the function x(t) = exp(-$\alpha$*t) is X(s) 1/(s+$\alpha$). The factor $\alpha$ is given for time in nanoseconds. The signal response we get with the inverse Laplace transform:

$$y(t) = Y(s)^{-1} = (H(s) * X(s))^{-1} \tag{2}$$

$$\begin{aligned} Y(t) &= (A * sin(z'_{Im,1} * t) + B * cos(z'_{Im,1} * t) * e^{z'_{Re,1} * t} \\ &+ (C * sin(z'_{Im,2} * t) + D * cos(z'_{Im,2} * t) * e^{z'_{Re,2} * t} \\ &+ E * e^{p' * t} + F * e^{-\alpha * t} \end{aligned} \tag{3}$$

where: $p' = p * 10^{-9}$ $z'_{1,2} = z_{1,2} * 10^{-9}$

TABLE I
COEFFICIENTS FOR EQ. 3 , WHERE TIME GIVEN IN NANOSECONDS

| $\alpha$ | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 0.048 | 1.1258 | 1.2349 | -10.080 | 3.4118 | -9.7361 | 5.0893 |
| 0.030 | 1.2354 | 1.0147 | -8.1538 | 4.0062 | -7.7082 | 2.6873 |
| 0.018 | 1.2697 | 0.8677 | -7.1206 | 4.1119 | -6.7684 | 1.7888 |
| 0.010 | 1.2778 | 0.7740 | -6.5323 | 4.1101 | -6.2596 | 1.3754 |

Fig. 3 shows responses on an unity jump of the input signal with various exponential attenuation factors. We can see that:

- pulses are delayed on 40 - 50 ns,
- an amplitude of pulses depends on the exponential factor.

A dependence of output signal amplitude on an exponential factor causes that short pulses have to have higher amplitudes to be above the threshold level. Fig. 4 shows pulses with tuned amplitudes to reach at least the T1 threshold level (1.75 VEM ~85 ADC-units).

Let us notice that for all investigated values of $\alpha$ we get : $\alpha < |z_{Re,1}| < |z_{Re,2}| < |p|$. This causes that the exponential term $e^{-\alpha * t}$ attenuates slower than other terms in Eq. 3and practically it only survives for $t$ >~50 ns (after a maximum). So, the attenuation (exponential) factor for a output tail remains the same as in the input signal with rapid jump and exponential tail.

This means that all estimations of attenuation factors for ADC traces from the shower simulations remains valid and the Bessel filter does not affect the attenuation characteristics after the maximum of pulse. The rise time still depends significantly on an order of the filter and the cutoff frequency.

## IV. BESSEL FILTERS FOR THE SDE UPGRADE

According to golden rules the new Front-End electronics with the sampling frequency of 120 MHz should be equipped in the 60 MHz cutoff anti-aliasing filter. Higher order of the filter provides sharper frequency characteristic in the stop-band, but it introduces slightly bigger suppression and few nanoseconds additional delay (Fig. 5).

Fig. 5 shows that 9-pole filter in comparison to 5-pole one:

- introduces bigger delay,
- suppresses stronger the signal,
- provides the same attenuation factor for tails.

5-pole Bessel filter for the upgrade design seems to be enough. Nevertheless, the test system installed in the Auger Engineering Array is considered to verify various configurations for sampling frequencies, data resolutions, trigger conditions etc. The sampling frequencies determine the cutoff in the anti-aliasing filters. For 120 MSps the cutoff should equals 60 MHz. On the other hand this cutoff would be too low for the 160 MSps. The hardware implementation of the Bessel filter on the PCB is difficult (parasitic capacitors have to be taken into account, a tolerance of components should be better than 1%. For resistors it is no problem, however for capacitors it is. Namely, a dynamic switching of cutoff frequencies in hardware filters is impossible. A requirement of wide spectrum of sampling frequencies excludes any fixed filter to be implemented on the PCB.

For this reason the Front-End has been designed without internal anti-aliasing filter at all.

## V. RISE TIMES OF THE SIGNAL WAVEFORMS

Fig. 5 suggests that a rise time of a unity jump takes at least 2-3 time-bins (20 - 25 ns) for the 120 MSps. This is fully confirmed by the CORSIKA + Offline simulations. Fig. 6a shows a contribution of signals waveforms with 1-, 2- and more time bins, respectively. The contribution of 1-time-bin is

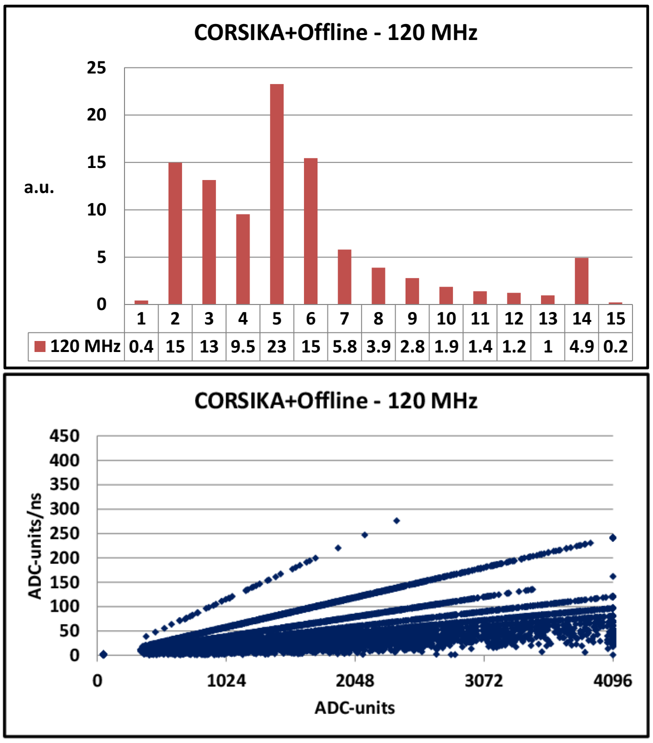
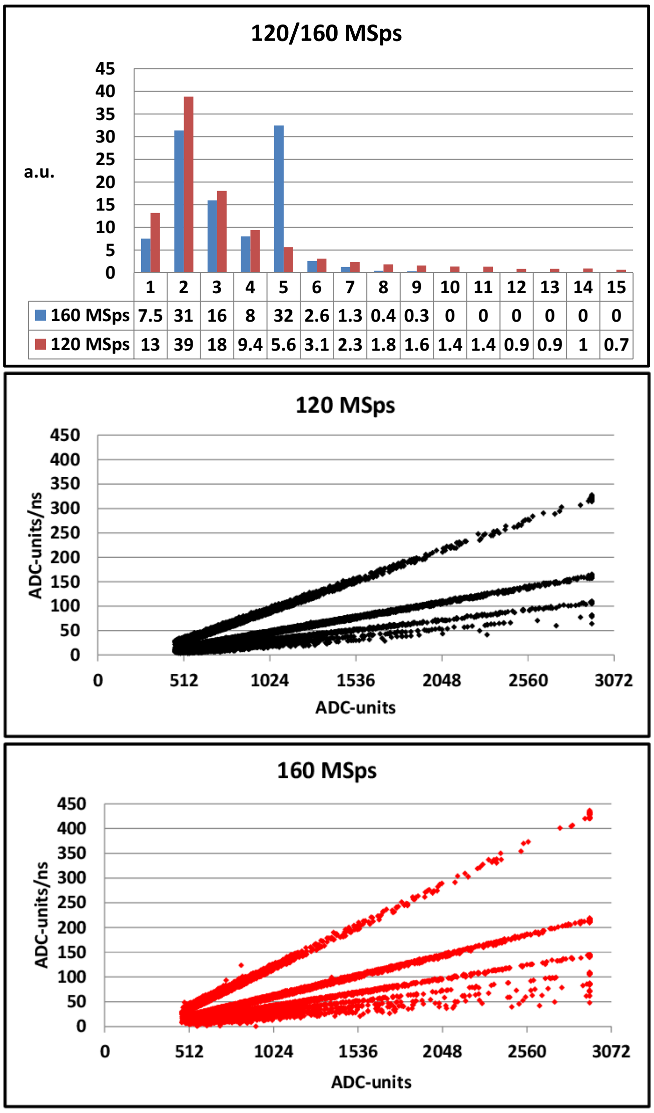


Fig. 6. A histogram showing a contribution of rise times for 1-,2- and more time bins for CORSIKA + Offline simulations with 120 MHz sampling frequency (upper graph) as well as a slope distribution of rising edges of signal waveforms vs. an amplitudes for 120 MSps (lower graph). Branches correspond to 1-, 2- and more time bins slopes from a pedestal level to maximum a maximum value.



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 160 MSps | 7.5 | 31 | 16 | 8 | 32 | 2.6 | 1.3 | 0.4 | 0.3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 🟥 120 MSps | 13 | 39 | 18 | 9.4 | 5.6 | 3.1 | 2.3 | 1.8 | 1.6 | 1.4 | 1.4 | 0.9 | 0.9 | 1 | 0.7 |

Fig. 7. A histogram showing a contribution of rise times for 1-,2- and more time bins for real measurements with 120 and 160 MSps (upper graph) as well as a slope distribution of rising edges of signal waveforms vs. an amplitudes for 120 MSps (middle graph) and 160 MSps (lower graph). Branches correspond to 1-, 2- and more time bins slopes from a pedestal level to a maximum value.

negligible (∼0.4%). Real measurements without anti-aliasing filters give much bigger 1-time-bin contribution: 13% for 120 MSps and 7.5% for 160 MSps (Fig. 7a).

Fig. 6b and Fig. 7b show slope distributions for simulations and real measurements, respectively, both for 120 MHz sampling frequency. We see that anti-aliasing filter introduces a significant inclination of the signal rising edges, which may dramatically affect a timing analysis based on rising times distributions and important for e.g. splitting of muon "pancakes" generated in very inclined showers. Charged muons are deflected in the Earth magnetic field and create two sub-showers. A precise timing analysis allows a significant improvement of a geometrical shower reconstruction.

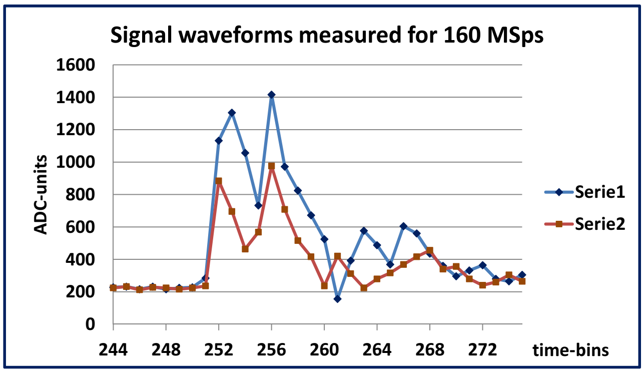


Fig. 8. Samples of signal waveforms registered with 160 MSps, which manifest light reflections from the Tyvec® walls.

Fig. 7c shows that for the 160 MSps a contribution signal waveforms with 1-time-bin rising edges is still high. It means a real rise time of many signal from the pedestal level to the maximal value may be even shorten than an interval of a single time bin.

Rising times for 160 MSps reach even ∼440 ADC-units/ns, which is 2-3 times faster than with even 5-pole Bessel filter. This may suggest that the rising edge is in fact steeper and with

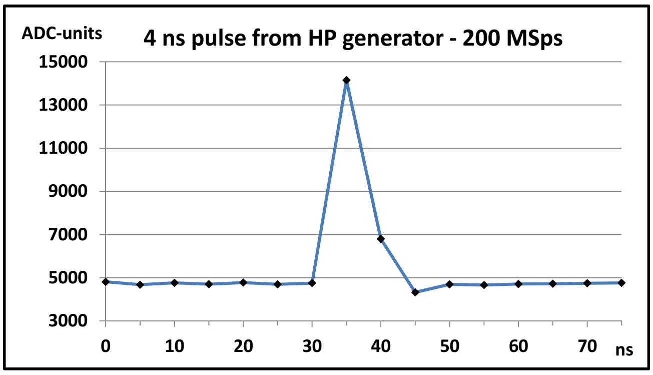


Fig. 9. A signal waveform for 4 ns pulse from the HP generator for the sampling frequency 200 MHz.

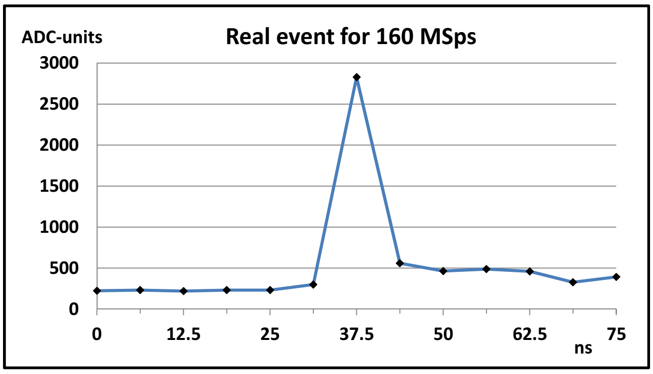


Fig. 10. A signal waveform of a real very short pulse registered for 160 MSps.

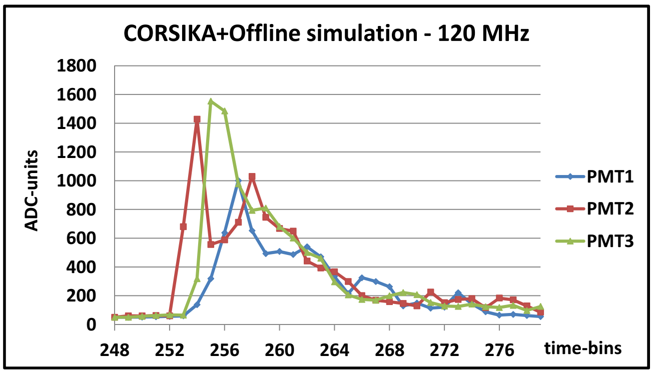


Fig. 11. Samples of signal waveforms simulated by the CORSIKA and Offline packages with 120 MSps, which manifest light reflections from the Tyvec® walls.

a faster sampling frequencies we would get still non-negligible amount of events located on the 1-time-bin branch.

A significant contribution of events with 5-time-bin rising edge comes from a light reflections from a Tyvec® walls (see samples of measured signal waveforms on Fig. 8 and simulated on Fig. 11). 3-4 time-bins for 160 MSps correspond to 18.75 - 25 ns and to 4 - 5 m of an additional geometrical distance for a light by n = 1.333 for a pure water.

The Front-End Board with Cyclone® V E is able to register even a 1-time-bin very short pulse. Fig. 9 shows a very short 4 ns pulse registered in the lab for a signal provided from the HP generator. ADC was working with 200 MHz sampling frequency. A similar pulse (Fig. 10) has been registered in real conditions for a signal obtained from PMTs of the tested surface detector. If the Front-End Board had been equipped in the anti-aliasing filter such pulses would have been strongly suppressed and probably not registered due to too low amplitude below a detection threshold.

## VI. DCT TRIGGER CONFIGURATIONS

16-point DCT algorithm for 120, 160 or 200 MHz sampling frequencies corresponds to 125, 93.75 or 75 ns, respectively, of sliding windows. We are interested in signal waveforms with a jump from a pedestal level to a maximum value and with the exponentially attenuated tails. As a pattern we can select 2, 3 or more time bins (Fig. 12) on a pedestal level following by the jump. DCT coefficients significantly depend on an amount of initial (pedestal level) time bins (Fig. 13).

The DCT coefficients are scaled to the $1^{st}$ harmonics. For the variant of 4 time bins, the DCT[1] is sometimes (for a particular tail slope $\alpha$ = 0.16667) very close to zero and as denominator in a scaling arbitrary units (a.u.) gives a huge value of a fraction (Fig. 14). So, the 4 time-bin variant provides an almost negligible sensitivity for $\alpha \neq 0.16667$.

For 160 and 200 MHz sampling frequencies 3 time-bins variants manifest the same undesirable coefficient characteristics similar to shown on Fig. 13c due to an almost singularity because of DCT[1]. for final test we select 3 time-bin variant for 120 MSps and 2 time-bin variants for 160 and 200 MSps, respectively.

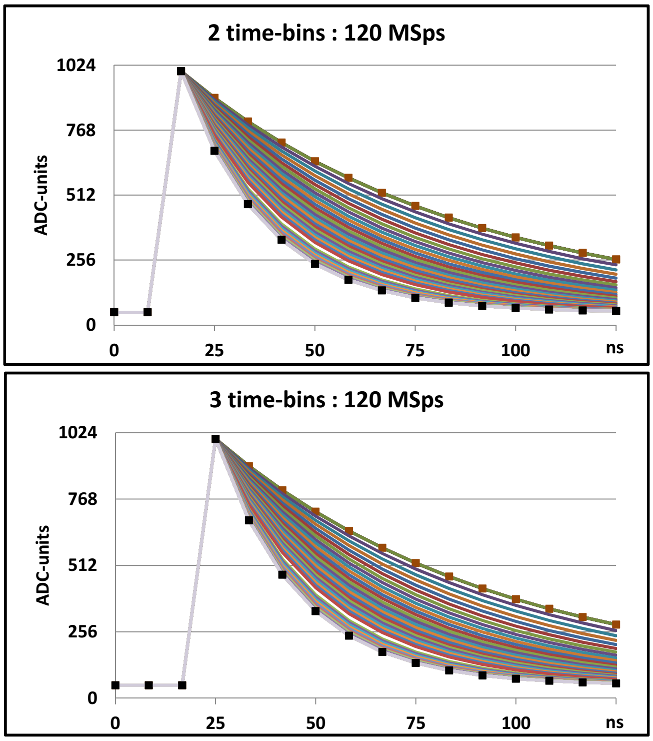


Fig. 12. Shapes of signal waveforms (for 120 MSps) giving the DCT trigger. The upper graph corresponds to an analysis of a signal waveform, where a jump appears on the $2^{nd}$ time-bin, the lower one for the $3^{rd}$ one.

DCT coefficients are preliminary calculated in an external C program for various $\alpha$ slopes. They are stored in the FPGA ROM and multiplexed in real time to keep arbitrary selected DCT rate (~6 Hz). In order to provide a sufficient calculation precision with a reasonably resources occupancy the DCT coefficients for ROM are selected as 12-bit fixed-point data.

For 160 and 200 MSps the DCT[8]/DCT[1] is almost independent of the tail slopes and are almost zeroed. We will neglect this coefficients as almost insensitive on signal waveform shapes.

The AHDL code for the FPGA contains a sigma-delta algorithm multiplexing several variants of bordering slopes to keep the stable DCT trigger rate in huge daily temperature variation [14]. This algorithm has been already successfully tested in Nov. 2014

## VII. CONCLUSIONS

We planned to deploy 7 Front-End Boards in a hexagon of the surface detectors in the Pierre Auger Engineering Array. As we see from Fig. 13 and Fig. 14 a configuration with 4 time-bins on a pedestal level is not suitable for a detection. A scaling fails due to a small DCT[1] coefficient for 15th k index. The variant with 3 time bins seems to be more appropriate, although 2 time bin variant also was tested in 2014 and passed successfully tests on the FEB with Cyclone III.

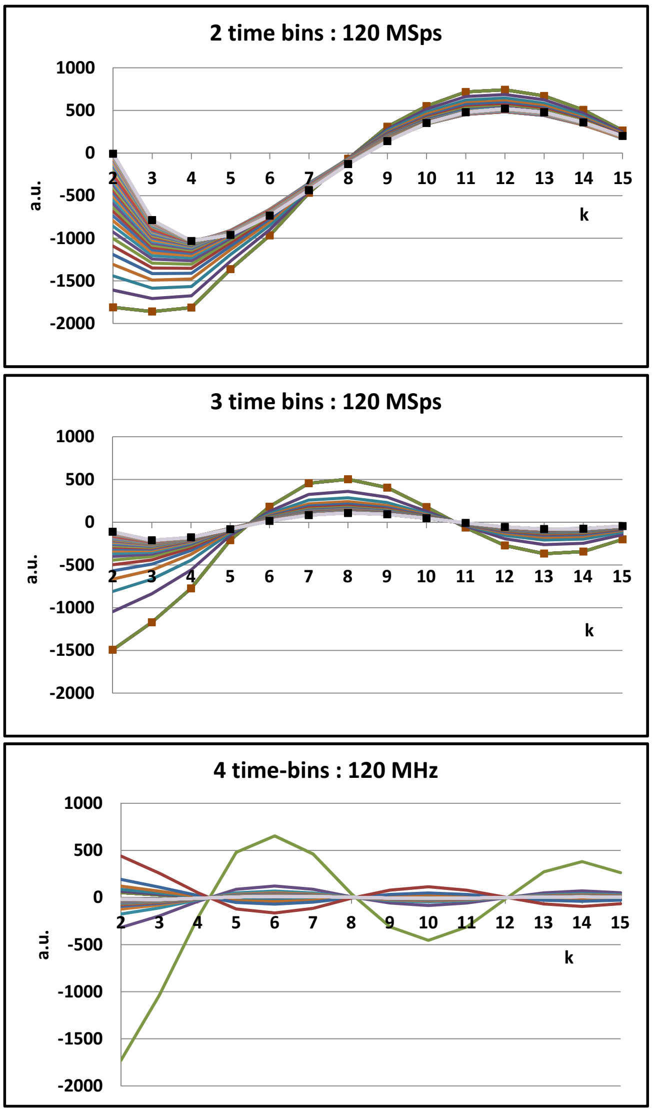


Fig. 13. DCT coefficients (for 120 MSps) giving the DCT trigger. The upper graph corresponds to an analysis of a signal waveform, where a jump appears on the $2^{nd}$ time-bin, the middle one for the $3^{rd}$ one and the lower graph for the $4^{th}$ one.

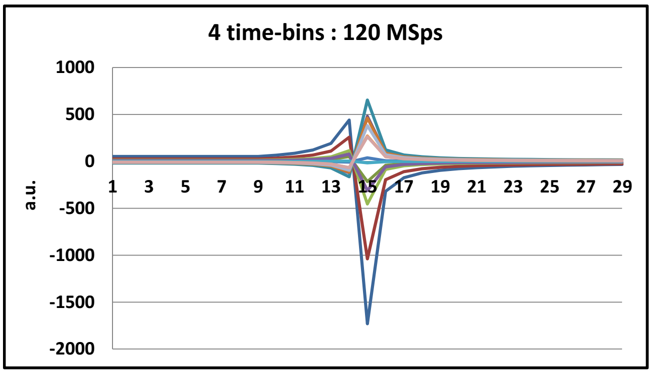


Fig. 14. DCT coefficients (for 120 MSps) vs. index of the tail slope. For k=15 the DCT[1] is close to zero (as denominator in a scaling) and DCT[k]/DCT[1] become huge values in comparison the other scaled DCT cofficients.

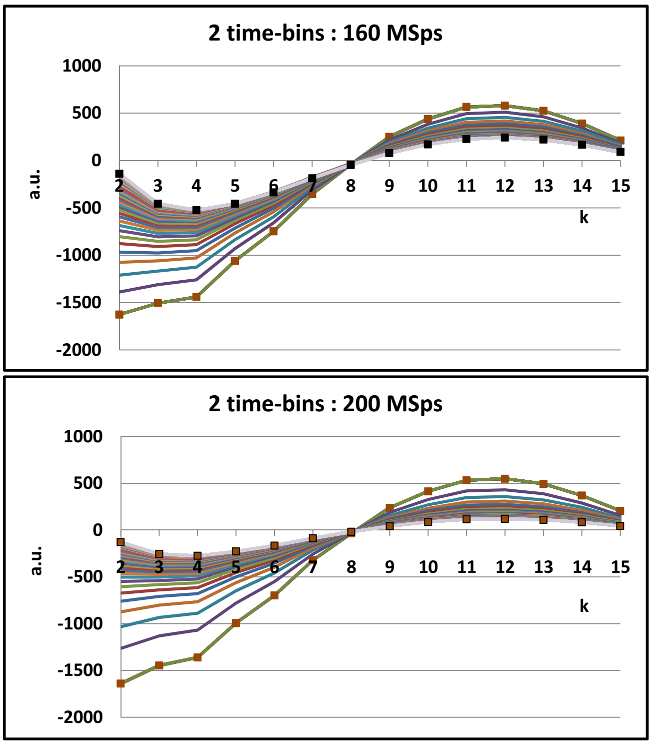


Fig. 15. DCT coefficients for 160 and 200 MSps giving the DCT trigger.

## Acknowledgment

This work was supported by the National Science Centre (Poland) under NCN Grant No. 2013/08/M/ST9/00322

## References

[1] J. Abraham et al., [Pierre Auger Collaboration], "Properties and Performance of the Prototype Instrument for the Pierre Auger Observatory", *Nucl. Instr. Meth.*, ser. A, vol. 523, pp. 50-95, May 2004.
DOI: 10.1016/j.nima.2003.12.012

[2] P. Abreu et al.,[ Pierre Auger Collaboration], "The exposure of the hybrid detector of the Pierre Auger Observatory", *Astroparticle Phys.* vol. 34 Issue: 6 pp. 368-381, Jan. 2011.
DOI: 10.1016/j.astropartphys.2010.10.001

[3] M. Settimo et al.,[ Pierre Auger Collaboration], "Measurement of the cosmic ray energy spectrum using hybrid events of the Pierre Auger Observatory", *European Phys. Journal Plus* vol. 127 Issue: 8 Article Number: 87, 15 pages, Aug 2012.
DOI: 10.1140/epjp/i2012-12087-9

[4] J. Abraham et al., [Pierre Auger Collaboration], "Trigger and aperture of the surface detector array of the Pierre Auger Observatory", *Nucl. Instr. Meth.*, ser. A, vol. 613, pp. 29-39, Jan. 2010.
DOI: 10.1016/j.nima.2009.11.018

[5] P. Abreu et al.,[ Pierre Auger Collaboration], "Search for ultrahigh energy neutrinos in highly inclined events at the Pierre Auger Observatory", *Phys. Rev. D* vol. 84 Issue: 12, No: 122005, Dec. 2011.
DOI: 10.1103/PhysRevD.84.122005

[6] P. Abreu et al.,[ Pierre Auger Collaboration], "Search for point-like sources of ultra-high energy neutrinos at the Pierre Auger Observatory and improved limit on the diffuse flux of tau neutrinos", *Astrophys. Journal Lett.* vol. 755 Issue: 1 Article Number: L4 7 pages, Aug. 2012.
DOI: 10.1088/2041-8205/755/1/L4

[7] Z. Szadkowski, K. Pytel, *Artificial neural network as a FPGA trigger for a detection of very inclined "young" showers*, IEEE Trans. on Nucl.

Science, vol. 63, Issue 3, pp. 1002-1009, June 2015.
DOI: 10.1109/TNS.2015.2421412

[8] Z. Szadkowski, "A spectral $1^{st}$ level FPGA trigger for detection of very inclined showers based on a 16-point Discrete Cosine Transform for the Pierre Auger Observatory", *Nucl. Instr. Meth.*, ser. A, vol. 606, pp. 330-343, July 2009.
DOI: 10.1016/j.nima.2009.03.255

[9] Z. Szadkowski, "Trigger Board for the Auger Surface Detector with 100 MHz Sampling and Discrete Cosine Transform", *IEEE Trans. on Nucl. Science*, vol. 58, pp. 1692-1700, Aug. 2011.
DOI: 10.1109/TNS.2011.2115252

[10] Z. Szadkowski "An optimization of 16-point Discrete Cosine Transform Implemented into a FPGA as a Design for a Spectral First Level Surface Detector Trigger in Extensive Air Shower Experiments", (2011) *Applications of Digital Signal Processing*, InTech, ISBN 978-953-307-406-1, Croatia.

[11] Z. Szadkowski, "Optimization of the detection of very inclined showers using a spectral DCT trigger in arrays of surface detectors", *IEEE Trans. on Nucl. Science*, vol. 60, no 5, pp. 3647-3653, Oct. 2013.
DOI: 10.1109/TNS.2013.2280639

[12] Z. Szadkowski, "The concept of an ACEX® cost-effective first level surface detector trigger in the Pierre Auger Observatory", *Nucl. Instr. Meth.*, ser. A, vol 551, pp. 477-486, Oct. 2005.
DOI: 10.1016/j.nima.2005.06.049

[13] Z. Szadkowski, K.-H. Becker, K.-H. Kampert, "Development of a new first level trigger for surface array in the Pierre Auger Observatory based on the Cyclone™ Altera® FPGA", *Nucl. Instr. Meth.*, ser. A, vol. 545, pp. 793-802, June 2005.
DOI: 10.1016/j.nima.2005.03.118

[14] J. Abraham et al., [Pierre Auger Collaboration], "Atmospheric effects on extensive air showers observed with the surface detector of the Pierre Auger Observatory", *Astroparticle Phys.* vol. 32 Issue: 2 pp. 89-99, Sept. 2009.
DOI: 10.1016/j.astropartphys.2009.06.004

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 709–716

DOI: 10.15439/2015F41
ACSIS, Vol. 5

# A hypothetical way to compute an upper bound for the heights of solutions of a Diophantine equation with a finite number of solutions

Apoloniusz Tyszka
University of Agriculture
Faculty of Production and Power Engineering
Balicka 116B, 30-149 Kraków, Poland
Email: rttyszka@cyf-kr.edu.pl

***Abstract*—Let**

$$f(n) = \begin{cases} 1 & \text{if} \quad n = 1 \\ 2^{2^{n-2}} & \text{if} \quad n \in \{2,3,4,5\} \\ \left(2+2^{2^{n-4}}\right)^{2^{n-4}} & \text{if} \quad n \in \{6,7,8,\dots\} \end{cases}$$

**We conjecture that if a system**

$$T \subseteq \{x_i + 1 = x_k,\ x_i \cdot x_j = x_k :\ i,j,k \in \{1,\dots,n\}\}$$

**has only finitely many solutions in positive integers $x_1,\dots,x_n$, then each such solution $(x_1,\dots,x_n)$ satisfies $x_1,\dots,x_n \leqslant f(n)$. We prove that the function $f$ cannot be decreased and the conjecture implies that there is an algorithm which takes as input a Diophantine equation, returns an integer, and this integer is greater than the heights of integer (non-negative integer, positive integer, rational) solutions, if the solution set is finite. We show that if the conjecture is true, then this can be partially confirmed by the execution of a brute-force algorithm.**

***Index Terms*—bound for integer solutions, Diophantine equation, finite-fold Diophantine representation, height of a solution, integer arithmetic.**

IN THIS article, we present a conjecture on integer arithmetic which implies a positive answer to all versions of the following open problem:

**Problem.** *Is there an algorithm which takes as input a Diophantine equation, returns an integer, and this integer is greater than the heights of integer (non-negative integer, positive integer, rational) solutions, if the solution set is finite?*

The height of a rational number $\frac{p}{q}$ is defined by $\max(|p|,|q|)$ provided $\frac{p}{q}$ is written in lowest terms. The height of a rational tuple $(x_1,\dots,x_n)$ is defined as the maximum of $n$ and the heights of the numbers $x_1,\dots,x_n$.

**Theorem 1.** *Only $x_1 = 1$ solves the equation $x_1 \cdot x_1 = x_1$ in positive integers. Only $x_1 = 1$ and $x_2 = 2$ solve the system $\{x_1 \cdot x_1 = x_1,\ x_1 + 1 = x_2\}$ in positive integers. For each integer $n \geqslant 3$, the following system*

$$\begin{cases} x_1 \cdot x_1 = x_1 \\ x_1 + 1 = x_2 \\ \forall i \in \{2,\dots,n-1\}\ x_i \cdot x_i = x_{i+1} \end{cases}$$

*has a unique solution in positive integers, namely $\left(1,2,4,16,256,\dots,2^{2^{n-3}},2^{2^{n-2}}\right)$.*

**Theorem 2.** *For each positive integer n, the following system*

$$\begin{cases} \forall i \in \{1,\dots,n\}\ x_i \cdot x_i = x_{i+1} \\ x_{n+2} + 1 = x_1 \\ x_{n+3} + 1 = x_{n+2} \\ x_{n+3} \cdot x_{n+4} = x_{n+1} \end{cases}$$

*is soluble in positive integers and has only finitely many integer solutions. Each integer solution $(x_1,\dots,x_{n+4})$ satisfies $|x_1|,\dots,|x_{n+4}| \leqslant \left(2+2^{2^n}\right)^{2^n}$. The maximal solution in positive integers is given by*

$$\begin{cases} \forall i \in \{1,\dots,n+1\}\ x_i = \left(2+2^{2^n}\right)^{2^{i-1}} \\ x_{n+2} = 1 + 2^{2^n} \\ x_{n+3} = 2^{2^n} \\ x_{n+4} = \left(1 + 2^{2^n} - 1\right)^{2^n} \end{cases}$$

*Proof.* The system equivalently expresses that $(x_1 - 2) \cdot x_{n+4} = x_1^{2^n}$. By this and the polynomial identity

$$x_1^{2^n} = 2^{2^n} + (x_1 - 2) \cdot \sum_{k=0}^{2^n-1} 2^{2^n-1-k} \cdot x_1^k$$

we get that $x_{n+3} = x_1 - 2$ divides $2^{2^n}$ and $x_{n+4} = \dfrac{x_1^{2^n}}{x_1 - 2}$. Hence, $x_1 \in \left[2 - 2^{2^n}, 2 + 2^{2^n}\right] \cap \mathbb{Z}$, the system has only finitely many integer solutions, and $|x_1|,\dots,|x_{n+4}| \leqslant \left(2+2^{2^n}\right)^{2^n}$. □

In [10, p. 719], the author proposed the upper bound $2^{2^{n-1}}$ for positive integer solutions to any system

$$T \subseteq \{x_i + x_j = x_k,\ x_i \cdot x_j = x_k :\ i,j,k \in \{1,\dots,n\}\}$$

978-83-60810-66-8/$25.00©2015, IEEE

which has only finitely many solutions in positive integers $x_1, \dots, x_n$. The bound $2^{2^{n-1}}$ is not correct for any $n \geqslant 8$ because the following system

$$\begin{cases} \forall i \in \{1,\dots,k\} \; x_i \cdot x_i &= x_{i+1} \\ x_{k+2} + x_{k+2} &= x_{k+3} \\ x_{k+2} \cdot x_{k+2} &= x_{k+3} \\ x_{k+4} + x_{k+3} &= x_1 \\ x_{k+4} \cdot x_{k+5} &= x_{k+1} \end{cases}$$

provides a counterexample for any $k \geqslant 3$. In [11, p. 96], the author proposed the upper bound $2^{2^{n-1}}$ for modulus of integer solutions to any system

$$T \subseteq \{x_k = 1, \;\; x_i + x_j = x_k, \;\; x_i \cdot x_j = x_k : \;\; i,j,k \in \{1,\dots,n\}\}$$

which has only finitely many solutions in integers $x_1, \dots, x_n$. The bound $2^{2^{n-1}}$ is not correct for any $n \geqslant 9$ because the following system

$$\begin{cases} \forall i \in \{1,\dots,k\} \; x_i \cdot x_i &= x_{i+1} \\ x_{k+2} &= 1 \\ x_{k+3} + x_{k+2} &= x_1 \\ x_{k+4} + x_{k+2} &= x_{k+3} \\ x_{k+4} \cdot x_{k+5} &= x_{k+1} \end{cases}$$

provides a counterexample for any $k \geqslant 4$. Let

$$f(n) = \begin{cases} 1 & \text{if} \quad n = 1 \\ 2^{2^{n-2}} & \text{if} \quad n \in \{2,3,4,5\} \\ \left(2 + 2^{2^{n-4}}\right)^{2^{n-4}} & \text{if} \quad n \in \{6,7,8,\dots\} \end{cases}$$

**Conjecture.** *If a system*

$$T \subseteq \{x_i + 1 = x_k, \;\; x_i \cdot x_j = x_k : \;\; i,j,k \in \{1,\dots,n\}\}$$

*has only finitely many solutions in positive integers $x_1, \dots, x_n$, then each such solution $(x_1, \dots, x_n)$ satisfies $x_1, \dots, x_n \leqslant f(n)$.*

Theorems 1 and 2 imply that the function $f$ cannot be decreased. Let $\mathcal{R}ng$ denote the class of all rings $\boldsymbol{K}$ that extend $\mathbb{Z}$, and let

$$E_n = \{x_k = 1, \;\; x_i + x_j = x_k, \;\; x_i \cdot x_j = x_k : \;\; i,j,k \in \{1,\dots,n\}\}$$

Th. Skolem proved that any Diophantine equation can be algorithmically transformed into an equivalent system of Diophantine equations of degree at most 2, see [6, pp. 2–3] and [2, pp. 3–4]. The following result strengthens Skolem's theorem.

**Lemma 1.** *([10, p. 720]) Let $D(x_1, \dots, x_p) \in \mathbb{Z}[x_1, \dots, x_p]$. Assume that* $\deg(D, x_i) \geqslant 1$ *for each $i \in \{1, \dots, p\}$. We can compute a positive integer $n > p$ and a system $T \subseteq E_n$ which satisfies the following two conditions:*

`Condition 1.` *If $\boldsymbol{K} \in \mathcal{R}ng \cup \{\mathbb{N}, \; \mathbb{N} \setminus \{0\}\}$, then*

$$\forall \tilde{x}_1, \dots, \tilde{x}_p \in \boldsymbol{K} \; \Big(D(\tilde{x}_1, \dots, \tilde{x}_p) = 0 \Longleftrightarrow$$

$$\exists \tilde{x}_{p+1}, \dots, \tilde{x}_n \in \boldsymbol{K} \; (\tilde{x}_1, \dots, \tilde{x}_p, \tilde{x}_{p+1}, \dots, \tilde{x}_n) \textit{ solves } T\Big)$$

`Condition 2.` *If $\boldsymbol{K} \in \mathcal{R}ng \cup \{\mathbb{N}, \; \mathbb{N} \setminus \{0\}\}$, then for each $\tilde{x}_1, \dots, \tilde{x}_p \in \boldsymbol{K}$ with $D(\tilde{x}_1, \dots, \tilde{x}_p) = 0$, there exists a unique tuple $(\tilde{x}_{p+1}, \dots, \tilde{x}_n) \in \boldsymbol{K}^{n-p}$ such that the tuple $(\tilde{x}_1, \dots, \tilde{x}_p, \tilde{x}_{p+1}, \dots, \tilde{x}_n)$ solves $T$.*

*Conditions 1 and 2 imply that for each $\boldsymbol{K} \in \mathcal{R}ng \cup \{\mathbb{N}, \; \mathbb{N} \setminus \{0\}\}$, the equation $D(x_1, \dots, x_p) = 0$ and the system $T$ have the same number of solutions in $\boldsymbol{K}$.*

For a positive integer $n$, let $S(n)$ denote the successor of $n$.

**Lemma 2.** *Let $T$ be a finite system of equations of the forms: $x = 1$, $x + y = z$, and $x \cdot y = z$. If the equation $x = 1$ belongs to $T$, then the system $T \cup \{x \cdot x = x\} \setminus \{x = 1\}$ has the same solutions in positive integers.*

**Lemma 3.** *Let $T$ be a finite system of equations of the forms: $S(x) = y$, $x + y = z$, and $x \cdot y = z$. If the equation $x + y = z$ belongs to $T$ and the variables $z_1$, $z_2$, $\widetilde{z_1}$, $\widetilde{z_2}$, $\widetilde{v}$, $u$, $t$, $\widetilde{t}$, $v$ are new, then the following system*

$$T \cup \{z \cdot x = z_1, \;\; z \cdot y = z_2, \;\; S(z_1) = \widetilde{z_1}, \;\; S(z_2) = \widetilde{z_2}, \;\; \widetilde{z_1} \cdot \widetilde{z_2} = \widetilde{v},$$

$$z \cdot z = u, \;\; x \cdot y = t, \;\; S(t) = \widetilde{t}, \;\; u \cdot \widetilde{t} = v, \;\; S(v) = \widetilde{v}\} \setminus \{x + y = z\}$$

*has the same solutions in positive integers and a smaller number of additions.*

*Proof.* According to [5, p. 100], for each positive integers $x, y, z$, $x + y = z$ if and only if

$$S(z \cdot x) \cdot S(z \cdot y) = S((z \cdot z) \cdot S(x \cdot y))$$

Indeed, the above equality is equivalent to

$$\left(z^2 \cdot x \cdot y + 1\right) + z \cdot (x + y) = \left(z^2 \cdot x \cdot y + 1\right) + z^2$$

□

Lemmas 1–3 imply the next theorem.

**Theorem 3.** *If we assume the Conjecture and a Diophantine equation $D(x_1, \dots, x_p) = 0$ has only finitely many solutions in positive integers, then an upper bound for these solutions can be computed.*

**Corollary 1.** *If we assume the Conjecture and a Diophantine equation $D(x_1, \dots, x_p) = 0$ has only finitely many solutions in non-negative integers, then an upper bound for these solutions can be computed by applying Theorem 3 to the equation $D(x_1 - 1, \dots, x_p - 1) = 0$.*

**Corollary 2.** *If we assume the Conjecture and a Diophantine equation $D(x_1, \dots, x_p) = 0$ has only finitely many integer solutions, then an upper bound for their modulus can be computed by applying Theorem 3 to the equation*

$$\prod_{(i_1, \dots, i_p) \in \{1,2\}^p} D((-1)^{i_1} \cdot (x_1 - 1), \dots, (-1)^{i_p} \cdot (x_p - 1)) = 0$$

**Lemma 4.** *([10, p. 720]) If there is a computable upper bound for the modulus of integer solutions to a Diophantine equation with a finite number of integer solutions, then there is a computable upper bound for the heights of rational solutions to a Diophantine equation with a finite number of rational solutions.*

**Theorem 4.** *The Conjecture implies that there is a computable upper bound for the heights of rational solutions to a Diophantine equation with a finite number of rational solutions.*

*Proof.* It follows from Corollary 2 and Lemma 4. □

The Davis-Putnam-Robinson-Matiyasevich theorem states that every recursively enumerable set $\mathcal{M} \subseteq \mathbb{N}^n$ has a Diophantine representation, that is

$$(a_1, \dots, a_n) \in \mathcal{M} \Longleftrightarrow$$
$$\exists x_1, \dots, x_m \in \mathbb{N} \quad W(a_1, \dots, a_n, x_1, \dots, x_m) = 0 \qquad \texttt{(R)}$$

for some polynomial $W$ with integer coefficients, see [2]. The polynomial $W$ can be computed, if we know the Turing machine $M$ such that, for all $(a_1, \dots, a_n) \in \mathbb{N}^n$, $M$ halts on $(a_1, \dots, a_n)$ if and only if $(a_1, \dots, a_n) \in \mathcal{M}$, see [2]. The representation (R) is said to be finite-fold, if for any $a_1, \dots, a_n \in \mathbb{N}$ the equation $W(a_1, \dots, a_n, x_1, \dots, x_m) = 0$ has only finitely many solutions $(x_1, \dots, x_m) \in \mathbb{N}^m$. Yu. Matiyasevich conjectures that each recursively enumerable set $\mathcal{M} \subseteq \mathbb{N}^n$ has a finite-fold Diophantine representation, see [1, pp. 341–342], [3, p. 42], and [4, p. 745]. Matiyasevich's conjecture implies a negative answer to the Problem, see [3, p. 42].

**Theorem 5.** *(cf. [10, p. 721]) The Conjecture implies that if a set $\mathcal{M} \subseteq \mathbb{N}$ has a finite-fold Diophantine representation, then $\mathcal{M}$ is computable.*

*Proof.* Let a set $\mathcal{M} \subseteq \mathbb{N}$ has a finite-fold Diophantine representation. It means that there exists a polynomial $W(x, x_1, \dots, x_m)$ with integer coefficients such that

$$\forall b \in \mathbb{N} \ \Big(b \in \mathcal{M} \Longleftrightarrow \exists x_1, \dots, x_m \in \mathbb{N} \quad W(b, x_1, \dots, x_m) = 0\Big)$$

and for any $b \in \mathbb{N}$ the equation $W(b, x_1, \dots, x_m) = 0$ has only finitely many solutions $(x_1, \dots, x_m) \in \mathbb{N}^m$. By Corollary 1, there is a computable function $g \colon \mathbb{N} \to \mathbb{N}$ such that for each $b, x_1, \dots, x_m \in \mathbb{N}$ the equality $W(b, x_1, \dots, x_m) = 0$ implies $\max(x_1, \dots, x_m) \leqslant g(b)$. Hence, we can decide whether or not a non-negative integer $b$ belongs to $\mathcal{M}$ by checking whether or not the equation $W(b, x_1, \dots, x_m) = 0$ has an integer solution in the box $[0, g(b)]^m$. □

In this paragraph, we follow [9] and we explain why Matiyasevich's conjecture although widely known is less widely accepted. Let us say that a set $\mathcal{M} \subseteq \mathbb{N}^n$ has a bounded Diophantine representation, if there exists a polynomial $W$ with integer coefficients such that

$$(a_1, \dots, a_n) \in \mathcal{M} \Longleftrightarrow \exists x_1, \dots, x_m \in \{0, \dots, \max(a_1, \dots, a_n)\}$$
$$W(a_1, \dots, a_n, x_1, \dots, x_m) = 0$$

Of course, any bounded Diophantine representation is finite-fold and any subset of $\mathbb{N}$ with a bounded Diophantine representation is computable. A simple diagonal argument shows that there exists a computable subset of $\mathbb{N}$ without any bounded Diophantine representation, see [1, p. 360]. The authors of [1] suggest a possibility (which contradicts Matiyasevich's conjecture) that each subset of $\mathbb{N}$ which has a finite-fold Diophantine representation has also a bounded Diophantine representation, see [1, p. 360].

For a positive integer $n$, let $\tau(n)$ denote the smallest positive integer $b$ such that for each system

$$T \subseteq \{x_i + 1 = x_k, \ x_i \cdot x_j = x_k : \ i, j, k \in \{1, \dots, n\}\}$$

with a finite number of solutions in positive integers $x_1, \dots, x_n$, all these solutions belong to $[1, b]^n$. By Theorems 1 and 2, $f(n) \leqslant \tau(n)$ for every positive integer $n$. The Conjecture implies that $f(n) = \tau(n)$ for every positive integer $n$.

**Theorem 6.** *(cf. [9, Theorem 4]) If a function $h \colon \mathbb{N} \setminus \{0\} \to \mathbb{N} \setminus \{0\}$ has a finite-fold Diophantine representation, then there exists a positive integer $m$ such that $h(n) < \tau(n)$ for every integer $n > m$.*

*Proof.* There exists a polynomial $W(x_1, x_2, x_3, \dots, x_r)$ with integer coefficients such that for each positive integers $x_1, x_2$,

$$(x_1, x_2) \in h \Longleftrightarrow$$
$$\exists x_3, \dots, x_r \in \mathbb{N} \setminus \{0\} \quad W(x_1, x_2, x_3 - 1, \dots, x_r - 1) = 0$$

and for each positive integers $x_1, x_2$ at most finitely many tuples $(x_3, \dots, x_r)$ of positive integers satisfy $W(x_1, x_2, x_3 - 1, \dots, x_r - 1) = 0$. By Lemmas 1–3, there is an integer $s \geqslant 3$ such that for any positive integers $x_1, x_2$,

$$(x_1, x_2) \in h \Longleftrightarrow$$
$$\exists x_3, \dots, x_s \in \mathbb{N} \setminus \{0\} \quad \Psi(x_1, x_2, x_3, \dots, x_s) \qquad \texttt{(E)}$$

where $\Psi(x_1, x_2, x_3, \dots, x_s)$ is a conjunction of formulae of the forms $x_i + 1 = x_k$ and $x_i \cdot x_j = x_k$, the indices $i, j, k$ belong to $\{1, \dots, s\}$, and for each positive integers $x_1, x_2$ at most finitely many tuples $(x_3, \dots, x_s)$ of positive integers satisfy $\Psi(x_1, x_2, x_3, \dots, x_s)$. Let $[\cdot]$ denote the integer part function, and let an integer $n$ is greater than $m = 2s + 2$. Then,

$$n \geqslant \left[\frac{n}{2}\right] + \frac{n}{2} > \left[\frac{n}{2}\right] + s + 1$$

and $n - \left[\frac{n}{2}\right] - s - 2 \geqslant 0$. Let $T_n$ denote the following system with $n$ variables:

$$\begin{cases} \multicolumn{4}{c}{\text{all equations occurring in } \Psi(x_1, x_2, x_3, \dots, x_s)} \\ \forall i \in \left\{1, \dots, n - \left[\frac{n}{2}\right] - s - 2\right\} & u_i \cdot u_i & = & u_i \\ & t_1 \cdot t_1 & = & t_1 \\ \forall i \in \left\{1, \dots, \left[\frac{n}{2}\right] - 1\right\} & t_i + 1 & = & t_{i+1} \\ & t_2 \cdot t_{\left[\frac{n}{2}\right]} & = & u \\ & u + 1 & = & x_1 \text{ (if } n \text{ is odd)} \\ & t_1 \cdot u & = & x_1 \text{ (if } n \text{ is even)} \\ & x_2 + 1 & = & y \end{cases}$$

By the equivalence (E), the system $T_n$ is soluble in positive integers, $2 \cdot \left[\frac{n}{2}\right] = u$, $n = x_1$, and

$$h(n) = h(x_1) = x_2 < x_2 + 1 = y$$

Since $T_n$ has at most finitely many solutions in positive integers, $y \leqslant \tau(n)$. Hence, $h(n) < \tau(n)$. □

Below is the excerpt from page 135 of the book [7]:

*Folklore. If a Diophantine equation has only finitely many solutions then those solutions are small in 'height' when compared to the parameters of the equation.*

*This folklore is, however, only widely believed because of the large amount of experimental evidence which now exists to support it.*

Below is the excerpt from page 12 of the article [8]:

*Note that if a Diophantine equation is solvable, then we can prove it, since we will eventually find a solution by searching through the countably many possibilities (but we do not know beforehand how far we have to search). So the really hard problem is to prove that there are no solutions when this is the case. A similar problem arises when there are finitely many solutions and we want to find them all. In this situation one expects the solutions to be fairly small. So usually it is not so hard to find all solutions; what is difficult is to show that there are no others.*

That is, mathematicians are intuitively persuaded that solutions are small when there are finitely many of them. It seems that there is a reason which is common to all the equations. Such a reason might be the Conjecture whose consequences we have already presented.

For a positive integer $b$, let $\Phi(b)$ denote the Conjecture restricted to systems whose all solutions in positive integers are not greater than $b$. Obviously,

$$\Phi(1) \Leftarrow \Phi(2) \Leftarrow \Phi(3) \Leftarrow \dots$$

and the Conjecture is equivalent to $\forall b \in \mathbb{N} \setminus \{0\}\ \Phi(b)$. The Conjecture is true for $n = 1$ and $n = 2$. Therefore, the sentence $\Phi(4)$ is true. For each positive integer $n$, there are only finitely many systems

$$T \subseteq \{x_i + 1 = x_k,\ x_i \cdot x_j = x_k :\ i, j, k \in \{1, \dots, n\}\}$$

Hence, for each positive integer $n$ there exists a positive integer $m$ such that the Conjecture restricted to systems with at most $n$ variables is equivalent to the sentence $\Phi(m)$.

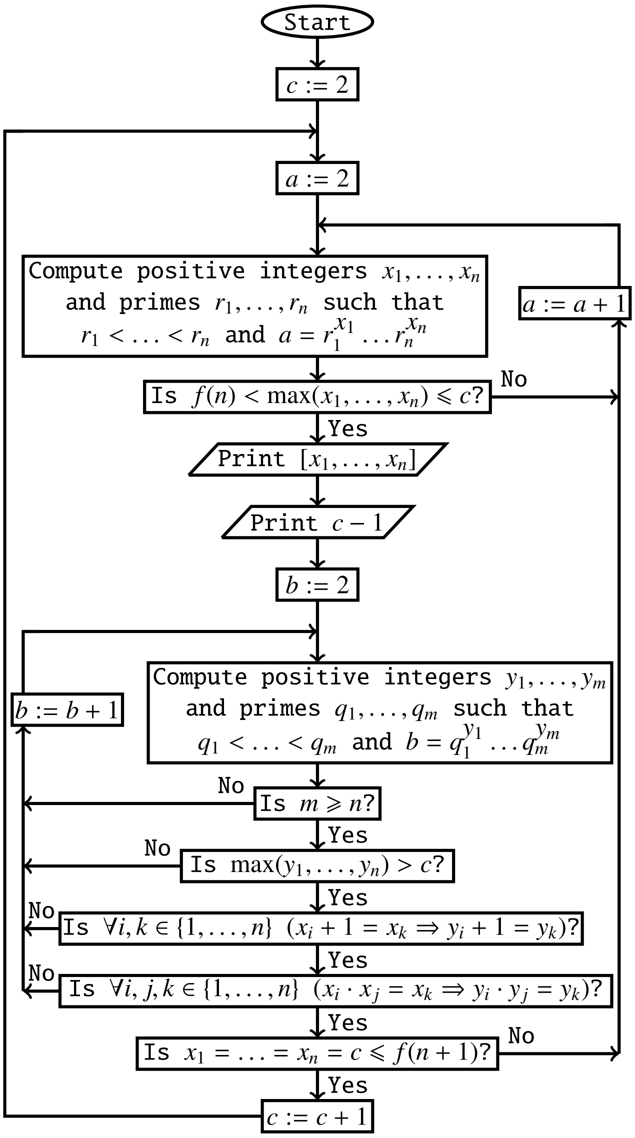
**Theorem 7.** *The Conjecture is equivalent to the following conjecture on integer arithmetic: if positive integers $x_1, \dots, x_n$ satisfy $\max(x_1, \dots, x_n) > f(n)$, then there exist positive integers $y_1, \dots, y_n$ such that*

$$\Big(\max(x_1, \dots, x_n) < \max(y_1, \dots, y_n)\Big) \wedge$$

$$\Big(\forall i, k \in \{1, \dots, n\}\ (x_i + 1 = x_k \Longrightarrow y_i + 1 = y_k)\Big) \wedge$$

$$\Big(\forall i, j, k \in \{1, \dots, n\}\ (x_i \cdot x_j = x_k \Longrightarrow y_i \cdot y_j = y_k)\Big)$$

The execution of the following flowchart never terminates.


Start
$c := 2$
$a := 2$
Compute positive integers $x_1, \dots, x_n$ and primes $r_1, \dots, r_n$ such that $r_1 < \dots < r_n$ and $a = r_1^{x_1} \dots r_n^{x_n}$
$a := a + 1$
Is $f(n) < \max(x_1, \dots, x_n) \leqslant c$?
No
Yes
Print $[x_1, \dots, x_n]$
Print $c - 1$
$b := 2$
Compute positive integers $y_1, \dots, y_m$ and primes $q_1, \dots, q_m$ such that $q_1 < \dots < q_m$ and $b = q_1^{y_1} \dots q_m^{y_m}$
$b := b + 1$
No
Is $m \geqslant n$?
Yes
No
Is $\max(y_1, \dots, y_n) > c$?
Yes
No
Is $\forall i, k \in \{1, \dots, n\}$ $(x_i + 1 = x_k \Rightarrow y_i + 1 = y_k)$?
Yes
No
Is $\forall i, j, k \in \{1, \dots, n\}$ $(x_i \cdot x_j = x_k \Rightarrow y_i \cdot y_j = y_k)$?
Yes
Is $x_1 = \dots = x_n = c \leqslant f(n + 1)$?
No
Yes
$c := c + 1$

**Theorem 8.** *If the Conjecture is true, then the execution of the flowchart provides an infinite sequence $X_1, c_1, X_2, c_2, X_3, c_3, \dots$ where $\{c_1, c_2, c_3, \dots\} = \mathbb{N} \setminus \{0\}$, $c_1 \leqslant c_2 \leqslant c_3 \leqslant \dots$ and each $X_i$ is a tuple of positive integers. Each returned number $c_i$ indicates that the performed computations confirm the sentence $\Phi(c_i)$. If the Conjecture is false, then the execution provides a similar finite sequence $X_1, c_1, \dots, X_k, c_k$ on the output. In this case, for the tuple $X_k = (x_1, \dots, x_n)$ an appropriate tuple $(y_1, \dots, y_n)$ does not exist, $\{c_1, \dots, c_k\} = [1, c_k] \cap \mathbb{N}$, $c_1 \leqslant \dots \leqslant c_k$, the sentences $\Phi(1), \Phi(2), \Phi(3), \dots, \Phi(c_k)$ are true, and the sentences $\Phi(c_k + 1), \Phi(c_k + 2), \Phi(c_k + 3), \dots$ are false.*

*Proof.* Let $p_n$ denote the $n^{\text{th}}$ prime number ($p_1 = 2$, $p_2 = 3$, etc.), and let $c$ stands for any integer greater than 1. The function $f$ is strictly increasing and there exists the smallest positive integer $n$ such that $c \leqslant f(n+1)$. Hence, if positive integers $x_1, \dots, x_i$ satisfy $f(i) < \max(x_1, \dots, x_i) \leqslant c$, then $i \leqslant n$ and $2 \leqslant p_1^{x_1} \dots p_i^{x_i} \leqslant p_1^c \dots p_n^c$. Therefore, if the sentence $\Phi(c)$ is true, then the flowchart algorithm checks all tuples of positive integers needed to confirm the sentence $\Phi(c)$. □

The following *MuPAD* code implements a simplified flowchart's algorithm which checks the following conjunction

$$\Big(m \geqslant n\Big) \wedge \Big(\max(y_1, \dots, y_n) > c\Big) \wedge$$

$$\Big(\forall i,k \in \{1,\dots,n\}\ (x_i + 1 = x_k \Longrightarrow y_i + 1 = y_k)\Big) \wedge$$

$$\Big(\forall i,j,k \in \{1,\dots,n\}\ (x_i \cdot x_j = x_k \Longrightarrow y_i \cdot y_j = y_k)\Big)$$

instead of four separate conditions.

```
c:=2:
while TRUE do
a:=2:
repeat
S:=op(ifactor(a)):
n:=(nops(S)-1)/2:
u:=min(S[2*i+1] $i=1..n):
v:=max(S[2*i+1] $i=1..n):
X:=[S[2*i+1] $i=1..n]:
if n=1 then f:=1 end_if:
if n>1 then f:=2^(2^(n-2)) end_if:
if n>5 then f:=(2+2^(2^(n-4)))^(2^(n-4))
end_if:
g:=2^(2^(n-1)):
if n>4 then g:=(2+2^(2^(n-3)))^(2^(n-3))
end_if:
if f<v and v<=c then
print(X):
print(c-1):
b:=2:
repeat
T:=op(ifactor(b)):
m:=(nops(T)-1)/2:
Y:=[T[2*i+1] $i=1..m]:
r:=min(m-n+1,max(Y[i] $i=1..m)-c):
for i from 1 to min(n,m) do
for j from 1 to min(n,m) do
for k from 1 to min(n,m) do
if X[i]+1=X[k] and Y[i]+1<>Y[k] then
r:=0 end_if:
if X[i]*X[j]=X[k] and Y[i]*Y[j]<>Y[k] then
r:=0 end_if:
end_for:
end_for:
end_for:
b:=b+1:
until r>0 end_repeat:
end_if:
a:=a+1:
until c=u and c=v and c<=g end_repeat:
c:=c+1:
end_while:
```

We attempt to confirm the sentence $\Phi(256)$. Since the execution of the flowchart algorithm (or its any variant) proceeds slowly, we must confirm the sentence $\Phi(256)$ in a different way. For integers $a_1, \dots, a_n$, let $P(a_1, \dots, a_n)$ denote the following system of equations:

$$\begin{cases} x_i + 1 &= x_k \quad (\text{if } a_i + 1 = a_k) \\ x_i \cdot x_j &= x_k \quad (\text{if } a_i \cdot a_j = a_k) \end{cases}$$

**Lemma 5.** *For each positive integer n, there exist positive integers $a_1, \dots, a_n$ such that $a_1 \leqslant \dots \leqslant a_n = \tau(n)$ and the system $P(a_1, \dots, a_n)$ has only finitely many solutions in positive integers. Each such numbers $a_1, \dots, a_n$ satisfy $a_1 < \dots < a_n$.*

*Proof.* If $a_1 < \dots < a_n$ does not hold, then we remove the first duplicate and insert $a_n + 1$ after $a_n$. Since $a_n + 1 > a_n = \tau(n)$, we get a contradiction. □

Let $\mathcal{F}$ denote the family of all systems $P(a_1, a_2, a_3)$, where integers $a_1, a_2, a_3$ satisfy $1 < a_1 < a_2 < a_3$.

**Theorem 9.** *The Conjecture is true for $n = 3$.*

*Proof.* By Lemma 5, there exist positive integers $a_1$, $a_2$, $a_3$ such that $a_1 < a_2 < a_3 = \tau(3)$ and the system $P(a_1, a_2, a_3)$ has only finitely many solutions in positive integers. If $a_1 = 1$, then $a_2 = 2$ and $a_3 \in \{3, 4\}$. Let $a_1 > 1$. Since $a_1 < a_2 < a_3$, we get $a_1 \cdot a_1 < a_1 \cdot a_2 < a_2 \cdot a_2$. Hence,

$$\operatorname{card}\Big(P(a_1, a_2, a_3) \cap \Big\{x_1 \cdot x_1 = x_3,\ x_1 \cdot x_2 = x_3,\ x_2 \cdot x_2 = x_3\Big\}\Big) \leqslant 1$$

Each integer $a_1$ satisfies $a_1 + 1 \neq a_1 \cdot a_1$. Hence,

$$\operatorname{card}\Big(P(a_1, a_2, a_3) \cap \Big\{x_1 + 1 = x_2,\ x_1 \cdot x_1 = x_2\Big\}\Big) \leqslant 1$$

Since $a_1 < a_2 < a_3$, the equation $x_1 + 1 = x_3$ does not belong to $P(a_1, a_2, a_3)$. By these observations, the following table shows all solutions in positive integers to any system that belongs to $\mathcal{F}$.

| | $\emptyset$ | $\{x_1{\cdot}x_1{=}x_3\}$ | $\{x_1{\cdot}x_2{=}x_3\}$ | $\{x_2{\cdot}x_2{=}x_3\}$ |
|---|---|---|---|---|
| $\emptyset \cup$ | any triple $(s,t,u)$ solves this system | any triple $\left(s,t,s^2\right)$ solves this system | any triple $(s,t,s{\cdot}t)$ solves this system | any triple $\left(s,t,t^2\right)$ solves this system |
| $\{x_1{+}1{=}x_2\} \cup$ | any triple $(s,s{+}1,u)$ solves this system | any triple $\left(s,s{+}1,s^2\right)$ solves this system | any triple $(s,s{+}1,s{\cdot}(s{+}1))$ solves this system | any triple $\left(s,s{+}1,(s{+}1)^2\right)$ solves this system |
| $\{x_1{\cdot}x_1{=}x_2\} \cup$ | any triple $\left(s,s^2,u\right)$ solves this system | $\notin \mathcal{F}$ | any triple $\left(s,s^2,s^3\right)$ solves this system | any triple $\left(s,s^2,s^4\right)$ solves this system |
| $\{x_2{+}1{=}x_3\} \cup$ | any triple $(s,t,t{+}1)$ solves this system | any triple $\left(s,s^2{-}1,s^2\right)$ solves this system | $\notin \mathcal{F}$ | $\notin \mathcal{F}$ |
| $\{x_1{+}1{=}x_2,$ $x_2{+}1{=}x_3\} \cup$ | any triple $(s,s{+}1,s{+}2)$ solves this system | only the triple $(2,3,4)$ solves this system | $\notin \mathcal{F}$ | $\notin \mathcal{F}$ |
| $\{x_1{\cdot}x_1{=}x_2,$ $x_2{+}1{=}x_3\} \cup$ | any triple $\left(s,s^2,s^2{+}1\right)$ solves this system | $\notin \mathcal{F}$ | $\notin \mathcal{F}$ | $\notin \mathcal{F}$ |

The table indicates that the system

$$\left\{x_1 + 1 = x_2,\ x_2 + 1 = x_3,\ x_1 \cdot x_1 = x_3\right\} =$$

$$\left\{x_1 + 1 = x_2,\ x_2 + 1 = x_3\right\} \cup \left\{x_1 \cdot x_1 = x_3\right\}$$

has a unique solution in positive integers, namely $(2,3,4)$. The other presented systems do not belong to $\mathcal{F}$ or have infinitely many solutions in positive integers. □

**Corollary 3.** *Since the Conjecture is true for $n \in \{1,2,3\}$, the sentence $\Phi(16)$ is true.*

**Theorem 10.** *The sentence $\Phi(256)$ is true.*

*Proof.* By Corollary 3, it suffices to consider quadruples of positive integers. The next *MuPAD* code returns 63 quadruples $(a_i,b_i,c_i,d_i)$ of positive integers, where $a_i < b_i < c_i < d_i \leqslant 256$ and $\max(a_i,b_i,c_i,d_i) = d_i > 16$. These quadruples have the following property: if positive integers $a,b,c,d$ satisfy $a < b < c < d \leqslant 256$ and $\max(a,b,c,d) = d > 16$, then there exists $i \in \{1,\dots,63\}$ such that $P(a,b,c,d) \subseteq P(a_i,b_i,c_i,d_i)$.

```
TEXTWIDTH:=60:
S:={}:
G:=[]:
T:={}:
H:=[]:
for a from 1 to 256 do
for b from 1 to 256 do
for c from 1 to 256 do
Y:=[1,a+1,a*a,a*b]:
for l from 1 to 4 do
X:=sort([a,b,c,Y[l]]):
u:=nops({a,b,c,Y[l]}):
v:=max(a,b,c,Y[l]):
if u=4 and 16<v and v<257 then
M:={}:
for i from 1 to 4 do
for j from i to 4 do
for k from 1 to 4 do
if X[i]+1=X[k] then
M:=M union {[i,k]} end_if:
if X[i]*X[j]=X[k] then
M:=M union {[i,j,k]} end_if:
end_for:
end_for:
end_for:
d:=nops(S union {M})-nops(S):
if d=1 then
S:=S union {M}:
G:=append(G,M):
T:=T union {X}:
H:=append(H,X):
end_if:
end_if:
end_for:
end_for:
end_for:
end_for:
for w from 1 to nops(G) do
for z from 1 to nops(G) do
p:=nops(G[w] minus G[z]):
q:=nops(G[z] minus G[w]):
if p=0 and 0<q then T:=T minus {H[w]}
end_if:
end_for:
end_for:
print(T):
```

The next table displays the quadruples $[a_1,b_1,c_1,d_1],\dots,[a_{63},b_{63},c_{63},d_{63}]$ and shows that for each $i \in \{1,\dots,63\}$ the system $P(a_i,b_i,c_i,d_i)$ has infinitely many solutions in positive integers, which completes the proof by Lemma 5. □

| | | |
|---|---|---|
| $(1, 2, 3, t)$<br>[1, 2, 3, 17] | $(1, 2, 4, t)$<br>[1, 2, 4, 17] | $(2, 3, 4, t)$<br>[2, 3, 4, 17] |
| $\left(t, t{+}1, t\,(t{+}1)\,, t(t{+}1)^2\right)$<br>[2, 3, 6, 18] | $(1, 2, t, 2t)$<br>[1, 2, 9, 18] | $(1, t, t{+}1, t(t{+}1))$<br>[1, 4, 5, 20] |
| $\left(t, t^2, t^2{+}1, t^2\left(t^2{+}1\right)\right)$<br>[2, 4, 5, 20] | $\left(t, t{+}1, (t{+}1)^2, t(t{+}1)^2\right)$<br>[2, 3, 9, 18] | $(t, t{+}1, t{+}2, (t{+}1)(t{+}2))$<br>[3, 4, 5, 20] |
| $\left(1, 2, t, t^2\right)$<br>[1, 2, 5, 25] | $\left(1, t, t{+}1, (t{+}1)^2\right)$<br>[1, 4, 5, 25] | $(t, t{+}1, t{+}2, t(t{+}1))$<br>[4, 5, 6, 20] |
| $(1, 2, t, t{+}1)$<br>[1, 2, 16, 17] | $\left(t, t^2, t^2{+}1, \left(t^2{+}1\right)^2\right)$<br>[2, 4, 5, 25] | $\left(1, t, t{+}1, t^2\right)$<br>[1, 5, 6, 25] |
| $\left(t, t{+}1, t{+}2, (t{+}2)^2\right)$<br>[3, 4, 5, 25] | $\left(1, t, t^2, t^2{+}1\right)$<br>[1, 4, 16, 17] | $\left(t, t^2, t^4, t^4{+}1\right)$<br>[2, 4, 16, 17] |
| $(t, t{+}1, t{+}2, t(t{+}2))$<br>[4, 5, 6, 24] | $\left(1, t, t^2, t^3\right)$<br>[1, 3, 9, 27] | $\left(t, t{+}1, (t{+}1)^2, (t{+}1)^2{+}1\right)$<br>[3, 4, 16, 17] |
| $\left(t, t{+}1, t{+}2, (t{+}1)^2\right)$<br>[4, 5, 6, 25] | $\left(t, t{+}1, (t{+}1)^2, (t{+}1)^3\right)$<br>[2, 3, 9, 27] | $\left(t, t{+}1, t^2, t^2{+}1\right)$<br>[4, 5, 16, 17] |
| $\left(t, t{+}1, t^2, t^3\right)$<br>[3, 4, 9, 27] | $\left(t, t{+}1, t{+}2, t^2\right)$<br>[5, 6, 7, 25] | $\left(t, t^2{-}1, t^2, t\left(t^2{-}1\right)\right)$<br>[3, 8, 9, 24] |
| $\left(t, t{+}1, t^2, t(t{+}1)\right)$<br>[4, 5, 16, 20] | $\left(t, t^2, t^3, t^5\right)$<br>[2, 4, 8, 32] | $\left(t, t{+}1, t(t{+}1), t^2(t{+}1)^2\right)$<br>[2, 3, 6, 36] |
| $\left(t, t^2{-}1, t^2, t^3\right)$<br>[3, 8, 9, 27] | $(t, t{+}1, t(t{+}1){-}1, t(t{+}1))$<br>[4, 5, 19, 20] | $(1, t, t{+}1, t{+}2)$<br>[1, 15, 16, 17] |
| $\left(t, t^2, t^2{+}1, t^3\right)$<br>[3, 9, 10, 27] | $\left(t, t{+}1, t^2, (t{+}1)^2\right)$<br>[4, 5, 16, 25] | $(t, t{+}1, t(t{+}1), t(t{+}1){+}1)$<br>[4, 5, 20, 21] |
| $\left(t, t{+}1, t^2, (t{+}1)t^2\right)$<br>[3, 4, 9, 36] | $\left(t, t^2, t^2{+}1, t\left(t^2{+}1\right)\right)$<br>[3, 9, 10, 30] | $\left(t, t^2{-}1, t^2, t^2{+}1\right)$<br>[4, 15, 16, 17] |
| $\left(t, t^2, t^4, t^5\right)$<br>[2, 4, 16, 32] | $\left(t, t{+}1, t(t{+}1), (t{+}1)^2\right)$<br>[4, 5, 20, 25] | $\left(1, t, t^2{-}1, t^2\right)$<br>[1, 5, 24, 25] |
| $\left(t, t{+}1, t(t{+}1), t^2(t{+}1)\right)$<br>[3, 4, 12, 36] | $\left(t, t^2, t^2{+}1, t^2{+}2\right)$<br>[4, 16, 17, 18] | $\left(t, t{+}1, (t{+}1)^2{-}1, (t{+}1)^2\right)$<br>[4, 5, 24, 25] |
| $\left(t, t{+}1, t^2{-}1, t^2\right)$<br>[5, 6, 24, 25] | $(t, t{+}1, t{+}2, t{+}3)$<br>[14, 15, 16, 17] | $\left(t, t^2, t^3{-}1, t^3\right)$<br>[3, 9, 26, 27] |
| $\left(t, t^2, t^3, t^3{+}1\right)$<br>[3, 9, 27, 28] | $\left(t, t^2{-}2, t^2{-}1, t^2\right)$<br>[5, 23, 24, 25] | $\left(t, t^2, t^3, t^6\right)$<br>[2, 4, 8, 64] |
| $\left(t, t^2{-}1, t^2, \left(t^2{-}1\right)^2\right)$<br>[3, 8, 9, 64] | $\left(t, t^2, t^4, t^6\right)$<br>[2, 4, 16, 64] | $\left(t, t^2{-}1, t^2, \left(t^2{-}1\right)t^2\right)$<br>[3, 8, 9, 72] |
| $\left(t^2, t^3, t^4, t^6\right)$<br>[4, 8, 16, 64] | $\left(1, t, t^2, t^4\right)$<br>[1, 3, 9, 81] | $\left(t, t{+}1, (t{+}1)^2, (t{+}1)^4\right)$<br>[2, 3, 9, 81] |
| $\left(t, t{+}1, t^2, t^4\right)$<br>[3, 4, 9, 81] | $\left(t, t^2{-}1, t^2, t^4\right)$<br>[3, 8, 9, 81] | $\left(t, t^2, t^2{+}1, t^4\right)$<br>[3, 9, 10, 81] |
| $\left(t, t^2, t^3, t^4\right)$<br>[3, 9, 27, 81] | $\left(t, t^2, t^4{-}1, t^4\right)$<br>[3, 9, 80, 81] | $\left(t, t^2, t^4, t^8\right)$<br>[2, 4, 16, 256] |

Of course, the Conjecture restricted to integers $n \in \{1,2,3,4\}$ is intuitively obvious and implies Theorem 10. Formally, the Conjecture remains unproven for $n = 4$. We explain why a hypothetical brute-force proof of the Conjecture for $n = 4$ is much longer than the proof of Theorem 10. By Lemma 5, it suffices to consider only the systems $P(a,b,c,d)$, where positive integers $a,b,c,d$ satisfy $a < b < c < d$.

`Case 1:` $a = 1$. Obviously,

$$\mathrm{card}\Big(\Big\{x_1 + 1 = x_2\Big\} \cap P(a,b,c,d)\Big) \leqslant 1$$

and

$$\mathrm{card}\Big(\Big\{x_3 + 1 = x_4\Big\} \cap P(a,b,c,d)\Big) \leqslant 1$$

Since $b + 1 < b \cdot b$, we get

$$\mathrm{card}\Big(\Big\{x_2 + 1 = x_3,\ x_2 \cdot x_2 = x_3\Big\} \cap P(a,b,c,d)\Big) \leqslant 1$$

Since $b \cdot b < b \cdot c < c \cdot c$, we get

$$\mathrm{card}\Big(\Big\{x_2 \cdot x_2 = x_4,\ x_2 \cdot x_3 = x_4,\ x_3 \cdot x_3 = x_4\Big\} \cap P(a,b,c,d)\Big) \leqslant 1$$

The above inequalities allow one to determine $(1+1) \cdot (1+1) \cdot (2+1) \cdot (3+1) = 48$ systems which need to be solved.

`Case 2:` $a > 1$. Obviously,

$$\mathrm{card}\Big(\Big\{x_2 + 1 = x_3\Big\} \cap P(a,b,c,d)\Big) \leqslant 1$$

Since $a + 1 < a \cdot a$, we get

$$\mathrm{card}\Big(\Big\{x_1 + 1 = x_2,\ x_1 \cdot x_1 = x_2\Big\} \cap P(a,b,c,d)\Big) \leqslant 1$$

Since $c + 1 < a \cdot c$, we get

$$\mathrm{card}\Big(\Big\{x_3 + 1 = x_4,\ x_1 \cdot x_3 = x_4\Big\} \cap P(a,b,c,d)\Big) \leqslant 1$$

Since $a \cdot a < a \cdot b < b \cdot b$, we get

$$\mathrm{card}\Big(\Big\{x_1 \cdot x_1 = x_3,\ x_1 \cdot x_2 = x_3,\ x_2 \cdot x_2 = x_3\Big\} \cap P(a,b,c,d)\Big) \leqslant 1$$

Since $a \cdot a < a \cdot b < b \cdot b < b \cdot c < c \cdot c$, we get

$$\mathrm{card}\Big(\Big\{x_1 \cdot x_1 = x_4,\ x_1 \cdot x_2 = x_4,\ x_2 \cdot x_2 = x_4,$$
$$x_2 \cdot x_3 = x_4,\ x_3 \cdot x_3 = x_4\Big\} \cap P(a,b,c,d)\Big) \leqslant 1$$

The above inequalities allow one to determine $(1+1) \cdot (2+1) \cdot (2+1) \cdot (3+1) \cdot (5+1) = 432$ systems which need to be solved.

*MuPAD* is a computer algebra system whose syntax is modelled on *Pascal*. The commercial version of *MuPAD* is no longer available as a stand-alone product, but only as the *Symbolic Math Toolbox* of *MATLAB*. Fortunately, the presented codes can be executed by *MuPAD Light*, which was offered for free for research and education until autumn 2005.

## References

[1] M. Davis, Yu. Matiyasevich, J. Robinson, *Hilbert's tenth problem. Diophantine equations: positive aspects of a negative solution,* in: Mathematical developments arising from Hilbert problems (ed. F. E. Browder), Proc. Sympos. Pure Math., vol. 28, Part 2, Amer. Math. Soc., 1976, 323–378; reprinted in: The collected works of Julia Robinson (ed. S. Feferman), Amer. Math. Soc., 1996, 269–324.

[2] Yu. Matiyasevich, *Hilbert's tenth problem,* MIT Press, Cambridge, MA, 1993.

[3] Yu. Matiyasevich, *Hilbert's tenth problem: what was done and what is to be done.* Hilbert's tenth problem: relations with arithmetic and algebraic geometry (Ghent, 1999), 1–47, Contemp. Math. 270, Amer. Math. Soc., Providence, RI, 2000.

[4] Yu. Matiyasevich, *Towards finite-fold Diophantine representations,* J. Math. Sci. (N. Y.) vol. 171, no. 6, 2010, 745–752, http://dx.doi.org/10.1007%2Fs10958-010-0179-4.

[5] J. Robinson, *Definability and decision problems in arithmetic,* J. Symbolic Logic 14 (1949), 98–114; reprinted in: The collected works of Julia Robinson (ed. S. Feferman), Amer. Math. Soc., 1996, 7–23.

[6] Th. Skolem, *Diophantische Gleichungen*, Julius Springer, Berlin, 1938.

[7] N. P. Smart, *The algorithmic resolution of Diophantine equations,* Cambridge University Press, Cambridge, 1998, http://dx.doi.org/10.1017/CBO9781107359994.

[8] M. Stoll, *How to solve a Diophantine equation,* in: An invitation to mathematics: From competitions to research (eds. M. Lackmann and D. Schleicher), Springer, Berlin-Heidelberg-New York, 2011, 9–19, http://dx.doi.org/10.1007/978-3-642-19533-4_2.

[9] A. Tyszka, *All functions $g\colon \mathbb{N} \to \mathbb{N}$ which have a single-fold Diophantine representation are dominated by a limit-computable function $f\colon \mathbb{N} \setminus \{0\} \to \mathbb{N}$ which is implemented in MuPAD and whose computability is an open problem,* in: Computation, cryptography, and network security (eds. N. J. Daras and M. Th. Rassias), Springer, Berlin-Heidelberg-New York, 2015, 577–590, http://dx.doi.org/10.1007/978-3-319-18275-9_24.

[10] A. Tyszka, *Conjecturally computable functions which unconditionally do not have any finite-fold Diophantine representation,* Inform. Process. Lett. 113 (2013), no. 19–21, 719–722, http://dx.doi.org/10.1016/j.ipl.2013.07.004.

[11] A. Tyszka, *Does there exist an algorithm which to each Diophantine equation assigns an integer which is greater than the modulus of integer solutions, if these solutions form a finite set?* Fund. Inform. 125 (2013), no. 1, 95–99, http://dx.doi.org/10.3233/FI-2013-854.

DOI: 10.15439/2015F40
ACSIS, Vol. 5

# Image Processing and Analysis of Textile Fibers by Virtual Random Walk

Hafedh Zghidi,
Maksym Walczak
Silesian University of Technology,
Faculty of Automatic Control,
Electronics and Computer Science,
44-100 Poland
E-mail: hafed.zghidi@polsl.pl,
maksym.walczak@gmail.com

Tomasz Blachowicz,
Krzysztof Domino
Silesian University of Technology,
Institute of Physics – Center for
Science and Education, 44-100
Poland
E-mail:
tomasz.blachowicz@polsl.pl,
krzysztof.domino@polsl.pol

Andrea Ehrmann
Bielefeld University of Applied
Sciences, Faculty of Engineering
Sciences and Mathematics, 33609
Germany
E-Mail: andrea.ehrmann@fh-bielefeld.de

***Abstract*—An algorithm for extracting material shape and spatial information from non-uniform background, and for generating object skeletons for statistical two-dimensional experiments using random walk approach, is presented. This finds applications in textile analysis and microscopic analysis of various materials like hairs and allows for further precise determination of textile yarn dimensions as well as other geometrical characteristics like a fractal dimension.**

## I. Introduction

HEMP fibers can be used for a variety of textile products, e.g. clothing or home textiles, but also for technical purposes in fiber composites. For all applications, it is important to determine the fiber lengths and diameters (averaged along with a standard deviation), a task which can be handled by optical equipment and respective software (e.g. [1,2]). Another important factor, however, is the identification of the degree of bifurcation, which influences the spinning process, the fiber-matrix bond strength etc. For this property there is no standard measurement procedure available. Our article aims at presenting a novel method to define the bifurcation of fibers, based on the evaluation of microscopic fiber images based on random walk analysis. The hemp fibers investigated in our study are pre-treated by hammer milling or by combing. Afterwards, the fibers are separated chemically by alkaline cooking [3].

Photographic images of different magnifications were taken using a digital optical microscope VHX-600D by Keyence with an objective VH-Z20R. Captured images were then processed giving monochromatic representative maps. Next, random walk statistical experiments on these maps were carried out. Subsequent chapters describe all mentioned steps, before conclusions are finally provided.

## II. Image Processing

The algorithm described here consists of two denoising and four extraction steps. In the first step, the image is being smoothed using a median filter [4]. This step slightly reduces fine-grained noise and also improves the efficiency of filters used in the first extraction step of the algorithm. The next step is an unsharp masking where it is necessary to define the proportional factor k of the operation [5]. This step can be defined as follows:

$$A_{um}=A_{mf}-Gauss\left(A_{mf}\right)$$
$$A_{um}=k\cdot A_{um}+A_{mf} \quad , \qquad (1)$$

where $A_{um}$ is the unsharp masked image, $A_{mf}$ is the median filtered image, and *Gauss* is the Gaussian filter [4], for which we use the Gaussian filter kernel $k = 0.4$. The window size and the standard deviation parameter σ equal 3. The parameters were chosen experimentally.

The first extraction step is an algorithm that can be split into two functional parts: a background sampling, and a background removal. The algorithm removes the background using a special filter that uses a background sample. The filter window is defined as a vector $h=\{h_1, h_2, h_3, h_4\}$, where $h_1$ is the average intensity, $h_2$ is the standard deviation of intensity, $h_3$ is the maximum intensity, and $h_4$ the minimum intensity. Contrary to the results of Ref. 6, all the values from $h_1$ to $h_4$ are calculated from a Moore neighborhood of each pixel. This improves the filter's quality as consecutive neighborhoods partially overlap each other.

In the first functional part, the algorithm sums Euclidean distances from reference vector, created from upper left corner of the background sample ($h_{ref}$), to reference vectors of all the other pixels from the sample ($h_{ib}$). Next, the algorithm calculates the reference pattern $\bar{d}$ that can be expressed as follows

$$\bar{d}=\frac{\sum_{i=0}^{i=M\cdot N} d\left(h_{ref},h_{ib}\right)}{s} \quad , \qquad (2)$$

where $s$ is the experimentally chosen value, $M$ is the number of rows in the background sample, and $N$ is the number of columns in the background sample.

There are two options for choosing the $s$ value. The default value is $1/4\,MN$, which makes $\bar{d}$ an average of all the distances, taking into consideration the window size = 3.

978-83-60810-66-8/$25.00©2015, IEEE

The other option is to divide by an experimentally chosen value to either increase or decrease the algorithm's sensitivity. For the pictures presented in this article we chose the second option. Thus, it was necessary to divide the sum by a number smaller than $1/4MN$ . This choice increased the filter's sensitivity, since the background in our images was far less uniform than in the images in [6]. Once the $s$ value is properly chosen, the algorithm will give similarly good results for other images provided their background has the same texture and the lighting conditions are similar. This procedure will be proved in the next section of this paper.

In the second part of the algorithm, the distance between reference vectors, generated for each pixel of the image, is compared to the previously calculated reference pattern. The filter function is defined as

$$\begin{matrix} A_{dn}\ if\ \ d\left(h_{ref},h_i\right)\ge \bar{d} \\ 0\ if\ \ d\left(h_{ref},h_i\right)< \bar{d} \end{matrix}\ , \qquad (3)$$

where $h_{ref}$ is the reference vector created from the upper left corner of the sampled region, $h_i$ is the $i$-th region of the whole image, and $A_{dn}$ is the denoised image. When the distance is greater than the reference pattern, we leave the processed image untouched, otherwise we remove the currently sampled pixel. Eventually any objects in the image are separated from the background.

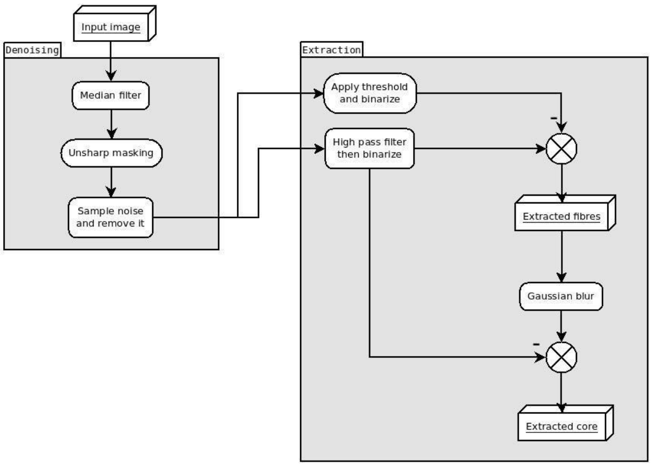


Fig. 1 Pipeline of the algorithm

The algorithm requires several parameters to be specified in order to process input images with expected quality. These parameters are: the background sample, the extraction threshold, the sensitivity, the median filter size, the unsharp-masking proportional-factor, the Gaussian filter window size, and the Gaussian and high-pass filter window-size. For the purpose of this paper, however, we analyzed all the parameters and chose to present only those which have the greatest impact on image quality. The results comparing influence of parameters for the same input image are presented in Fig. 2. In Fig. 3 only input images and final results needed for random walk experiments are depicted.

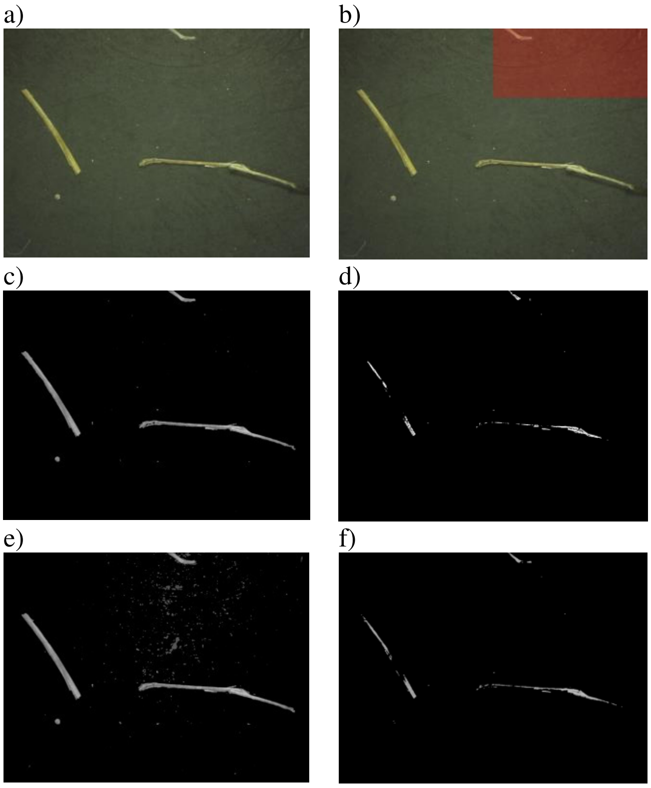


Fig. 2 A first exemplary set with optimally chosen parameters (segmentation threshold = 100, $s$ = 0.06 · width · height): input image for processing (a), background sample (b), image separated from background (c), second result set for image separated from the background with too high sensitivity, the same background sample as above (segmentation threshold = 100, s = 0.05 · width · height) (d), third set of results for the image separated from the background with too low sensitivity, the same background sample as above (segmentation threshold = 100, $s$ = 0.07 · width · height) (e), a fourth set of results for the image separated from the background with bad noise sample (segmentation threshold = 100, $s$ = 0.06 · width · height) (f).

As it is seen in the examples, the presented parameters have very high impact on the image quality. Although the parameter $s$ needs to be chosen experimentally, different input images having similar background texture and lighting conditions are being processed properly. This proves the algorithm's applicability in microscopic photography and industrial photography, where different objects are photographed in similar, but not necessarily identical conditions.

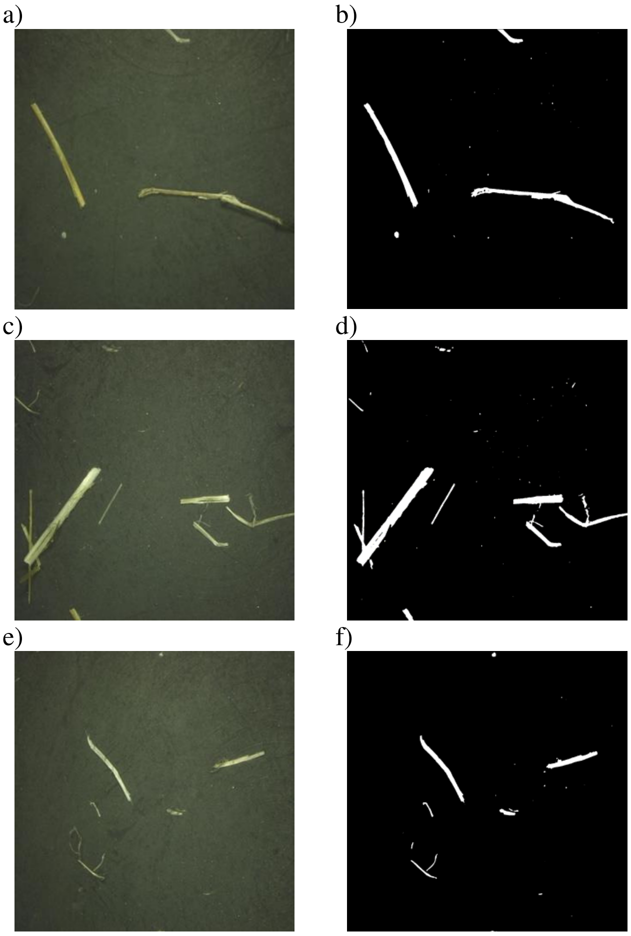


Fig. 3 Different examples of input images (a, c, e) and final results of monochrome maps needed for random walk experiments (b, d, f), respectively

## III. Random Walk Experiments

In this chapter the use of the image statistical analysis to examine the time series of the displacement of the random walker is discussed. Let us consider a uniform two-dimensional 'black' plane. A random walk can be applied and a walker is allowed to move randomly (with the same probability) up, down, to the left or to the right by a unit length in a unit time. This is the Markovian type random walk and the direction of the $n$-th step is chosen at random and does not depend on directions of previous ones. Here the mean square displacement of the probe $\left\langle R^2(n)\right\rangle$ is linearly dependent on the number of steps $n$ (time) which is proportional to the squared averaged random walk distance:

$$\left\langle R^2(n)\right\rangle \propto n. \tag{4}$$

Similarly, the pictures under investigation can be examined by means of the random walk procedure. At every iteration, a random direction is being chosen and then the algorithm proceeds one step into the chosen direction. However, the position at the $n$-th step depends on the previous positions due to the non-uniformity of the texture. Finally, the mean squared displacement is no longer the linear function of $n$ [7-9], but the following relation appears for large n:

$$\left\langle R^2(n)\right\rangle \propto n^{2H}, \tag{5}$$

where $H$ is the self-similarity coefficient called Hurst exponent [7,10]. From Hurst exponent distribution, unambiguous information about yarn hairness, and in general, information about yarn quality can be obtained. Such quantitative evaluations are on demand for textile industry. In practice, for a random walk on two-dimensional monochromatic pictures the typical result is $H < 1/2$; however, the information obtained for the exponent distribution is a unique information of a given picture (Fig. 4). Details about this type of analysis, applied for knitted fabrics, can be found in [11].

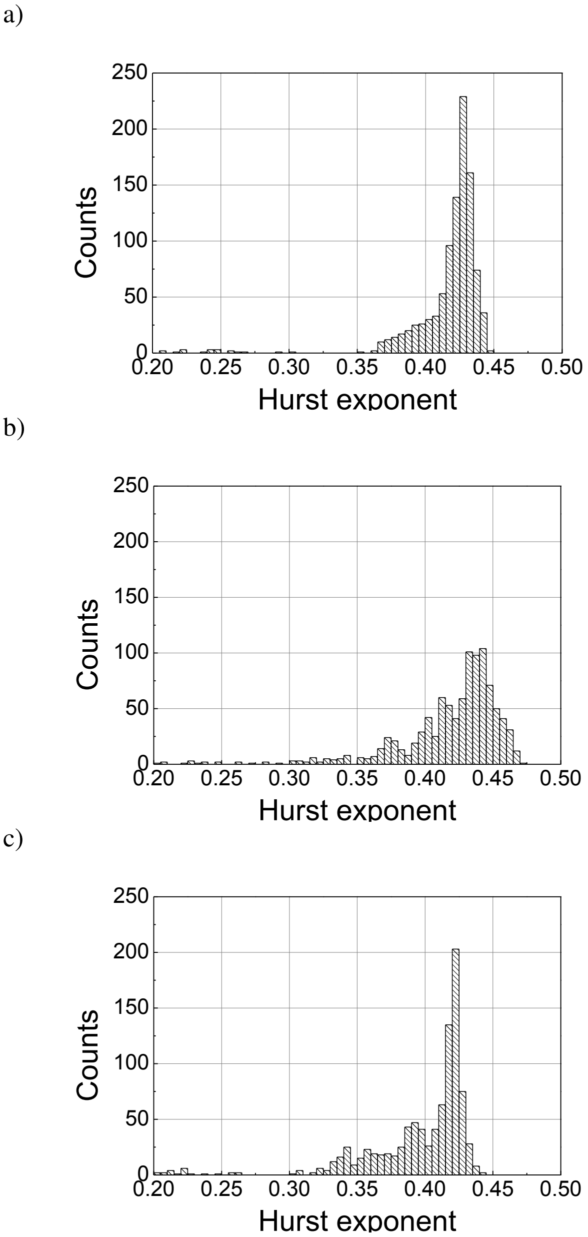


Fig. 4 Results of random walk experiments (a, b, c) carried out on the three monochromatic images from Fig. 3 (b, d, f), respectively

## IV. Conclusions

The paper gives an overview of the necessary steps used to create reliable monochromatic images from microscopic images by separating objects under examination from the background. This algorithm was applied on microscopic pictures of hemp fibers. Comparison between random walks performed on the monochromatic pictures gives a quantitative measure of differences between the hemp fibers, dependent on the fiber diameters and bifurcations which are important parameters in fiber manufacturing processes such as spinning or composite production. Thus, the image processing algorithm presented here is a fundamental requirement in the optical evaluation of fiber quality using the self-similarity coefficient.

## References

[1] H. L. Bos, J. Müssig, and M. J. A. van den Oever, "Mechanical properties of short-flax-fibre reinforced compounds, Composites Part A," *Applied Science and Manufacturing* 37, pp. 1591-1604, Oct. 2006. DOI:10.1016/j.compositesa.2005.1-0.011

[2] J. L. Thomason, J. Carruthers, J. Kelly, and G. Johnson, "Fibre cross-section determination and variability in sisal and flax and its effects on fibre performance characterization," *Composites Science and Technology* 71, pp. 1008-1015, May 2011. DOI:10.1016/j.compscitech.2011.03.007

[3] A. Gutiérrez and J. C. del Rio, "Lipids from Flax Fibers and Their Fate in Alkaline Pulping," *J. Agric. Food Chem.* 51, pp. 4965-4971, Jul. 2003. DOI: 10.1021/jf034370t

[4] Mark Nixon, "Feature Extraction & Image Processing for Computer Vision", *Academic Press*, Third edition p. 109, 2012

[5] Mark Nixon, "Feature Extraction & Image Processing for Computer Vision", *Academic Press*, Third edition p. 114, 2012

[6] A. Fabiańska, and L. Jackowska-Strumiłło, "Image processing and analysis algorithms for yarn hairiness determination," *Machine Vision and Applications* 23, pp. 527-540, Feb. 2012. DOI: 10.1007/s00138-012-0411-y

[7] S. Havlin, and D. Ben-Avraham, "Diffusion in disordered media," *Advances in Physics*, 51, pp. 187-292, Jan. 2002. DOI: 10.1080/00018730110116353

[8] J. W. Haus, and K. W. Kehr, "Diffusion in Regular and Disordered Lattices," *Phys. Rep.* 150 , pp. 263-406, Jun. 1987. DOI: 10.1016/0370-1573(87)90005-6

[9] D. Ben-Avraham, and S. Havlin, Difffusion and Reactions in Fractals and Disordered Systems, Cambridge Univ. Press, Cambridge 2000.

[10] S. Alexander, and S. R. Orbach, "Density of states on fractals: fractons," *J. Phys. Lett., Paris*, 43, pp. 625-631, Sept. 1982. DOI: 10.1051/jphyslet:019820043017062500

[11] A. Ehrmann, T. Blachowicz, K. Domino, S. Aumann, M. O. Weber, and H. Zghidi, "Examination of hairiness changes due to washing in knitted fabrics using a random walk approach," *Textile Research Journal*, online first, April 16, 2015. DOI: 10.1177/0040517515581591

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 721–728

DOI: 10.15439/2015F86
ACSIS, Vol. 5

# Reproducible floating-point atomic addition in data-parallel environment

David Defour
Laboratoire DALI-LIRMM
52 avenue Paul Alduy
66860 Perpignan Cerdex - France
Email: david.defour@univ-perp.fr

Sylvain Collange
INRIA – Centre de recherche Rennes – Bretagne Atlantique
Campus de Beaulieu, F-35042 Rennes Cedex, France
Email: sylvain.collange@inria.fr

***Abstract*—Floating-point additions in concurrent execution environment are known to be hazardous, as the result depends on the order in which operations are performed. This problem is encountered in data parallel execution environments such as GPUs, where reproducibility involving floating-point atomic addition is challenging. This problem is due to the rounding error or cancellation that appears for each operation, combined with the lack of control over execution order. In this article we propose two solutions to address this problem: work reassignment and fixed-point accumulation. Work reassignment consists in enforcing an execution order that leads to weak reproducibility. Fixed-point accumulation consists in avoiding rounding errors altogether thanks to a long accumulator and enables strong reproducibility.**

## I. Introduction

Efficient exploitation of modern multicore architectures relies on a hierarchical structuration of computation as well as execution concurrency. This affects determinism and numerical reproducibility making software development tedious. For example, GPUs manage several thousands of concurrent threads thanks to hardware resources such as warp and block schedulers. The design complexity of those processors is such that thread scheduling is mostly unknown and can be considered unpredictable. A common workaround is to enforce interaction between tasks using memory consistency with synchronization mechanisms such as locks, atomics or barriers.

Atomic operations are designed to perform a read-modify and write operation in one instruction. For example, atomicAdd() reads a word at a given address, adds a number to it, and writes the result back to the same address. The operation is atomic in the sense that it is guaranteed to be performed without interference from other threads. In other words, no other thread can access this address until the operation is complete. Although atomic operations can address the problem of memory consistency, they do not solve the problem of numerical reproducibility when dealing with floating-point numbers. This is a major issue as non-determinism of floating-point calculations in parallel programs causes validation and debugging issues, and may even lead to deadlocks [1].

The numerical non-reproducibility of floating-point atomic additions is due to the combination of two phenomena: rounding-error and the order in which operations are executed. This problem can be depicted with the simplified following CUDA kernel which computes the sum of $N$ floating-point numbers stored in table *i_val* according to their address *i_adr* in a table *res* located in global memory.

```
__global__ void GlobalSum(float *i_val,
             int *i_adr, float *res, int N){
  int gid =  blockDim.x*blockIdx.x+threadIdx.x;

  for(uint i=0; i<N; i+=GridDim.x*blockDim.x)
     atomicAdd(&res[i_adr[i+gid]],
             i_val[i+gid]);
}
```

Listing 1.  Floating-point atomic accumulation

The problem with this simple code, is that we do not have any information on the order in which threads will acquire access to the datum *Res*. For example, on a set of $N = 2^{16}$ values with a condition number[1] of $10^8$ and a single output address, we measured that out of over 1000 runs with 1 block of 1024 threads on a GTX680 we obtain 1000 different results.

One can argue that in the case where there is a single accumulator, weak reproducibility could be achieved by replacing atomicadd with standard addition combined with a reduction algorithm. However this solution does not hold when some threads are not producing any value (not executing the atomic addition), or when there are multiple accumulator, as in the bin counting or histogram problem as encountered in Nbody [3],

[1] The condition number characterizes the numerical stability of a problem [2].

978-83-60810-66-8/$25.00©2015, IEEE

Real-time simulation [4], accurate reduction scheme [5], or SQL query [6]. In these applications, floating-point atomic addition is used to accumulate values while preserving memory consistency.

In this article, we will use two concepts regarding numerical reproducibility in the context of data parallelism. *Weak numerical reproducibility* consists in being able to reproduce the same result between two executions when input parameters are identical. *Strong numerical reproducibility* consists in being able to reproduce the same results between two executions independently of the execution parameters or the architecture. Strong numerical reproducibility can be further subdivided in two classes of algorithms. Those which are producing correctly rounded results such as the one based on long accumulators [7], and others which provide reproducible results without any guaranty on accuracy [8].

Following these two definitions, we propose two solutions to address reproducibility involving floating-point atomics addition in the context of GPU programming by attacking the problem at its two roots. The first solution, work reassignment, consists in enforcing an evaluation order by overloading the block index assignment. With this solution, results are identical among runs for identical execution parameters. It is considered weakly reproducible as the number of threads per block affects the evaluation scheme and therefore the result. The second solution, fixed-point accumulation, avoids the rounding errors that occur during floating-point addition by using a long accumulator. As the addition is now exact, thus associative, the result is independent on the evaluation order or the hardware. In addition, the result is as accurate as possible as the accumulation is performed exactly. This solution is considered strongly reproducible.

The rest of this article is organized as follow. Section II introduces the necessary background about floating-point arithmetic and model of execution of GPUs. Section III presents the first solution we propose, based on block reordering. Section IV presents the second solution based on long accumulators. Section V analyses the theoretical cost of both methods and Section VI presents performance measurements on Nvidia GPUs.

## II. Background

### A. Floating-point arithmetic

Floating-point (FP) numbers approximate real numbers with a significand, an exponent, and a sign. The IEEE-754 standard, which was revised in 2008, specifies floating-point formats and operations. In this paper, we consider the `binary32` or single precision format. The floating-point number system can represent a wide range of numbers with nearly-constant precision.

Floating-point addition is not associative, due to the rounding error that occur when adding numbers with different exponents. It leads to the absorption of the lower bits of the sum. For example the exact mathematical result of $(1 + 2^{100} - 2^{100})$ is equal to 1 whereas the computed result is either 0 or 1 depending on the order of operations. Thus, the final accuracy of a floating-point summation depends on the order of evaluation. More details can be found in the main references related to floating-point arithmetic [9], [10]. This problem of numerical reproducibility linked to the order of floating-point operations, is amplified when executed in massively parallel environments like GPUs.

### B. GPU execution model

In this article we consider CUDA capable Nvidia GPUs used for the execution of tasks exhibiting data parallelism. These tasks are divided in threads operating in SIMT mode and executed by specific hardware. We must distinguish the software and hardware organization of threads. From the developer point of view, threads are divided into three hierarchical levels: a *grid* of *blocks* of *threads*. The same code, or kernel, is executed by multiple threads running in parallel on different data. *Threads* are grouped in set of *block_size* elements in order to make so-called *blocks*. Blocks are packed in set of *grid_size* elements in order to make a so-called *grid*. Threads in a block and blocks of a grid are uniquely identified by their coordinates in the blocks and the grid. In this model, threads in a block and the blocks of a grid are virtually launched in parallel, which implies that no assumption shall be made regarding the execution order.

In CUDA terminology, GPU hardware consists of CUDA cores organized hierarchically. These CUDA cores are grouped in streaming multiprocessors (SMs). The number of SMs varies depending on the architecture of the GPU and the CUDA compute capability. An additional and transparent level of grouping is introduced at the hardware level: the warp. These warps, corresponding to 32 threads, are created, managed, launched and executed by SIMT units. Multiprocessors share the instruction fetch, decoding and control logic across all the threads in a warp, so they run in lock-step [11].

Blocks are dispatched among the available multiprocessor by the block schedulers. This step consists in launching a new block with a unique identifier according to available resources. The number of concurrent blocks depends on the number and version of SMs and the resources such as registers and shared memory required by the executed kernel. This step impacts determinism as no assumption can be made on how indexes are generated and is subject to variations from one run to another [12].

Traditional thread synchronization primitives used in concurrent programming such as mutexes, barriers, and semaphores are limited on GPU to intra-block synchronizations. Atomic operations provide basic support for inter-block communication. Atomic instructions were first introduced on Nvidia hardware with compute capability 1.1, and atomic addition operating on 32-bit floating-point values in global and shared memory were introduced with compute capability 2.0. Atomic floating-point operations are necessary, first to provide the substrate for high performance floating-point operations, and second, to preserve the memory consistency necessary to deal with thousands of threads in flight.

The variety and performance of synchronization primitives available on GPUs have continuously increased over the years. This has lead to numerous works that have been focusing on efficient inter-block synchronization. For example, in [13] Volkov et al. propose a global software synchronization method that does not use atomic operations to accelerate dense linear-algebra constructs. In [14], [15], Xiao and Feng propose a mechanism for inter-block communication via global memory. In [16], Stuart and Owens are evaluating various implementations of barriers, mutexes and semaphores applied to Nvidia's GPU. Collange et al. use long accumulators to enable reproducible floating-point reductions [7]. However, none of those works have addressed the problem of reproducible atomic floating-point addition.

## III. Enforcing execution order

As discussed in the introduction, atomic operations enforce memory consistency, but not execution order. This means that atomic additions ensure that all data will be considered, but do not guarantee the order of operations. This is problematic for floating-point addition. The first solution we propose is to enforce an execution order for atomic addition, similarly to what can be achieved using a reduction algorithm. Threads and blocks are spawned and scheduled by hardware, following unspecified and implementation-specific policies. Thus, no assumption can be made on their real execution order. Therefore, we cannot rely on CUDA thread and block identifiers used at software level. Such solution would leads to an inefficient execution scheme and most likely to a deadlock situation [15].

The proposed solution uses two key concepts. The first one is based on a new software assignment of the block index, in order to guarantee a fine control over blocks effectively executed by the hardware. The second concept is based on atomic locks in order to ensure uniqueness of accesses at block level. We extend the solution proposed in [17] in order to control the order of execution. The lock used is similar to a fetch-and-add mutex algorithm in which the block index corresponds to the token acquired by a block.

### *A. Block re-ordering*

Block synchronization is challenging, as the CUDA programming model does not support it. The only safe solution consists on splitting kernels into subkernels, as a kernel launches involves an implicit synchronization barrier. Alternatively, resident kernel techniques take advantage of the fact that once launched, a block continue its execution until completion, freeing resources only at the end. For example, the block barrier proposed by Feng in [14] is working only when the number of launched blocks is less than the number of blocks that could be executed concurrently on the hardware. In case this assumption is not met, deadlocks may occur. For example, consider a GPU and a kernel such that only 8 blocks can run concurrently. If the kernel is launched on more than 9 blocks, then at least 1 block will not be scheduled. In that case running blocks will be waiting on the barrier for every block to complete, which will never happen as resources taken by those running block will never be released for others to complete.

In our case, we want to ensure that blocks are executed in a reproducible manner. This requires defining an execution order, which corresponds to a statically known order. This order can be, for example, the one corresponding to their block index. Besides requiring a known order, we need to prevent deadlock situation, which involves restrictions on the real scheduling of blocks.

The proposed solution consists in overriding the index generated by the block scheduler. This can be done by using a global variable *oBlkId* set to 0 at kernel launched which will be queried by one thread of each block at the beginning of the execution. The resulting index is stored in a shared variable *sBlkIdx* accessible by every thread of the block they belong to, and can be used as the new block index. An overview of the code is given in listing 2.

```
// Shared BlockIdx within a block
__shared__ uint sBlkIdx;

// Reindexing Block
struct BlkIdx{
  // New ordered index for BlockIdx
  uint *oBlkId;

  BlkIdx(void){
    cudaMalloc((void**) &oBlkId, sizeof(uint));
    cudaMemset(oBlkId, 0, sizeof(uint));
  }

  ˜BlkIdx(void){
    cudaFree( ordBlockId );
  }
  __device__ uint get_ordered_blockId(){
    if (threadIdx.x == 0)
      sBlkIdx = atomicInc(oBlkId, gridDim.x-1);

    __syncthreads();
    return sBlockIdx;
  }
};
```

Listing 2. Block reindexing

Thanks to this new block index, we can ensure a known execution order for block, which will be deadlock-free. With this technique, we chain together running blocks following a well-defined order. Therefore the overhead is solely concentrated in the access to the global index *sBlockIdx*. We now have to set a floating-point atomic addition which is reproducible within each block, which we will describe next.

### B. Atomic completion

Threads of blocks are scheduled as set of 32 consecutive threads, or warps. Again, as we do not have any control over the execution order of threads, we need to make sure that threads of a given block are always scheduled in a similar fashion when they are atomically accessing the destination address.

A solution is to encapsulate the atomic addition within a fetch-and-add mutex (FA) algorithm similarly to the one described in listing 3. With this solution, each thread of a given block is waiting for its turn on a variable shared at block level. These results ultimately to a serialization of each atomic addition at both thread-level and block-level as there are launched in an order corresponding to their global identifier. This solution is simple and provides strong numerical reproducibility. However accesses cannot happen concurrently and will perform poorly. We should mention that a syncthreads barrier is required at the end of the outer loop to ensure that no warp will go faster than any other warp with lower thread identifier.

```
struct Lock{
  uint *g_lock;

  Lock(void){
    cudaMalloc( (void**) &g_lock, sizeof(uint));
    cudaMemset( g_lock, 0, sizeof(uint));
  }

  ~Lock(void){
    cudaFree( g_lock );
  }

  __device__ void acquire_lock(int goalval){
    if (threadIdx.x == 0)
      while(atomicAdd(g_lock, 0) != goalval);
    __syncthreads();
  }

  __device__ void release_lock(){
    __syncthreads();
    if (threadIdx.x == 0){
      (*g_lock) = ((*g_lock)+1)%(gridDim.x);
    }
  }
};
```

Listing 3. Fetch-and-Add mutex

To improve performance of the previous solution, we can relax the constraint on the execution order in order to allow the atomic additions within a block to be replaced by simple additions executed in parallel. This results in a reduction scheme that only depends on the block size and is therefore considered as weakly reproducible. This solution requires 6 steps as described in listings 4. First, floating-point atomic data and addresses are stored in parallel in two tables located in shared memory. Then, both tables are sorted using the corresponding destination address as sorting key. In our implementation we used a bitonic sort as it preserve the order in case of equality. Once data are sorted according to their address, we perform a segmented sum for data with identical address. Then additions to the destination adresses are done in parallel for each different addresses within a block. We should mention that the number of additions is bounded by the number of threads per block. Those additions do not require atomic operations as atomic access is guaranteed with a FA lock on the new block identifier.

```
__device__ void ordered_concurrent_ AtomicAdd(
        float *dest, float val, Lock &lock){
  int tidx = threadIdx.x;
  // 1: Store in shared mem
  s_adr[tidx] = dest;
  s_val[tidx] = val;
  __syncthreads();

  // 2: Sort Element
  bitonicSort( s_adr, s_val);

  // 3: Segmented Sum
  segmented_sum_per_block(s_adr, s_val);

  // 4: Aquire the lock in order
  lock.acquire_lock(sBlockIdx);

  // 5: Final write in
  if (threadIdx.x < blockDim.x-1){
    if (s_adr[tidx] != s_adr[tidx+1]){
      *(s_adr[tidx]) += s_val[tidx];
    }
  }else{
      // For the last thread !
      *(s_adr[tidx]) += s_val[tidx];
  }
  // 6: Release the lock
  lock.release_lock();
}
```

Listing 4. Second solution: Serialization of atomic access at block level

## IV. Making addition associative

To achieve numerical reproducibility of atomic addition, we have seen in section III a solution based on enforcing an execution order. This section describes an alternative solution that consists in avoiding rounding errors to compute the correctly-rounded result.

To enforce strong reproducibility, we replace floating-point additions by fixed-point additions, which are associative. To avoid losing any precision compared to the floating-point format, we use a fixed-point accumulator that covers the whole range of values representable in the considered floating-point format. For instance, an accumulator with 127 integral bits and 127 fractional bits covers the range of single-precision floating-point. Such counter is known as a long accumulator or superaccumulator [18].

While the use of such a wide accumulator may seem extremely costly at first sight, three considerations make its performance actually attractive in the context of atomic operations:

- First, the average complexity of adding one floating-point value to a superaccumulator does not depend on the width of the superaccumulator [7]. Assuming the accumulator is represented as a vector of machine words (32-bit in the single-precision case), an accumulation generally only affects two words in the vector. Although carries may need to be propagated, the probability that a carry propagates across multiple words quickly gets negligible assuming low-order digits follow a uniform distribution [19].
- Second, updates to individual words of the superaccumulator can happen in any order without bearing any impact on the result, as long as all carries are detected and eventually propagated (including carries that occur during carry propagation). This enable regular integer atomic operations to update each part of the accumulator independently, without the need to implement any locking or coarser-grain transactions. Carries are detected *a posteriori* from the older value returned by the atomic operation and the value that was accumulated. Carries that occur are iteratively propagated to the higher-order words of the superaccumulator. Atomic operations from other threads may modify the superaccumulator during carry propagation; however, it bears no impact on the final result as all carry are eventually propagated in an arbitrary order.
- Third, this solution only requires hardware support for atomic integer addition, and does not need support for atomic floating-point addition. Therefore, the long accumulator allows to design reproducible atomic addition for any floating-point representation format (e.i. half, single or double precision of the IEEE-754 standard). By contrast, current GPUs only support floating-point atomic operations on single-precision data. The only difference lies in the size of the long accumulator.

We propose a thread-safe generalization of the long accumulation algorithm [20]. This accumulation algorithm is illustrated on Figure 1. The exponent and mantissa are extracted from the input number. High-order bits $e_h$ of the exponent are used to select the words that are affected in the superaccumulator. Lower-order bits $e_l$ select a cutoff point to split the mantissa into (at most) two parts, $m_h$ and $m_l$, whose weights are aligned with the words of the superaccumulators. This has the effect of shifting the mantissa to align it with the superaccumulator words. Words $m_h$ and $m_l$ are independently accumulated atomically to their respective accumulator words. Carries are detected and propagated using independent atomic operations.

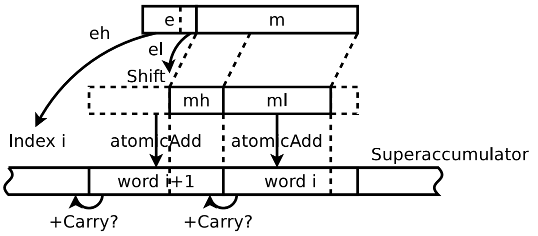


Fig. 1. Accumulation of a floating-point number to a superaccumulator.

## V. Cost

In this section we details the memory and operation cost of the two proposed solutions.

### A. Work reassignment

The solution based on enforcing an execution order requires one integer for the global index, one for the lock and one integer in shared memory for each block to store the shared index. In addition to these elements it requires a table of $N$ integers and $N$ floating-point numbers in shared memory per block to store temporary values, with $N$ corresponding to the number of threads per block.

In terms of computational overhead, this solution first requires to get a new block index at the beginning of the kernel. This involves one atomic instruction. Then, the kernel is left unmodified except when floating-point atomic additions need to be performed. In that case, each atomic addition is replaced by a write of both the address and the value to shared memory, plus a bitonic sort and a segmented sum per block over $N$ elements achieved in $O(log(N))$.

Then, threads of each block are waiting for their turn by spinning over the lock index. Once this condition is met, threads with the last address to be written send their write to global memory along the normal store path. As mentioned before these writes do not require atomic addition and can be done in parallel with an increased probability that global memory accesses are coalesced as data are sorted according to their destination addresses. For these reasons, this solution will perform well when many threads of the same block attempt to atomically add a value to the same location.

### B. Fixed-point accumulation

The fixed-point accumulator solution adds overhead in storage, computation, memory transfers and adds an extra rounding step. The most obvious cost of a long accumulator is its memory overhead. An accumulator able to represent exactly every single-precision floating-point value requires 280 bits of storage, as opposed to 32 bits using a floating-point accumulator. Likewise, a double-precision long accumulator needs about 2100 bits. Fortunately, long accumulators often contain sparse data: when accumulating numbers of the same order of magnitude, only a small part of the accumulator will be accessed, while the remaining part stays null. The hot parts can fit inside caches and benefit from fast cached atomic operations, while the cold part will remain in the large, off-chip memory.

Adding a floating-point value to a long accumulator also requires extra computations. The algorithm described in section IV extracts the exponent and mantissa of the floating-point number, then splits the mantissa by scaling and rounding it twice. However, binning algorithms are memory intensive and their performance is usually not limited by computations. This means the computational of long accumulation may be hidden by the memory access delays.

In terms of memory accesses, fixed-point accumulation requires two atomic operations per number in the common case, while a floating-point accumulation will only need one atomic add on platforms that support it in hardware. On the other hand, conflicts are reduced as atomic accesses span a larger memory area.

Finally, the fixed-point result has to be rounded to floating-point at the end of the accumulation. Rounding involves finding the first significant bit, then scanning the rest of the accumulator to compute the round, guard and sticky bits, that are necessary for IEEE-754 compliant rounding. The rounding phase can be performed in parallel between accumulators as they have no dependencies.

## VI. RESULTS

We tested both methods on a GTX480, a GTX560 and a GTX680 architectures with characteristics described in Table I. We made comparisons for a number of output addresses ranging from 1 (which corresponds to an accumulation) to 16384 different addresses randomly generated. Each thread was responsible for the accumulation of one floating-point value at a given address in global memory. Each kernel was launched with 100 blocks of 512 threads.

We compared the proposed methods against the *unordered solution* which consists in atomically adding the generated data at a given address as described in listing 1. For comparison purposes, we also included the fully sequential solution (block and warp sequential). Both block and warp sequential and block sequential methods relies on enforcing an execution order whereas superaccumulators avoid rounding errors altogether.

Figure 2(a), 2(b) and 2(c) describes the execution time on the GTX480, GTX560 and GTX680 respectively. One can observe that the execution time of the unordered solution quickly decreases as the number of output addresses increases. This is due to reduced contention : as the atomic accesses are spread across a wider range of accesses, the probability of conflicts decreases and the global performance improves. On the other hand, both sequential methods exhibit an execution time that is almost insensitive to the number of output addresses. The solution based on the long accumulator has an execution cost that decreases on the GTX480 and GTX560 and quickly increases after 64 to 128 different addresses. The decrease is due to concurrent accesses that can happen simultaneously as the number of different accumulators increases. However after a certain bound, the overhead of the accumulator do not compensate the gain obtained by the increase in concurrent accesses.

On the GTX 680, the unordered algorithm is always best according to figure 2(c), which is consistent with the fact that atomic operations were improved on this architecture compared to previous generations. Atomic performance is less sensitive to conflicts. The long accumulator also benefits from the improving atomic performance.

On GTX480 and GTX560, one can notice that the block sequential solution can even be faster than the unordered solution for small numbers of output addresses. This is because the small number of atomic access compensates the overhead of the segmented summation and sorting. By contrast, for the unordered solution almost every atomic access generated by each thread is conflicting on the output address, leaving no room for concurrent execution.

Out of these results, we can observe that enforcing strong numerical reproducibility is between 1.3 and 21 times more expensive than the unordered summation, whereas weak numerical reproducibility corresponds to 0.4 to 10 times the execution time of the unordered summation.

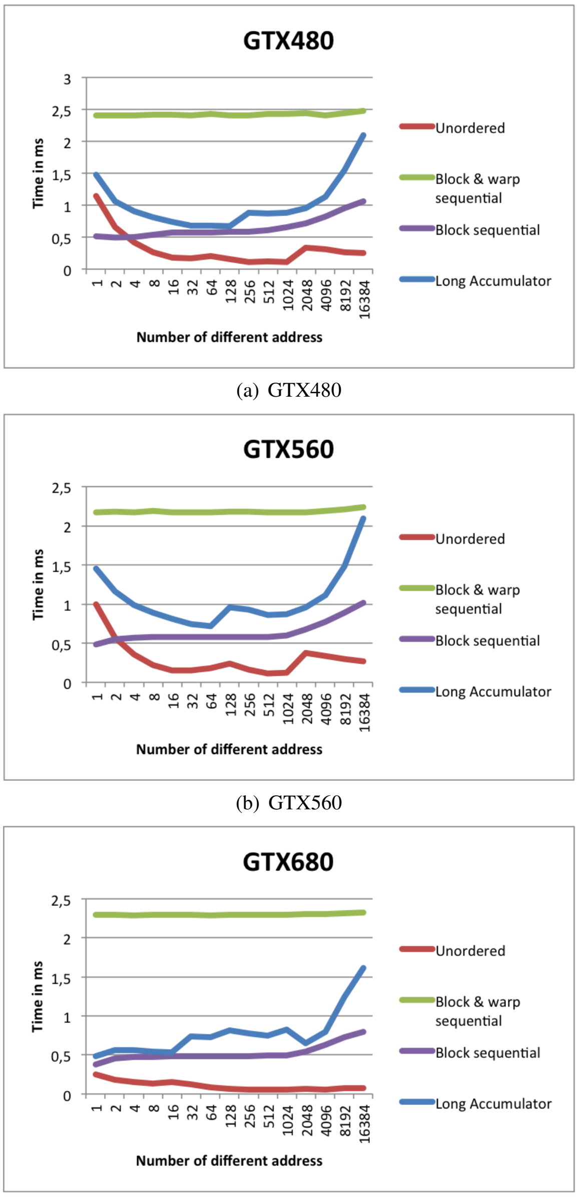
We should mention that the solution based on enforcing execution order uses block-wide synchronization barriers. This implies that atomics using the weakly reproducible solution need to be invoked from uniform control-flow regions. Every thread has to execute the instruction FPatomicAdd, otherwise the behavior is unspecified.

## VII. Conclusion

Atomic operations are very helpful in data parallel programming to enforce memory consistency. However when dealing with floating point values, this concurrency can lead to high variability, even when input data and execution parameters are identical. This behavior, due to rounding errors combined with the lack of control over execution order, is problematic as it causes validation and debugging issues, as well as producing hard-to-diagnose bugs in distributed environments.

In this article we propose two solutions to tackle this problem, each one addressing one source of the problem. We have described a first solution that enforces an execution order at block level thanks to new block synchronization primitive. We have shown that this solution is competitive in terms of performance (0.4-10 times) with the traditional hardware-based solution which is not reproducible. However, numerical reproducibility is ensured only for similar execution parameters as this solution depends on the number of threads per block and is therefore considered weakly reproducible.

We have proposed another solution based on a very long accumulator, which eliminates every rounding error and makes floating-point addition associative. This solution avoids the need to enforce an execution order and provides strong numerical reproducibility. It provides a correctly-rounded result with optimal accuracy. We have shown that this solution is between 1.3 and 21 times more expensive than the unordered solution. On the other hand, this solution does not require hardware support for floating-point atomic addition and can be applied to any floating-point representation format.



(a) GTX480

(b) GTX560

(c) GTX680

Fig. 2. Execution time to atomically add 1 floating-point value per threads to a number of different addresses ranging from 1 to 16384. Kernel are launched with 100 blocks of 512 threads.

## References

[1] K. Doertel, “Best known method: Avoid heterogeneous precision in control flow calculations,” Intel, Tech. Rep., 2013.

[2] N. J. Higham, *Accuracy and stability of numerical algorithms*. SIAM, 2002, second edition. [Online]. Available: http://www.maths.manchester.ac.uk/~higham/asna

[3] (2014, july) N-body: Fp atomics v. recomputation. [Online]. Available: http://blog.cudahandbook.com/2012/11/02/n-body-fp-atomics-v-recomputation.aspx

[4] J. Allard, S. Cotin, F. Faure, P.-J. Bensoussan, F. Poyer, C. Duriez, H. Delingette, and L. Grisoni, “Sofa an open source framework for medical simulation,” in *Medicine Meets Virtual*

TABLE I.
DESCRIPTION OF NVIDIA GPU' S OF DIFFERENT GENERATIONS .

| GPU | Arch. | CUDA Cap. | #MP/ GPC | #MP | CUDA Core /MP | Warp. Scheduler /MP | GPU Clock (Mhz) | Memory Clock (Mnz) |
|---|---|---|---|---|---|---|---|---|
| GTX 480 | GF100 | 2.0 | 4 | 15 | 32 | 2 | 1401 | 1848 |
| GTX 560 | GF114 | 2.1 | 4 | 7 | 48 | 2 | 1620 | 2004 |
| GTX 680 | K10 | 3.0 | 3 | 8 | 192 | 4 | 1059 | 3004 |

Reality (MMVR'15), Long Beach, USA, February 2007.

[5] W.-F. Chiang, G. Gopalakrishnan, Z. Rakamari´ c, D. H. Ahn, and G. L. Lee, "Determinism and reproducibility in large-scale HPC systems," in *Informal Proceedings of the 4th Workshop on Determinism and Correctness in Parallel Programming* (WoDet 2013), 2013.

[6] P. Bakkum and K. Skadron, "Accelerating SQL database operations on a GPU with CUDA," in *Proceedings of 3rd Workshop on General Purpose Processing on Graphics Processing Units,* GPGPU 2010, Pittsburgh, Pennsylvania, USA, March 14, 2010, ser. ACM International Conference Proceeding Series, D. R. Kaeli and M. Leeser, Eds., vol. 425. ACM, 2010, pp. 94–103. [Online]. Available: http://doi.acm.org/10.1145/1735688.1735706

[7] S. Collange, D. Defour, S. Graillat, and R. Iakymchuk, "Full-Speed Deterministic Bit-Accurate Parallel Floating-Point Summation on Multi- and Many-Core Architectures," INRIA, DALI–LIRMM, LIP6, ICS, Tech. Rep. HAL: hal-00949355, Feb. 2014.

[8] J. Demmel and H. D. Nguyen, "Parallel reproducible summation," *IEEE Trans. Computers,* vol. 64, no. 7, pp. 2060–2070, 2015. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TC.2014.2345391

[9] N. J. Higham, *Accuracy and stability of numerical algorithms,* 2nd ed. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2002.

[10] J.-M. Muller and al., *Handbook of floating-point arithmetic.* Birkhäuser, 2010.

[11] J. Nickolls and W. J. Dally, "The GPU computing era," *IEEE Micro,* vol. 30, pp. 56–69, March 2010. [Online]. Available: http://dx.doi.org/10.1109/MM.2010.41

[12] D. Defour, "Impacting predictability of gpu's," *HAL-CCSD, Tech.* Rep. hal-00951920, 2013. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00951920

[13] V. Volkov and J. W. Demmel, "LU , QR and Cholesky factorizations using vector capabilities of GPUs," Department of Electrical Engineering and Computer Science, University of California, Berkeley, inst-UCB-EECS:adr, LAPACK Working Note 202, May 2008. [Online]. Available: http://www.netlib.org/lapack/lawnspdf/lawn202.pdf

[14] W. chun Feng and S. Xiao, "To gpu synchronize or not gpu synchronize?" in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on,* 2010, pp. 3801–3804.

[15] S. Xiao and W. chun Feng, "Inter-block GPU communication via fast barrier synchronization," in IPDPS. IEEE, 2010, pp. 1–12. [Online]. Available: http://dx.doi.org/10.1109/IPDPS.2010.5470477

[16] J. A. Stuart and J. D. Owens, "Efficient synchronization primitives for GPUs," CoRR, vol. Abs/1110.4623, 2011. [Online]. Available: http://arxiv.org/abs/1110.4623

[17] J. Sanders and E. Kandrot, *CUDA by example: an introduction to general-purpose GPU programming.* pub-AW:adr: Addison-Wesley, 2010.

[18] U. W. Kulisch, *Computer arithmetic and validity,* 2nd ed., ser. de Gruyter Studies in Mathematics. Berlin: Walter de Gruyter & Co., 2013, vol. 33, theory, implementation, and applications.

[19] T. Grandlund, "GNU MP: The GNU Multiple Precision Arithmetic Library," http://gmplib.org.

[20] G. Bohlender and U. Kulisch, "Comments on fast and exact accumulation of products," in *Applied Parallel and Scientific Computing.* Springer, 2012, pp. 148–156.

# 2nd International Workshop on Cyber-Physical Systems

PROLIFERATION of computers in everyday life requires cautious investigation of approaches related to the specification, design, implementation, testing, and use of modern computer systems interfacing with real world and controlling their environment. Cyber-Physical Systems (CPS) are physical and engineering systems closely integrated with their typically networked environment. Modern airplanes, automobiles, or medical devices are practically networks of computers. Sensors, robots, and intelligent devices are abundant. Human life depends on them. Cyber-physical systems transform how people interact with the physical world just like the Internet transformed how people interact with one another.

The event is a continuation and extension of 2006-2010 Real-Time Software FedCSIS workshops and 2013 IWCPS. The objective of the workshop is to assemble and develop a community with main interest in cyber-physical systems.

## Topics

Due to an extensive scope of the topics, the workshop will accept papers in the following areas:

- Control Systems
  - embedded/networked/intelligent
  - wireless sensing/actuation
  - adaptive/predictive
- Scalability/Complexity
  - modularity
  - design methodology
  - legacy systems
  - tools
- Interoperability
  - concurrency
  - models of computation
  - networking
  - heterogeneity
- Validation and Verification
  - assurance
  - certification
  - simulation
- Cyber-security
  - intrusion detection
  - resilience
  - privacy
  - attack vectors
- Applications of CPS
  - robotics
  - transportation
  - military
  - medical
  - consumer
  - manufacturing
  - power systems
- CPS Education
  - curriculum development
  - web-based laboratories
  - academic courses
  - pedagogy issues

## Event Chairs

**Grega, Wojciech,** AGH University of Science and Technology, Poland

**Kornecki, Andrew J.,** Embry Riddle Aeronautical University, United States

**Nigro, Libero,** Università della Calabria, Italy

**Szmuc, Tomasz,** AGH University of Science and Technology, Poland

**Zalewski, Janusz,** Florida Gulf Coast University, United States

## Program Committee

**Angiulli, Fabrizio,** University of Calabria

**Babiceanu, Radu,** ERAU

**Cicirelli, Franco,** Università della Calabria, Italy

**Crespo, Alfons,** Universitat Politecnica de Valencia, Spain

**Golatowski, Frank,** University of Rostock, Germany

**Gomes, Luis,** Universidade Nova de Lisboa, Portugal

**Halang, Wolfgang A.,** Fernuniversitaet, Germany

**Letia, Tiberiu,** Technical University of Cluj-Napoca, Romania

**Malec, Jacek,** Lund University, Sweden

**Marwedel, Peter,** Technische Universität Dortmund, Germany

**Motus, Leo,** Tallinn University of Technology, Estonia

**Nadjm-Tehrani, Simin,** Linköping University, Sweden

**Nigro, Libero,** Universite della Calabria

**Rysavy, Ondrej,** Brno University of Technology, Czech Republic

**Sanden, Bo,** Colorado Technical University, United States

**Schagaev, Igor,** London Metropolitan University, United Kingdom

**Seker, Remzi,** Embry Riddle Aeronautical University, United States

**Sveda, Miroslav,** Brno University of Technology, Czech Republic

**Trybus, Leszek,** Rzeszow University of Technology, Poland

**Vardanega, Tullio,** University of Padova, Italy

**Villa, Tiziano,** Università di Verona, Italy

**Zoebel, Dieter,** University Koblenz-Landau, Germany

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 731–742

DOI: 10.15439/2015F132
ACSIS, Vol. 5

# Simulation and Formal Modelling of Yaw Control in a Drive-by-Wire Application

Richard Banach
School of Computer Science,
University of Manchester
Oxford Road, M13 9PL, U.K.
Email: banach@cs.man.ac.uk

Pieter Van Schaik, Eric Verhulst
Altreonic,
Gemeentestraat 61/A,
Linden B3210, Belgium
Email: {pieter.vanschaik, eric.verhulst}@altreonic.com

*Abstract*—**Cyberphysical systems, with their interdependence between physical behaviour and digital control, need insights from frequency domain control engineering, state space control engineering and discrete formal systems theory for their proper description. Neglecting any of these, results in descriptions that omit essential details. Hybrid Event-B is a formalism that enables all the relevant detail to be assimilated. A case study based on yaw control for the KURT e-vehicle is used as a testbed to explore the effective interaction between the various needed disciplines in exploring a specific design issue, the formalisation of yaw control discretization, using Hybrid Event-B.**

## I. Introduction

TODAY, the low cost, small size, low energy consumption and wide availability of digital processors, together with the ready availability of a wide variety of stadardised control components, makes the embedding of computing components and digital control into what was previously purely analogue equipment, ubiquitous. This has now given rise to the burgeoning field of cyberphysical systems, in which computing systems are intimately connected to equipment that acts in the physical space. Increasingly, the impact of such systems is safety critical, and in such cases, the techniques by which the systems are developed demand scrutiny — at least in those spheres where it is recognised that safety and dependability (more precisely, their potential lack) merit certification processes that have to be successfully passed before systems can be deployed in the field. In reality, what is required to be certified often lags well behind what is attempted and shown to be feasible technically.

An essential element of many such certification processes is a verifiable audit trail of mathematical models of the system and of their relevant properties. Ideally, all models and properties should be verifiably consistent with one another, and demonstrably possess the properties needed for safe operation. In the case of cyberphysical systems this ideal is challenging, for the following reasons.

Traditionally, control design is done in the frequency domain [1], [2], [3]. This readily yields the quantities needed by the engineer, using a mixture of rigorous results and design heuristics. Although the rigorous results often do not hold with mathematical precision in reality (e.g. needed bandwidth assumptions), the degree of inexactitude is not harmful in practice. A major element of this approach is the use of simulation to judge the suitability of a design, using tools like Modelica [4].

Traditionally, computing systems, which proceed by discrete steps, are modelled and analysed within a discrete state space — there is no notion of frequency domain for an arbitrarily constructed discrete space. Since there is an enormous variety in the aspects of behaviour that can be modelled by the discrete steps of a computing system, correspondingly, depending on what the elements of the state space represent, we find an enormous variety of approaches to the formalisation of computing systems [5].

Following on from this observation, traditionally, formalisms for computing systems (e.g. the many surveyed in [5]) do not engage with continuous mathematics at all. To address this shortcoming in the context of cyberphysical systems, two different approaches are seen. In the first, the formalism does not engage with continuous mathematics, stays essentially discrete, and incorporates facts concerning continuous aspects of the modelled system as inputs or axioms. In the second, special purpose formalisms are designed to include continuous phenomena in suitable ways alongside the discrete ones.

More recently, a more rigorous approach to control has emerged within 'mathematical control theory' [6], [7], [8], that emphasises the state based approach to control. This perspective is able to relate more directly to the state based perspective of computing formalisms in a way that the frequency domain perspective (although equivalent to it via transform theory) would struggle to do. Moreover, the rigour of the proofs in mathematical control theory typically matches much better with the style of argument in computing formalisms, both being ultimately grounded in set theory. Then again, much of the useful engineering information that is evident and easily manipulated in the frequency domain, and is inferred smoothly from simulation, can become greatly obscured in the state based approach. Thus there is a conflict between the usual approaches to the various disciplines that contribute to the cyberphysical systems agenda.

In this paper we examine a case study that hosts an encounter (we hesitate to say collision) between the various approaches and issues mentioned. Although it is a simplified case study, it is not a toy, in that it is drawn directly from a genuine system, the KURT e-vehicle from Altreonic. We specifically look at yaw control and its stability in KURT. We embed the abridged development in the Hybrid Event-B (HEB) formalism [9], [10], and focus on the formal description and properties of the discretization step from a high level continuous design to a lower level time triggered discrete one.

978-83-60810-66-8/$25.00©2015, IEEE

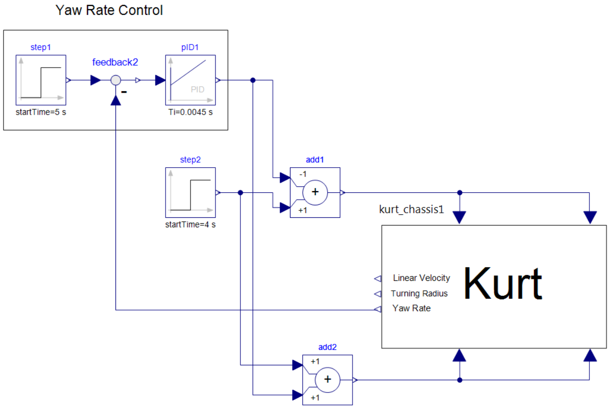


Fig. 1. KURT simulation: Modelica yaw rate control configuration.

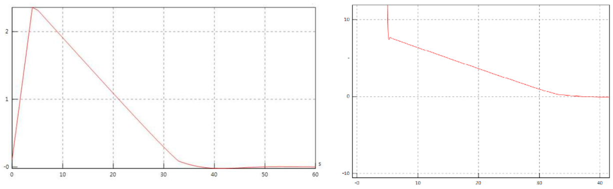


Fig. 2. KURT simulation: left, vehicle linear velocity (m/s) vs. simulation Time (s); right, vehicle turning radius (m) vs. simulation time (s).

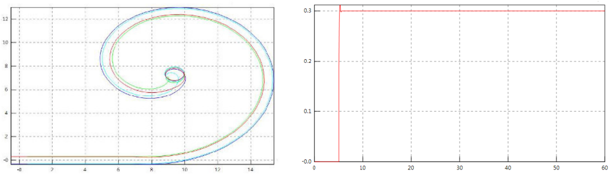


Fig. 3. KURT simulation: left, vehicle trajectory in Y vs. X axis (m); right, vehicle yaw rate response (rad/s) vs. simulation time (s).

From a rigorous point of view, discretization steps introduce copious amounts of low level detailed technical complexity. In order to keep the account within reasonable bounds, we do take some shortcuts in the development, commenting on the pros and cons as we go.

The rest of this paper is as follows. In Section II we overview KURT, and describe a Modelica simulation that was used to validate a number of design parameters for yaw control. In Section III we overview the HEB formalism, stressing the aspects that are most important for us. Section IV reformulates yaw control in HEB, and Section V examines the stability of the model in a state space based way. Section VI discusses the issues raised by going from a continuous to a discretized formulation in a formal manner. Section VII concludes.

## II. Yaw Control in the KURT E-Vehicle

The KURT e-Vehicle is an innovative vehicle concept based on a modular, scalable and fault tolerant architecture. The propulsion system of KURT utilises four independently controlled in-wheel motors and employs a differential steering technique combined with a drive-by-wire architecture. In its simplest form such a steering technique entails steering the vehicle by creating a difference in linear velocity between the left and right side of the vehicle. Such approaches are often utilised in unmanned robotic platforms as well as heavy earth moving machinery. In the absence of a mechanical steering mechanism, such as articulated steering, these vehicles are often referred to as skid-steer vehicles. The advantages of utilising a 4-wheel drive differential steered concept includes minimising the mechanical complexity of the steering mechanism as well as increased manoeuvrability. However, reducing mechanical complexity demands a more intelligent propulsion control system which in turn will be deployed on an embedded target platform. The employed steering control strategy will therefore do well by minimising implementation complexity specifically with regards to aspects such as required processing power and number of sensors. To this end a control strategy has been devised for the KURT e-Vehicle whereby the effective yaw rate of the vehicle is utilised as the control parameter.

According to the kinematic relations [11] of a differential steered vehicle, yaw rate is related to linear velocity and instantaneous turning radius, and can be relatively easily measured with inexpensive MEMS based sensors. The added benefit of yaw rate control is related to safety, more specifically, with regards to maintaining stability of the combined vehicle and driver centre of gravity (COG) when executing turning manoeuvres. In a drive-by-wire, and specifically a steer-by-wire system, there is not necessarily provision for a feedback mechanism which serves to cause the driver to limit the turning radius when executing a turn in accordance with the linear velocity of the vehicle. Simply put, if a turn is taken too sharply at too high speed the vehicle can topple over. Therefore, by controlling the yaw rate of the vehicle when executing a turn, the effective turning radius can be controlled. This in turn permits preventing the centrifugal force component from getting large enough to cause the vehicle to overturn.

A basic propulsion and steering control strategy is as follows: the throttle command issued by the user is interpreted as a thrust request which is translated into a torque command and is applied equally to all four wheels; a steering command issued by the user is translated into a yaw rate request which serves as the set point to a closed loop PID controller. The output of the PID controller is added to the torque command of the wheels on one side and subtracted on the opposite side. The side to which it is added or subtracted depends on whether the steering request is to the left or to the right. When steering to the left the output of the PID controller will be subtracted from the torque command of the wheels on the left and added to the torque commands of the wheels on the right. When steering to the right the situation will be reversed.

In order to simulate the proposed control strategy, a dynamic model of a skid-steer vehicle was created in Modelica [4]. For the purposes of this investigation, the model does not include complex tyre and surface interactions, but rather models the wheels as point masses, on which the propulsion, rolling resistance and friction forces act.[1] The vehicle is represented with a simple H-shaped geometry with a single point mass located equidistant from the rear and front wheel

[1] Tyre and surface interactions result in friction forces (and are even necessary to control the vehicle properly).

representing the combined vehicle and payload mass. It is also the yaw rate of this mass that is measured and utilised in the steering control loop. The PID steering controller was designed by applying the Cohen-Coon tuning method [12], [13], [14]. A typical yaw control simulation setup is depicted in Fig. 1.

The kurt_chassis1 component from Fig. 1 resembles the dynamic model of the skid-steer vehicle. The model receives four inputs namely the thrust applied to the left rear and front wheel masses as well the thrust applied to the right rear and front wheel masses. The model provides three outputs of which the yaw rate is of primary concern. The step input source element step2 simulates a thrust request issued by the user. The steering PID controller is implemented by pID1 of which the output is subtracted from the output of step2 by add1 and added to the output of step2 by add2. The output of add1 is applied equally to the left wheel masses whereas the output of add2 is applied equally to the right wheel masses. The purpose of step1 is to simulate a steering request issued by the user in the form of a yaw rate request. The output of step1 serves as the set point to the closed loop controller with the yaw rate output of the kurt_chassis1 being the measured variable. The simulation setup therefore represents a steering request to turn left. The simulation sequence commences by issuing a constant thrust request ($thr = 15\,\mathrm{N}$) for a duration of 4 seconds during which no steering request is present ($yrr = 0\,\mathrm{rad/s}$). At $t = 4$ seconds the thrust command is removed ($thr = 0\,\mathrm{N}$) and at $t = 5$ seconds a constant steering request is issued ($yrr = 0.3\,\mathrm{rad/s}$). The simulation continues to run until $t = 40$ seconds.

The vehicle is therefore expected to accelerate from $t = 0$ to $t = 4$ seconds after which it will decelerate. From $t = 5$ seconds onwards the vehicle is expected to turn to the left. According to the kinematic relations the turning radius of the vehicle is expected to decrease proportionally to the linear velocity provided that the yaw rate is held constant. Figs. 2 and 3 depict the results obtained from the simulation run. The RHS of Fig. 3 shows the step response of the measured yaw rate. From the results it is seen that the controller is sufficiently capable of maintaining a constant yaw rate with acceptable overshoot (4%) and settling time (0.1 seconds). From Fig. 2 it is seen that the turning radius decreases in proportion to the linear velocity in accordance with the kinematic relations. Fig. 3 also shows the trajectory of the four wheel masses. The trajectory indicates the vehicle follows a spiral path as the linear velocity and the turning radius decreases. Correlating the trajectory with the turning radius it is observed that the rear wheels progressively digress from the trajectory of the front wheels as the turning radius decreases. From $t = 32$ seconds onwards the vehicle starts to rotate around its own centre of mass resulting in near zero turning radius.

## III. An Outline of Hybrid Event-B

In this section we outline HEB, relating it to the more familiar Event-B [15]. The bulk of the material refers to a single machine. However, our models involve three machines: for the user, for KURT's behaviour, and for the control system, so we include what we need for multiple machines below.

### A. Single Hybrid Event-B Machines

In Fig. 4 we see a schematic HEB machine. It starts with declarations of time and of a clock. Time is a first class citizen in that all variables are functions of time (which is read-only), explicitly or implicitly. Clocks are assumed to increase like time, but may be set during mode events. Variables are of two kinds. There are mode variables (like $u$) which take their values in discrete sets and change their values via discontinuous assignment in mode events. There are also pliant variables (such as $x, y$), declared in the PLIANT clause, which typically take their values in topologically dense sets (normally $\mathbb{R}$) and which are allowed to change continuously, such change being specified via pliant events.

Next are the invariants. These resemble invariants in discrete Event-B, in that the types of the variables are asserted to be the sets from which the variables' values *at any given moment of time* are drawn. More complex invariants are similarly predicates that are required to hold *at all moments of time* during a run.

Then, the events. The *INITIALISATION* has a guard that synchronises time with the start of any run, while all other variables are assigned their initial values as usual.

Mode events are analogues of events in discrete Event-B. They can assign all machine variables (except time). The schematic *MoEv* of Fig. 4, has parameters $i?, l, o!$, (input, local, and an output), and a guard *grd*. It also has the after-value assignment specified by the before-after predicate *BApred*, which can specify the after-values of all variables (except time, inputs and locals).

Pliant events are new to HEB. They specify the continuous evolution of the pliant variables over an interval of time. Fig. 4 has a schematic pliant event *PliEv*. There are two guards: *iv*, for specifying enabling conditions on the pliant variables, clocks, and time; and *grd*, for specifying enabling conditions on the mode variables.

The body of a pliant event contains three parameters $i?, l, o!$, (again, input, local, and output) which are functions of time, defined over the duration of the pliant event. The behaviour of the event is defined by the COMPLY and SOLVE clauses. The SOLVE clause contains direct assignments, e.g. of $y$ and output $o!$ (to time dependent functions); and differential equations, e.g. specifying $x$ via an ODE (with $\mathcal{D}$ as the time derivative).

The COMPLY clause can be used to express any additional constraints that are required to hold during the pliant event via the before-during-and-after predicate *BDApred*. Typically, constraints on the permitted ranges of the pliant variables, can be placed here. The COMPLY clause can also be used to specify properties at an abstract level, e.g. stating safety properties for the event without going into detail.

Briefly, the semantics of a HEB machine consists of a set of *system traces*, each of which is a collection of functions of time, expressing the value of each machine variable over the duration of a system run.

Time is modeled as an interval $\mathcal{T}$ of the reals. A run starts at some initial moment of time, $t_0$ say, and lasts either for a finite time, or indefinitely. The duration of the run, $\mathcal{T}$, breaks up into a succession of left-closed right-open subintervals: $\mathcal{T} = [t_0 \ldots t_1), [t_1 \ldots t_2), [t_2 \ldots t_3), \ldots$. Mode events (with their discontinuous updates) take place at the isolated times corresponding to the common endpoints of these subintervals $t_i$,

```
MACHINE $HyEvBMch$
TIME $t$
CLOCK $clk$
PLIANT $x,y$
VARIABLES $u$
INVARIANTS
  $x,y,u \in \mathbb{R},\mathbb{R},\mathbb{N}$
EVENTS
  $INITIALISATION$
    STATUS ordinary
    WHEN
      $t = 0$
    THEN
      $clk,x,y,u := 1,x_0,y_0,u_0$
    END
... ...
```

```
... ...
  $MoEv$
    STATUS ordinary
    ANY $i?,l,o!$
    WHERE
      $grd(x,y,u,i?,l,t,clk)$
    THEN
      $x,y,u,clk,o! : |$
        $BApred(x,y,u,i?,l,o!,$
          $t,clk,x',y',u',clk')$
    END
... ...
```

```
... ...
  $PliEv$
    STATUS pliant
    INIT $iv(x,y,t,clk)$
    WHERE $grd(u)$
    ANY $i?,l,o!$
    COMPLY
      $BDApred(x,y,u,$
        $i?,l,o!,t,clk)$
    SOLVE
      $\mathcal{D}x =$
        $\phi(x,y,u,i?,l,o!,t,clk)$
      $y,o! :=$
        $E(x,u,i?,l,t,clk)$
    END
  END
```

Fig. 4. A schematic Hybrid Event-B machine.

and in between, the mode variables are constant, and the pliant events stipulate continuous change in the pliant variables.

We insist that on every subinterval $[t_i \ldots t_{i+1})$ the behaviour is governed by a well posed initial value problem $\mathcal{D}xs = \phi(xs\ldots)$ (where $xs$ is a relevant tuple of pliant variables). Within this interval, we seek the earliest time $t_{i+1}$ at which a mode event becomes enabled, and this time becomes the preemption point beyond which the solution to the ODE system is abandoned, and the next solution is sought after the completion of the mode event.

In this manner, assuming that the *INITIALISATION* event has achieved a suitable initial assignment to variables, a system run is *well formed*, and thus belongs to the semantics of the machine, provided that at runtime:

(1) Every enabled mode event is feasible, i.e. has an after-state, and on its completion enables a pliant event (but does not enable any mode event).[2]

(2) Every enabled pliant event is feasible, i.e. has a time-indexed family of after-states, and EITHER:
   (i) During the run of the pliant event a mode event becomes enabled. It preempts the pliant event, defining its end. ORELSE
   (ii) During the run of the pliant event it becomes infeasible: finite termination. ORELSE
   (iii) The pliant event continues indefinitely: nontermination.

Thus, in a well formed run mode events alternate with pliant events. The last event (if there is one) is a pliant event (whose duration may be finite or infinite). In reality, there are several semantic issues that we have glossed over in the framework just sketched. We refer to [9] for a more detailed presentation (and to [10] for the extension to multiple machines). The presentation just given is quite close to the modern formulation of hybrid systems. See e.g. [16], [17] — or [18] to get a perspective stretching further back.

If, from Fig. 4, we erase time, clocks, pliant variables and pliant events, we arrive at a skeleton (conventional) Event-B machine. This simple erasure process illustrates (in reverse) the way that HEB has been designed as a clean extension of the original Event-B framework. The only difference of note is that, now —at least according to the (conventional) way that Event-B is interpreted in the physical world— (the mode) events (left behind by the erasure) execute *lazily*, i.e. *not* at the instant they become enabled (which is, of course, the moment of execution of the previous event).

### B. Multiple Hybrid Event-B Machines

The principal objective in modelling complex systems in the B-Method is to start with small simple descriptions and to refine to richer, more detailed ones. This means that, at the highest levels of abstraction, the modelling must **abstract away from concurrency**. By contrast, at lower levels of abstraction, the events describing detailed individual behaviours of components become visible. In a purely discrete event framework, like conventional Event-B, there can be some leeway in deciding whether to hold all these low level events in a single machine or in multiple machines — because all events execute instantaneously, isolated from one another in time (in the usual interpretation).

In HEB the issue is more pressing. Because of the inclusion of continuous behaviour, *all* components are always executing *some* event. Thus an integrated representation risks hitting the combinatorial explosion of needing to represent each possible combination of concurrent activities within a separate event, and so there is a much stronger incentive to put each (relatively) independent component into its own machine, synchronised appropriately. Put another way, there is a very strong incentive to **not abstract away from concurrency**.

The same impulse is reinforced when we wish to construct systems out of components, e.g. a plant and a controller. There, it is also convenient to conceive the pieces separately and combine them appropriatelty. The key concept in achieving this is the INTERFACE. This is a syntactic contruct (adapted from the idea in [19]) that includes the declarations of a set of variables, the invariants that involve them, and also their initialisations. A community of machines may have access to the variables declared in an interface if each machine CONNECTS to the interface. All events in the machines must

[2] If a mode event has an input, the semantics assumes that its value only arrives at a time strictly later than the previous mode event, ensuring part of (2) automatically.

```
PROJECT Kurt_Prj
INTERACES
  YawCtrl_IF
MACHINES
  KurtUser_Mch
  Kurt_Mch
  YawCtrl_Mch
END
```

```
INTERFACE YawCtrl_IF
SEES Kurt_Ctx
TIME t
PLIANT
  yrr, yrm, stc,
  yreP, yreI, yreD,
  thr, tal, tar
INVARIANTS
  $yrr, yrm, stc \in \mathbb{R}, \mathbb{R}, \mathbb{R}$
  $yreD, yreP, yreI \in \mathbb{R}, \mathbb{R}, \mathbb{R}$
  $thr, tal, tar \in \mathbb{R}, \mathbb{R}, \mathbb{R}$
INITIALISATION
  WHEN
    $t = 0$
  THEN
    $yrr, yrm, stc := 0, 0, 0$
    $yreP, yreI, yreD := 0, 0, 0$
    $thr, tal, tar := 0, 0, 0$
  END
END
```

```
CONTEXT Kurt_Ctx
  ...   ...
AXIOMS
  ...   ...
END
```

```
MACHINE KurtUser_Mch
CONNECTS YawCtrl_IF
EVENTS
  SteerKurt
    STATUS pliant
    BEGIN
      $thr(t) := \Theta(4-t)$
      $yrr(t) := \Theta(t-5)$
    END
END
```

```
MACHINE Kurt_Mch
CONNECTS YawCtrl_IF
EVENTS
  KurtBehaves
    STATUS pliant
    SOLVE
      $\mathcal{D}\, yrm(t) := C_K stc(t)$
    END
END
```

```
MACHINE YawCtrl_Mch
CONNECTS YawCtrl_IF
EVENTS
  YawControl
    STATUS pliant
    SOLVE
      $yreP(t) := yrr(t) - yrm(t)$
      $yreD(t) := \mathcal{D}\, yreP(t)$
      $\mathcal{D}\, yreI(t) := yreP(t)$
      $stc(t) :=$
        $K_P[yreP(t) + yreI(t)/T_I + T_D yreD(t)]$
      $tal(t) := thr(t) - stc(t)$
      $tar(t) := thr(t) + stc(t)$
    END
END
```

Fig. 5. A Hybrid Event-B system for yaw control.

preserve all of the invariants in the interface, of course. An important point is that *all* invariants involving the interface's variables must be in the interface.

Multi-machine HEB systems need more than what we have just described, namely (at least) synchronisation and instantiation mechanisms. These, and other issues, are discussed in [10]. What we have mentioned will suffice for this paper.

## IV. A Hybrid Event-B Model of Yaw Control

In this section we take the model discussed in Section II and re-express it as a Hybrid Event-B PROJECT. The project itself appears in Fig. 5, where its overall structure is defined in the PROJECT *Kurt_Prj* file. This indicates the pieces that the system is constructed from. These consist of the INTERFACE *YawCtrl_IF* and the MACHINEs *KurtUser_Mch*, *Kurt_Mch* and *YawCtrl_Mch.*

The interface SEES the CONTEXT *Kurt_Ctx* which contains the definitions of all the constants and static mathematics that the project will need, and more importantly, it is also the home of any AXIOMS (concerning these static elements) that we may rely on for verification. The interface then names the (pliant) variables shared by the machines that connect to it, lists their invariants, and defines their intialisations. Table 1 lists the variables, and describes how they relate to the elements of the KURT simulation model in Fig. 1.

The three machines *KurtUser_Mch*, *Kurt_Mch* and *YawCtrl_Mch* are formal definitions of the three actors in the dynamics.

The *KurtUser_Mch* machine describes the behaviour of the user who drives KURT. The machine CONNECTS to the *YawCtrl_IF* interface, to access needed variables, and it has a single pliant event *SteerKurt*. This applies a constant thrust from time 0 to time 4 $thr(t) := \Theta(4-t)$, and a constant yaw request from time 5 onwards $yrr(t) := \Theta(t-5)$, where $\Theta$ is the Heaviside step function. This is consistent with the description in Section II.

Machine *Kurt_Mch* describes the intrinsic behaviour of the KURT e-vehicle. It also CONNECTS to *YawCtrl_IF*. In this simple model it is assumed that KURT will emit a measured yaw rate *yrm* whose derivative is proportional to the difference of the thrusts applied to left and right wheel sets $tar(t) - tal(t)$, and which is thus (via a positive constant $C_K$) proportional to the differential thrust $stc(t)$ (see Fig. 1):

$$\frac{d}{dt} yrm(t) = C_K \, stc(t) \tag{3}$$

Machine *Kurt_Mch* expresses this in HEB notation.

Machine *YawCtrl_Mch* describes the controller that turns the user's steering commands into thrust commands to KURT's wheels. Of course it CONNECTS to *YawCtrl_IF*. At its heart is the PID controller in Fig. 1 which calculates the differential steering thrust command $stc(t)$ from the value, integral and derivative of the yaw rate error $yer(t)$:

$$stc(t) = K_P \left[ yre(t) + \frac{1}{T_I} \int_0^t yre(s)\, ds + T_D \frac{d}{dt} yre(t) \right] \tag{4}$$

The formalism of HEB does not permit us to write this directly since (aside from implicit constraints in the COMPLY clause), it allows direct assignment and differential equations only, in the SOLVE clause. The formulation in the *YawCtrl_Mch* machine unwinds (4) into an acceptable form. Thus, separate variables are introduced for the proportional, integral and derivative of $yre(t)$: $yreP(t), yreI(t), yreD(t)$ (variable $yre(t)$ itself is not an element of the *Kurt* project). On this basis, equation (4) turns into the following lines of the SOLVE clause of the *YawControl* pliant event:

$$\begin{aligned} yreP(t) &:= yrr(t) - yrm(t) && (5)\\ yreD(t) &:= \mathcal{D}\, yreP(t) && (6)\\ \mathcal{D}\, yreI(t) &:= yreP(t) && (7)\\ stc(t) &:= K_P[yreP(t) + yreI(t)/T_I + T_D\, yreD(t)] && (8) \end{aligned}$$

The remaining assignments in the SOLVE clause of the *YawControl* event, quite faithfully mirror the relevant functions and connections of the yaw control model in Fig. 1, when the interpretation is mediated via the information in Table 1.

## V. Formal Properties of Yaw Control

Some properties can be easily checked from the text of Fig. 5. For instance, each of the variables of the *YawCtrl_IF* interface appears exactly once in the left hand side of any of the assignments or ODEs in any of the pliant events in the project. Since all these pliant events run concurrently, this property is a prerequisite for consistency.

The next obvious thing is the observation that all of the assignments and equations of the *Kurt* project are linear. This means that an analytic solution to the system's behaviour is within reach, which we examine now.

### A. Stability Analysis of the Simulation

Given that there are step functions in the system inputs $thr$ and $yrr$ at $t = 4$ and $t = 5$, the behaviour splits naturally into three intervals: $[0 \dots 4), [4 \dots 5), [5 \dots \infty)$.

During $[0 \dots 4)$ the vehicle accelerates from 0: thus $thr = 15 = tal = tar$, and all other variables remain at 0. During $[4 \dots 5)$, the thrust is switched off $thr = 0 = tal = tar$. So all variable values are 0. (In the simulation of Section II the vehicle begins to slow, although this depends on frictional forces not included in our formal model.)

**Table 1: Variables Used in the Yaw Control Models**

| Variable | Meaning |
|---|---|
| *yrr* | Yaw Rate Request (output of step1) |
| *yrm* | Yaw Rate Measured (output of Kurt) |
| *yre* | Yaw Rate Error, i.e. $yre = yrr - yrm$ (output of feedback2) |
| *yreD* | Time Derivative of Yaw Rate Error (derivative of output of feedback2) |
| *yreP* | Proportional Steering Yaw Rate Error, i.e. $yreP = yre$ (output of feedback2) |
| *yreI* | Time Integral of Yaw Rate Error (integral of output of feedback2) |
| *stc* | (Differential) Steering Thrust Command (output of pID1) |
| *thr* | Thrust Request (output of step2) |
| *tal* | Thrust Applied Left (output of add1) |
| *tar* | Thrust Applied Right (output of add2) |

Turning starts at $t = 5$, and we must solve the system of equations in the *Kurt* project. The *KurtBehaves* event in the *KurtMch* machine implies that $yrm(t)$ is the time integral of $C_K stc(t)$ up to a constant of integration $L_P$. This can be substituted into the right hand side of (5) which then yields $yreP(t)$. Differentiating this, in turn yields $yreD(t)$ via (6). Integrating it instead, yields $yreI(t)$ via (7), up to another constant of integration $L_I$. Substituting these relationships into (8) yields the integral equation:

$$\begin{aligned} stc(t) &= K_P(0.3 - L_P) - C_K K_P \int_5^t stc(s)\, ds \\ &\quad + \frac{K_P}{T_I}\left[(0.3 - L_P)(t-5) - L_I - C_K \int_5^t\!\!\int_5^s stc(u)\, du ds\right] \\ &\quad - C_K K_P T_D\, stc(t) \qquad (9) \end{aligned}$$

Differentiating this twice yields the homogeneous ODE:

$$\left(T_D + \frac{1}{C_K K_P}\right)\frac{d^2}{dt^2}stc(t) + \frac{d}{dt}stc(t) + \frac{1}{T_I}stc(t) \;=\; 0 \qquad (10)$$

The only solutions of (10) are exponential. Putting in the ansatz $stc(t) = Re^{\lambda t}$ and integrating twice yields candidates for the integral terms in the RHS of (9). Since (9) must be an identity, equating coefficients of $e^{\lambda t}$ and of the linear terms allows $L_P$, $L_I$ and $R$ to be determined from initial conditions. And with $stc(t)$ determined, we can easily calculate the behaviour of all the other system variables if we wish.

For mechanical stability, we need the real part of either value of $\lambda$ to be negative. This yields two constraints on the family of constants in the *Kurt* system, each being of the form $expr > 0$. However, up to positive constant factors and an additional factor of $T_I$, one $expr$ is the reciprocal of the other. Therefore, the two cannot be consistent unless $T_I > 0$, whence we get:

$$T_I > 0 \quad \text{and} \quad T_D + \frac{1}{C_K K_P} > 0 \qquad (11)$$

We can add these as AXIOMS to the context *Kurt_Cxt*:

AXIOMS
$T_I > 0$
$T_D + 1/(C_K K_P) > 0$

With axioms like these included in the project, new invariants become provable. Specifically, because $yreP(t)$ in $[5 \dots \infty)$ is bounded by a negative exponential displaced by a constant, its maximum is finite, so that we can add:

INVARIANTS
$yreP(t) \leq yreP_{MAX}$

to the interface *YawCtrl_IF*, where $yreP_{MAX}$ can be calculated explicitly.

### B. More General Stability Analysis

The above analysis accurately reflected —though from a formal vantage point— the kind of evaluation that can be achieved by a simulation based approach, such as we had in Section II. In this section, we extend the formal analysis to the case of a more arbitrary yaw rate request input $yrr(t)$, provided it stays within specified bounds. We illustrate thereby the greater reach of a more symbolically based approach, in cases where the calculational challenges remain tractable.

With a relatively arbitrary $yrr(t)$, we can redo the derivation of the previous section. We arrive at an analogue of (10) in which the LHS is as before and the RHS is modified:

$$\begin{aligned} \ldots &= \frac{1}{C_K}\left(T_D\frac{d^3}{dt^3}yrr(t)+\frac{d^2}{dt^2}yrr(t)+\frac{1}{T_I}\frac{d}{dt}yrr(t)\right) \\ &\equiv inh(t) \end{aligned} \tag{12}$$

We see that the inhomogeneous term $inh(t)$ depends solely on the derivatives of $yrr(t)$.

Introducing the vector $\mathbf{stc}(t) = [stcP(t) \quad stcD(t)]^{\mathrm{T}}$ where $stcP(t) \equiv stc(t)$ and $stcD(t)$ is the time derivative of $stcP(t)$, we can write the second order ODE (12) as a first order system:

$$\frac{d}{dt}\mathbf{stc}(t) = \mathbf{A}\,\mathbf{stc}(t) + \mathbf{b}(t) \tag{13}$$

where:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -H/T_I & -H \end{bmatrix} \quad \text{and} \quad \mathbf{b}(t) = \begin{bmatrix} 0 \\ H\,inh(t) \end{bmatrix} \tag{14}$$

and $H = C_K K_P/(1+C_K K_P T_D)$, (the latter being the reciprocal of the constant appearing in (11)).

The form of (13) is standard (see, e.g. [20], [21], [22] and many other places), so the system can be integrated by applying a routine procedure. For $t \geq 5$ we have:

$$\mathbf{stc}(t) = e^{\mathbf{A}(t-5)}\mathbf{stc}(5) + \int_5^t e^{\mathbf{A}(t-s)}\mathbf{b}(s)\,ds \tag{15}$$

Since $\mathbf{b}(t)$ consists solely of derivatives of $yrr(t)$, we can integrate by parts repeatedly. To do so we introduce the notation $\mathbf{yrr}(t) = [0 \quad yrr(t)]^{\mathrm{T}}$, and we observe that:

$$\begin{aligned} &\int_5^t e^{\mathbf{A}(t-s)}\frac{d^k}{ds^k}\mathbf{yrr}(s)\,ds = \\ &\quad \left[e^{\mathbf{A}(t-s)}\sum_{j=0}^{k-1}\mathbf{A}^{k-j-1}\frac{d^j}{ds^j}\mathbf{yrr}(s)\right]_5^t \\ &\quad + \mathbf{A}^k\int_5^t e^{\mathbf{A}(t-s)}\mathbf{yrr}(s)\,ds \end{aligned} \tag{16}$$

So as not to have to deal with a large collection of boundary terms coming from (16), we now hypothesise a turning episode in which $yrr(t)$ starts at zero (for $t = 5$), smoothly increases and then smoothly decreases back to zero (for $t > 9$ say). Dropping the boundary terms, we get, in the $t > 9$ region:

$$\begin{aligned} \mathbf{stc}(t) &= e^{\mathbf{A}(t-5)}\mathbf{stc}(5) + \\ &\frac{1}{C_K}\left(\frac{1}{T_I}\mathbf{A}+\mathbf{A}^2+T_D\mathbf{A}^3\right)\int_5^t e^{\mathbf{A}(t-s)}\mathbf{yrr}(s)\,ds \end{aligned} \tag{17}$$

A result like (17) allows us to estimate in a symbolic manner the steering thrust command required for turning episodes corresponding to $yrr(t)$'s that behave in ways characterised by some generic pattern. For example we may be able to confirm that for the class of turning episodes considered, the magnitude of the steering command will not breach the physical boundaries engineered into the system.[3]

If this strategy is pursued, then the properties assumed for $yrr(t)$ can be introduced axiomatically in the interface *YawCtrl_IF*. Technically, constants would be introduced in *YawCtrl_IF*, e.g. a constant *YRR* naming a function of time, which would be endowed with the properties required to be assumed for $yrr(t)$, expressed via axioms. Then the behaviour of $yrr(t)$ would be set equal to *YRR* in machine *KurtUser_Mch*.

The properties concerning $yrr(t)$ derivable from this basis could be dealt with in various ways. For persistent properties, the most natural approach would be to recast them as invariants of the system. Properties not of this kind cold be expressed as THEOREMS in the syntax. Both kinds would then need to be proved.

### C. On Mechanical Verification

The previous two sections gave examples of what could be addressed within a formal development framework capable of treating continuous behaviour as first class citizen. But writing a desirable property is one thing, and mechanically discharging a proof of it is another. While proper mechanical support for HEB is, as yet, an aspiration, achieving the power to do the kind of mathematics indicated in a reasonable time would require the import of the capabilities of existing tools like *Mathematica* [23]. Such an approach is entirely practical, and would provide a good level of additional assurance, beyond what can be achieved by explorations of system behaviour via simulation.

For applications requiring an even higher level of assurance, the user would have to program the rules and tactics for the relevant portion of mathematics directly, so that the details of the derivation could be exposed to scrutiny, in contrast to tools like *Mathematica*, where the internal reasoning algorithms are commercial secrets. The capability to approach the verification task in both ways is part of the planned tool support for HEB.

## VI. Discretizing Hybrid Event-B Yaw Control

A major issue in turning a conceptual design into a reality in today's engineering environment, is going from the original continuous control model to a discretized control model. This is because, with today's components, analogue control is prohibitively expensive when compared to its discrete counterpart. (There are, of course, many other reasons for preferring discrete control which are well known, such as the flexibility of software, and the lack of drift in digital components.)

In some approaches, the design is initiated directly in the discrete sphere, bypassing the continuous world altogether. However, that forces the problem of deciding the sampling frequency, to be confronted immediately. The advantage of starting in the continuous world is that this issue is postponed in favour of engagement with the primary design challenges, which are most clearly viewed in the continuous world. This is what we do here, starting with the continuous model, and then contemplating the discretized version.

[3] A more realistic simulation of KURT than shown in Section II includes limiters to do just that.

```
PROJECT KurtD_Prj
REFINES –??– Kurt_Prj
INTERACES
  YawCtrlD_IF
MACHINES
  KurtUserD_Mch
  KurtD_Mch
  YawCtrlD_Mch
END
```

```
INTERFACE YawCtrlD_IF
REFINES –??– YawCtrl_IF
SEES KurtD_Ctx
TIME t
PLIANT
  $yrr_D, yrm_D$,
  $stc_D, stc_D^{pr}$,
  $yreP_D, yreP_D^{pr}$,
  $yreI_D, yreD_D$,
  $thr_D, tal_D, tar_D$
INVARIANTS
  $yrr_D, yrm_D \in \mathbb{R}, \mathbb{R}$
  $stc_D, stc_D^{pr} \in \mathbb{R}, \mathbb{R}$
  $yreP_D, yreP_D^{pr} \in \mathbb{R}, \mathbb{R}$
  $yreI_D, yreD_D \in \mathbb{R}, \mathbb{R}$
  $thr_D, tal_D, tar_D \in \mathbb{R}, \mathbb{R}, \mathbb{R}$
  $thr_D = thr$
  $yrr_D = yrr$
  $|yrm_D - yrm| < B_{yrm}$
  $|stc_D - stc| < B_{stc}$
  $|stc_D^{pr} - stc| < B_{stc}$
  $|yreP_D - yreP| < B_{yreP}$
  $|yreP_D^{pr} - yreP| < B_{yreP}$
  $|yreI_D - yreI| < B_{yreI}$
  $|yreD_D - yreD| < B_{yreD}$
  $|tal_D - tal| < B_{tal}$
  $|tar_D - tar| < B_{tar}$
...   ...
```

```
...   ...
  INITIALISATION
    WHEN
      $t = 0$
    THEN
      $yrr_D, yrm_D$ := $0, 0$
      $stc_D, stc_D^{pr}$ := $0, 0$
      $yreP_D, yreP_D^{pr}$ := $0, 0$
      $yreI_D, yreD_D$ := $0, 0$
      $thr_D, tal_D, tar_D$ := $0, 0, 0$
    END
  END
```

```
CONTEXT KurtD_Ctx
EXTENDS Kurt_Ctx
  ...   ...
AXIOMS
  $NT = 1$
  ...   ...
END
```

```
MACHINE KurtUserD_Mch
REFINES KurtUser_Mch
CONNECTS YawCtrlD_IF
EVENTS
  SteerKurt
    REFINES SteerKurt
    STATUS pliant
    BEGIN
      $thr_D(t)$ := $\Theta(4 - t)$
      $yrr_D(t)$ := $\Theta(t - 5)$
    END
END
```

```
MACHINE KurtD_Mch
REFINES –??– Kurt_Mch
CONNECTS YawCtrlD_IF
EVENTS
  KurtBehavesPli
    REFINES KurtBehaves
    STATUS pliant
    COMPLY skip
    END
  KurtBehavesMo
    STATUS ordinary
    WHEN $(\exists n \in \mathbb{N} \bullet t = nT)$
    THEN
      $yrm_D$ := $yrm_D + C_K T stc_D$
    END
END
```

```
MACHINE YawCtrlD_Mch
REFINES –??– YawCtrl_Mch
CONNECTS YawCtrlD_IF
EVENTS
  YawControlPli
    REFINES YawControl
    STATUS pliant
    COMPLY skip
    END
  YawControlMo
    STATUS ordinary
    WHEN $(\exists n \in \mathbb{N} \bullet t = nT)$
    THEN
      $yreP_D$ := $yrr_D - yrm_D$
      $yreP_D^{pr}$ := $yreP_D$
      $yreI_D$ := $yreI_D + T yreP_D$
      $yreD_D$ := $(yreP_D - yreP_D^{pr})/T$
      $stc_D$ := “$K_P[yreP_D + yreI_D/T_I + T_D yreD_D]$”
      $stc_D^{pr}$ := $stc_D$
      $tal_D$ := $thr_D - stc_D$
      $tar_D$ := $thr_D + stc_D$
    END
END
```

Fig. 6. A discretized Hybrid Event-B system for yaw control.

### A. Continuous and Discretized Systems

From a formal development standpoint, the most desirable relationship between a system model and its more idealised predecessor, is a refinement. Typically, a refinement enriches a more idealised model with detail taking it ‘closer to implementation’. The enriched model is proved consistent with its predecessor (normally, via a formal simulation relation). Done properly, a refinement has the potential to preserve valuable properties established earlier, in the new model. Unfortunately, in the context of the discretization issue, this strategy, applied naively, fails. The reasons are as follows.

A continuous description of a system contains an ‘infinite’ amount of information: i.e. the values of all system variables over a continuum of times. Any implementable sampling method will unavoidably ‘forget’ all but a tiny fraction of this information, i.e. all but the sampled values themselves. If the system response to the environment depends on the information it has about the system's behaviour, it is more or less inevitable that, in principle, the quality of a sampled system's response will be inferior compared with the continuous case. Thus the discretization process is not an enrichment of the continuous model but an impoverishment, and refinement, as a technique, struggles to cope with it, since the impoverishment degrades the information available for the consistency proof rather than enhancing it.

Still, the news is not all bad. Typically, the interaction between the system and the environment/plant is two way (closed loop). If, overall, both the continuous and discrete versions of the combined system are stable (with suitable choices of parameters etc.), then a reaction in the discretized system that is in some way undesirably increased compared to what it would be in the continuous system under similar circumstances, can be compensated for by the environment of the discretized system, which can increase suitably its input to the system to steer overall behaviour towards the desired regime. Doing this successfully depends on a number of things: good understanding of both system and environment; the deviations spoken of being moderate in magnitude; the overall system (in both the continuous and discrete versions) being stable; suitable choices of parameters being made.[4] See [16]

[4] Suitable parameter choice is heavily dependent on insight from the frequency domain. We return to this point below.

for a relevant technical discussion. However, if the interaction between the system and the environment is one way (open loop), it is much easier to see less acceptable devations.

### B. The Discretized Model

In Fig. 6 there is a discretized version of the previous continuous yaw control model. Each syntactic construct is replaced by its discretized counterpart; e.g. *Kurt_Prj* is replaced by *KurtD_Prj* which **REFINES –??–** it. The question marks qualifying the REFINES claim refer to a certain level of ignorance concerning the precision of the relationship between the continuous and discretized versions that we must endure, and that affects many components of the two models. We discuss this point in detail in Section VI-D.

In this exercise, for simplicity, we keep all the model constants (such as $C_K, K_P$ etc.) the same.[5] In addition, there is a further constant $T$, which represents the sampling period. For simplicity, $T$ is axiomatized to be $1/N$'th of a unit of time, so that the external stimuli to the system can remain the same as in the continuous model (and both, moreover, are open loop).

Variables $var_D$ are the discretized counterparts of their earlier predecessors $var$, sampled and updated every $T$ time units. Now, the semantics of Hybrid Event-B imposes a specific interpretation on the assignments that occur in mode events, e.g. $var_D := expr(var_D)$. When such an assignment is executed at a time $kT$ say, the LHS of the assignment denotes the new value $var_D(kT)$, which we write as $var_{D,k}$. However, the RHS is evaluated using the limiting value of $var_D$ just before $kT$. If we assume that $var_D$ does not change during any sampling interval, then the RHS is in fact $expr(var_{D,k-1})$, so the assignment implements the difference equation $var_{D,k} = expr(var_{D,k-1})$. We return to this point below.

In order to implement simple approximations to derivatives and integrals (done via backward differences and accumulated sums, respectively), preceding values of some variables need to be recorded: $var_D^{pr}$. (As with the model constants, the literature contains many approaches that tackle these issues in more sophisticated ways; see e.g. [24], [25], [26].)

We discuss the machines, one by one. The simplest is *KurtUserD_Mch*. This genuinely REFINES the earlier *KurtUser_Mch* machine in a manner which is easy to see. Namely, the original and discretized variables $thr$ and $yrr$ vs. $thr_D$ and $yrr_D$ have identical behaviours in the sole (pliant) event of the two machines *SteerKurt*. This is formalised via the equalities $thr_D = thr$ and $yrr_D = yrr$ in the invariants of the interface.

Next we have *KurtD_Mch*. This has both a pliant event *KurtBehavesPli* and a mode event *KurtBehavesMo*. The pliant event (continuously) skips. This models the zero order hold that characterises a simple sampling scheme. The pliant event REFINES –??– the *KurtBehaves* event of *Kurt_Mch*. The mode event models the periodic updates to $yrm_D$ at the sampling times, obtained by replacing the differential equation of the *KurtBehaves* event with a discretized approximation of the corresponding integral equation via $yrm_D := yrm_D + C_K T stc_D$. This is to be interpreted as discussed above, which means that the assignment represents the difference equation $yrm_{D,k} = yrm_{D,k-1} + C_K T stc_{D,k-1}$.

Then we have *YawCtrlD_Mch*. The conventions already described hold here too. Thus there is a pliant event *YawControlPli* that skips while it REFINES –??– the *YawControl* event of *YawCtrl_Mch*, and a mode event *YawControlMo* that models the periodic updates to the discretized counterparts of all the variables modified by *YawControl*. At this point a subtlety needs to be pointed out.

In a pliant event, there is no difference between the (parallel) direct assignments $x, y := y, z$ and $x, y := z, z$ because of the equality semantics of direct (instantaneous) assignment. However, when the two assignments are naively discretized, they turn into $x_D, y_D := y_D, z_D$ and $x_D, y_D := z_D, z_D$ respectively, which are to be interpreted as we discussed above. The first of these corresponds to the difference equation $x_{D,k}, y_{D,k} = z_{D,k-2}, z_{D,k-1}$, because the LHS and RHS of such assignments refer to values one sampling period apart, as noted above. Thus a chain of $n$ dependent equalities in a pliant event —which in the pliant event relate values at the same time point— can generate an $n$'th order difference equation upon discretization. This can have detrimental effects on the quality of the approximation and on its stability due to the use of older and older values — as is discussed extensively within numerical analysis, e.g. [27], [28].

In order to minimise the impact of this, we can back substitute to make use of values that are as fresh as possible.[6] Note that every different choice of scheme for doing the back substitutions results in a discretization scheme that is correspondingly different, amounting to a different design decision regarding what discretization means.

We can go further, designing the difference equations that we wish to use for the discretization *a priori*, and independently of the 'obvious' discretizations of the continuous model, and then work back to derive the discrete assignments that would implement them.

In the context of these remarks, the main impact on the *YawCtrlD_Mch* machine is to alter the detailed expressions that occur on the RHS of the assignments of the mode event. The assignment that is of most interest is the assignment for $stc_D$, where the feedback from the control strategy is most felt. Its RHS is enclosed in heavy quotes to allude to this.

### C. Discretized Stability Analysis

Let $k \in \mathbb{N}$ index the number of $1/N$'ths of a time unit elapsed since $t = 5$. We address the discretization of $stc_D$ in more detail. In the light of all the possibilities just discussed, we proceed as follows.

Looking at the assignments for $yrm_D$ and $yreP_D$ that appear in Fig. 6 and dropping the inhomogeneous terms, it is not difficult to derive the difference equation $yreP_{k+1} = -C_K T \sum_{r=0}^{k} stc_{D,r}$. Deriving the analogous expression for $yreI_D$ involves a double summation. Dealing with these directly in the assignment for $stc_D$ is certainly inconvenient. However, if we take second differences of the $stc_D$ assignment, the summations

[5] In many discretization approaches, model constants are adjusted, in order to better approximate the continuous model.

[6] This amounts to using the $x_D, y_D := z_D, z_D$ form in the earlier example.

cancel, and we obtain:

$$\begin{aligned} stc_{D,k+3} - 2stc_{D,k+2} + stc_{D,k+1} \; = \\ -C_K K_P [T_D(stc_{D,k+2} - 2stc_{D,k+1} + stc_{D,k}) \\ + T(stc_{D,k+2} - stc_{D,k+1}) + T^2 stc_{D,k+2}/T_I] \quad (18) \end{aligned}$$

which is much more amenable to analysis. Inserting the ansatz $stc_{D,k} = RW^k$ into (18) yields:

$$\begin{aligned} D(W) \; \equiv \; W^3 + C_K K_P [T^2/T_I + T + T_D - 2/C_K K_P] W^2 \\ + C_K K_P [1/C_K K_P - 2T_D - T] W + C_K K_P T_D \; = \; 0 \quad (19) \end{aligned}$$

This is a cubic equation for $W$. For stability in the system as a whole, we need $|W| < 1$ for all the solutions of $D(W) = 0$. Standard resources for cubics such as e.g. [29], [30], show the technical burden of trying to analyse this directly.

We observe that because of the sign of the $W^3$ term in (19), if all the roots of $D(W) = 0$ are real and of modulus $< 1$, then (1) $D(+1) > 0$, (2) $D(-1) < 0$, and if $D'$ is the derivative of $D$, then (3) the roots $r_\pm$ of $D'(W) = 0$ satisfy $-1 < r_- < r_+ < +1$, (4) $D(r_-) > 0$, (5) $D(r_+) < 0$. Since $D'(W) = 0$ is a quadratic it is a lot easier to handle. (We note that the constraints cited can be related to the Sturm technique for finding regions containing real roots of an arbitrary polynomial [31].)

Although the general form of the coefficients of (19) makes it cumbersome to test for the conditions (1)-(5) in full generality, we note that we are predominantly interested in the region $T \to 0$. If any necessary conditions do not hold in the limit of vanishing sampling interval, then they cannot be of interest for any engineering purpose. The $T \to 0$ limit simplifies the coefficients considerably.

Beyond this, we can rely on the physical properties of the system to simplify the case analysis further. From the Cohen-Coon tuning analysis of Section II, it emerges that $T_D \approx 10T$ and $T_I \approx 4T_D$. As well, the kinematics of the problem mean that $C_K$ is positive, and it also follows that $K_P$ is positive. So all the constants in our problem space are positive.

All of these observations make the corresponding tests relatively straightforward to carry out. Tests (1) and (2) are straightforward evaluations. The former yields a triviality while the latter yields the constraint:

$$1 > C_K K_P T_D \qquad (20)$$

which turns out to be necessary for the small $T$ limit to be feasible. Constraint (3) gives rise to a number of further conditions. However, in the small $T$ limit, they are all subsumed by the stronger condition (20).

The expressions for the roots $r_\pm$ of $D'(W) = 0$ are the standard formulae for a quadratic, and turn out to be expressions in the combination $C_K K_P T_D$. Accordingly, it is easiest to use (20) to substitute numerical values for $C_K K_P T_D$ and thence to check conditions (4) and (5). It turns out that values of $C_K K_P T_D$ around approximately 0.5 permit (4) and (5) to be satisfied in the small $T$ limit. We thus conclude that there is a stable regime in the small $T$ region, and hence furthermore, that there is a still larger region of stability when a pair of roots of $D(W) = 0$ fuses and bifurcates into a pair of complex roots (which will be close to the real axis for some range of values of the parameters).

We concede that the above analysis was somewhat *ad hoc*. More significantly, it was purposely confined entirely within the state space formulation of the problem. This is important to the extent that the HEB approach is lodged in the state space domain for reasons which were explained in the Introduction.

By contrast, most discretizations of continuous designs in engineering practice take place within the frequency domain. As mentioned previously, there are various approaches described in the literature cited earlier. Happily, one of them coincides with what we derived: the discrete equivalence approach. In that approach, a zero order hold is introduced into the model at the right point, standard $z$-transform elements are introduced for the PID components, and a transfer function is calculated by combining all of these. The poles of the transfer function give the characteristic frequencies of the system, which are checked for stability. It turns out that the denominator of the transfer function (which is a rational function in the $z$ plane), coincides exactly with (19) aside from the change of variable.

In the conventional approach to discretization, stability is analysed using the Jury test [32], [26], [8], which is applied in the $z$ domain. This generates a sequence of tests, all of which have to be passed to deduce stability (which is the property $|z| < 1$ for the characteristic frequencies). Happily once more, the first few of these coincide with the first few of the *ad hoc* tests we did above. All of this shows not only the desirability, but the feasibility of greater cooperation between the two formulations of control within the formal HEB approach.

### *D. Relating the Continuous and Discretized Models*

The cornerstone of any formal development technique like (Hybrid) Event-B is the idea of relating successive models via a formal refinement relation, which relates a more abstract model to a more concrete one. In practice this is always a simulation relation, which amounts to the statement: *IF* the invariants hold at a given moment *THEN* they hold after any update (mode or pliant) of the concrete variables — for a suitable choice of update of the abstract variables. Thus, suitably initialised, the implicational structure can be cascaded inductively into a statement that holds true at all times. Evidently, for the described approach to have force, the invariants mentioned must express a desired relationship between the two families of variables that is also expected to hold at all times.

Establishing this, when viewing discretization as an instance of refinement, proves to be very demanding in all but the simplest cases. A trivial case of discretization treated this way occurs in [9] — no technical difficulties occur there. From our own development, an equally trivial case of refinement occurs between machines *KurtUser_Mch* and *KurtUserD_Mch*, since in the INVARIANTS section of the *YawCtrlD_IF* interface we find the joint invariants $thr_D = thr$ and $yrr_D = yrr$, which express the equality of the relevant pairs of variables. Since the original and discretized variables are defined to behave in exactly the same way in their respective machines, establishing the required properties is indeed trivial, and *KurtUserD_Mch* is a genuine refinement of *KurtUser_Mch*.

For the other machines in the model, the situation is a lot less clear cut. The detailed behaviour of the continuous and discretized variables in each corresponding pair is known a lot less precisely. In particular, there is a significant lack of detail compared with what is stated regarding *thr* and *yrr* and their counterparts. A number of points arise concerning this.

Firstly, given a linear and third order discretization scheme, with sufficient additional effort an analytic solution could in principle be obtained for the various variables involved, at least in the form of summations over powers of the roots. The effort involved would be considerable, yet the perspicacity of the solution obtained would not be a given. With a relatively opaque formulation of the solution, its relationship with the analytic solution to the continuous system (which we did not pursue to its conclusion either) would also not be clear. This would lead to further technical difficulties in formulating relevant joint invariants for corresponding pairs of variables. It is not evident that they would enjoy the same level of precision that we were able to indicate for *thr* and *yrr*.

Secondly, the effort expended in achieving the goals indicated (if actually expended) would only apply for the single pair of abstract and concrete system trajectories given by the specific driving inputs specified by *thr* and *yrr* in *KurtUser_Mch* and its counterpart. They would not necessarily apply for any other inputs. But, for dependability, we would want a generic result that applied to a whole family of trajectories that was large enough to include all that could be expected to arise in practice. For that, much more generic and powerful results would be needed. And this mismatch between what typical formal techniques routinely demand, and what the analysis of control systems can routinely supply, rapidly grows greater the more complex the application system becomes.

Based on these observations, the relationships that we have quoted in the *YawCtrlD_IF* interface between continuous and discretized versions of our variables (excepting *thr* and *yrr*) is approximate equality. Some justification for this lies in the fact that we were able to show that both systems were stable (for suitably chosen static parameters).

For us, approximate equality meant $|var_D - var| < B_{var}$, for corresponding variable pairs and constants $B_{var}$. When the stability properties are sufficiently strong, such results can become provable in principle; see e.g. [16]. Still, it is not unquestionably the case that these results can be related directly to problem domain quantities, because of their reliance on existential properties.

The claim that $|var - var_D| < B_{var}$ can serve as a suitable joint invariant depends crucially on knowing that the dynamics is such that the difference between continuous and discretized versions of variables always strictly decreases, i.e. that we always have behaviour consistent with asymptotic stability, *regardless of the behaviour of any external input*. However, in Section VI-A, we acknowledged the possibility of impoverished information from sampling allowing the discrepancy between variables to grow, even if temporarily. In such cases, the arguments in [16] will not hold.

In cases like that, it is often not too hard to prove *local properties*, i.e. properties concerning a single sampling interval, that assert not-too-bad divergence. The downside of such properties though, is that they do not cascade inductively into statements that hold at all times.

A variation on the refinement technique that is focused on such local properties is retrenchment [33], [34], [35], [36]. It provides a formal framework in which such local properties can be expressed, and additionally, related to properties controlled by refinement (see particularly [35]). A judicious combination of refinement and retrenchment could prove to be the best approach to the technical difficulties mentioned.

Of course, we are not the first authors to discuss the challenges posed by discretization. As well as standard discrete control references such as [32], [26], [25], [37], [8], there are more specialised treatments aimed at specific aspects, e.g. [38], [39], [40], [41]. Frequently they focus on a frequency domain approach or on statistical properties. It is probably fair to say though, that there is no detailed treatment that is an ideal match to the needs of a formal approach to control systems design and development.

## VII. Conclusions

In the preceding sections, we took a simplified though non-trivial version of the yaw control problem for the KURT e-vehicle, and used it as a testbed for complementing and strengthening the assurance obtainable via conventional engineering approaches, by using a formal modelling and refinement approach to the development. The vehicle for the latter was the Hybrid Event-B formalism [9], [10], a formalism designed to capture hybrid system behaviour in way that is compatible with established formal development approaches from the computing sphere, and with state based approaches from the control sphere.

One immediate consequence of this is the evident tension between the state based perspective of formal approaches (which need to deal with arbitrarily structured state spaces) and the frequency domain based perspective of conventional engineering design. The frequency domain approaches, typically based on Laplace transforms and *z*-transforms, turn functional properties into algebraic ones — the latter are usually much easier to manipulate in practice, and thus are strongly favoured in practical engineering. The simplification of the design process coming from the use of algebraic techniques is amplified by the use of specific stimuli (such as response to step function inputs), and of simulation, as preferred techniques to gauge the appropriateness of a design. Of course, to quote a familiar truism, simulation can only show the presence of faults, not their absence, so enhancing the development methodology with formal techniques which potentially have broader coverage is a worthwhile aim.

Doing this seriously though, quickly invites back the technical obstacles that the algebraic and simulation based approaches strove to avoid. We rapidly saw this in our examination of the discretization of PID control in the KURT e-vehicle application — the technical difficulties inherent in moving from a relatively perspicuous continuous design to a suitable discretized version quickly proliferated.

These observations readily propose a series of topics that it would be very worthwhile to develop more fully in order to increase the utility of a combined approach.

[I] A closer cooperation is needed between, on the one hand, the state space methods typical in state based control and formal approaches, and on the other hand, the frequency domain based approaches so widely used in conventional engineering. Normally, discussions take place exclusively within one domain or the other, but a greater interaction could improve the feasibility of a combined approach. For this, and depending on the specific goals of a given analysis, more contact points would be identified between state space and frequency methods, such that the solution to a state space question could be derived in the frequency domain. However, it has to be appreciated that only certain questions translate well between the two domains.

[II] With issues such as discretization, it is clearly impractical to tackle each development from first principles in a practical methodology. A family of suitable generic results is needed that can be applied in a wide variety of contexts to inform the formal strand of the development. Similar remarks apply to other aspects of the development methodology — we could mention topics such as sensitivity and robustness, besides stability, which we concentrated on.

[III] To facilitate the efficient incorporation of the relevant kinds of formal derivation (done by hand in an *ad hoc* manner in this paper) additional special purpose syntactic support could be provided within the Hybrid Event-B formalism, especially in the context of mechanised tool support.

Besides these issues concerned with the technical interaction between different approaches, lies the application level scope of the formal modelling and refinement. We confined our attention to stability in an ideal environment, thereby neglecting many things. In reality, friction and sensitivity to external influences in the equipment affect the behaviour of the system. If these are not modelled properly, then the rigour of a formal refinement becomes spurious. Likewise statistical fluctuations in both the equipment and the environment of operation impacts the behaviour, and needs to be taken into account.

The traditional engineering approach to such questions is again via the frequency domain. It is assumed that influences such as these are each characterised by a suitable frequency domain response profile, often determined experimentally where needed. With a characterisation of that kind to hand, the controller can be designed to filter out stimuli that are undesired, and to respond appropriately to those that are important. With such a bandwidth limited frequency domain controller design, the Nyquist theorem and engineering heuristics give a good guide to the sampling frequency needed for a dependable discretization. However, this kind of frequency domain derivation is rather far from the state space approaches that could be directly placed in a formal development framework. All of these topics provide fertile territory for further work, to be pursued in other publications.

## REFERENCES

[1] K. Ogata, *Modern Control Engineering*. Pearson, 2008.

[2] R. Dorf and R. Bishop, *Modern Control Systems*. Pearson, 2010.

[3] K. Dutton, S. Thompson, and B. Barraclough, *The Art of Control Engineering*. Addison Wesley, 1997.

[4] Modelica Homepage, https://www.modelica.org/.

[5] J. Van Leeuven, *Handbook of Theoretical Computer Science, Vol. A and Vol. B*. Elsevier, 1990.

[6] E. Sontag, *Mathematical Control Theory*. Springer, 1998.

[7] N. Ahmed, *Dynamic Systems and Control With Applications*. World Scientific, 2006.

[8] D. Hinrichsen and A. Pritchard, *Mathematical Systems Theory I*. Springer, 2005.

[9] R. Banach, M. Butler, S. Qin, N. Verma, and H. Zhu, "Core Hybrid Event-B I: Single Hybrid Event-B Machines," *Sci. Comp. Prog.*, 2015, to appear.

[10] R. Banach, M. Butler, S. Qin, and H. Zhu, "Core Hybrid Event-B II: Multiple Cooperating Hybrid Event-B Machines," 2015, submitted.

[11] K. Kozłowski and D. Pazderski, "Modeling and Control of a 4-Wheel Skid-Steering Mobile Robot," *Int. J. Appl. Match Comput. Sci.*, vol. 14, pp. 477–496, 2004.

[12] ControlsWiki, https://controls.engin.umich.edu/wiki/index.php/PIDTuningClassical.

[13] K. Astrom and T. Hagglund, *Advanced PID Control*. ISA, 2006.

[14] K. Ogata, *System Dynamics*. Pearson, 2013.

[15] J. R. Abrial, *Modeling in Event-B: System and Software Engineering*. Cambridge University Press, 2010.

[16] P. Tabuada, *Verification and Control of Hybrid Systems: A Symbolic Approach*. Springer, 2009.

[17] A. Platzer, *Logical Analysis of Hybrid Systems: Proving Theorems for Complex Dynamics*. Springer, 2010.

[18] L. Carloni, R. Passerone, A. Pinto, and A. Sangiovanni-Vincentelli, "Languages and Tools for Hybrid Systems Design," *Foundations and Trends in Electronic Design Automation*, vol. 1, pp. 1–193, 2006.

[19] S. Hallerstede and T. Hoang, "Refinement by Interface Instantiation," in *Proc. ABZ-12*, Derrick, Fitzgerald, Gnesi, Khurshid, Leuschel, Reeves, Riccobene, Ed., vol. 7316. Springer, LNCS, 2012, pp. 223–237.

[20] W. Walter, *Ordinary Differential Equations*. Springer, 1998.

[21] C. Chicone, *Ordinary Differential Equations with Applications*, 2nd ed. Springer, 2006.

[22] P. Antsaklis and A. Michel, *Linear Systems*. Birkhauser, 2006.

[23] Mathematica Homepage, http://www.wolfram.com.

[24] J. D'Azzo and C. Houpis, *Linear Control System Analysis and Design: Conventional and Modern*. McGraw Hill, 1995.

[25] G. Franklin, J. Powell, and M. Workman, *Digital Control Systems*. Prentice Hall, 1996.

[26] P. Paraskevopoulos, *Digital Control Systems*. Prentice Hall, 1996.

[27] E. Isaacson, *Analysis of Numerical Methods*. Dover, 2003.

[28] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, 1996.

[29] Wikipedia, "Cubic function."

[30] F. Olver, D. Lozier, R. Boisvert, and C. Clark, *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[31] Wikipedia, "Sturm's theorem."

[32] B. Kuo, *Digital Control Systems*. Oxford University Press, 1992.

[33] R. Banach, M. Poppleton, C. Jeske, and S. Stepney, "Engineering and Theoretical Underpinnings of Retrenchment," *Sci. Comp. Prog.*, vol. 67, pp. 301–329, 2007.

[34] R. Banach, C. Jeske, and M. Poppleton, "Composition Mechanisms for Retrenchment," *J. Log. Alg. Prog.*, vol. 75, pp. 209–229, 2008.

[35] R. Banach and C. Jeske, "Retrenchment and Refinement Interworking: the Tower Theorems." *Math. Struc. Comp. Sci.*, vol. 25, pp. 135–202, 2015.

[36] Retrenchment Homepage, http://www.cs.man.ac.uk/retrenchment.

[37] M. Fadali and A. Visioli, *Digital Control Engineering: Analysis and Design*. Academic Press, 2009.

[38] B. Widrow and I. Kollar, *Quantization Noise*. Cambridge University Press, 2008.

[39] R. Marks, *Introduction to Shannon Sampling and Interpolation Theory*. Springer, 1991.

[40] R. Pytlak, *Numerical Methods for Optimal Control Problems with State Constraints*, ser. Lecture Notes in Mathematics. Springer, 1999, vol. 1707.

[41] L. Grune, *Asymptotic Behavior of Dynamical and Control Systems under Perturbation and Discretization*. Springer, 2002.

DOI: 10.15439/2015F190
ACSIS, Vol. 5

# Behavior-Preserving Abstraction of ESTEREL Programs

Nir Koblenc
Department of Mathematics and Computer Science
Open University of Israel
Ra'anana, Israel
Email: skoblenc@gmail.com

Shmuel Tyszberowicz
School of Computer Science
Academic College of Tel-Aviv Yaffo
Tel-Aviv, Israel
Email: tyshbe@tau.ac.il

***Abstract*—Reactive programs often control safety-critical systems, thus it is essential to verify their safety requirements. ESTEREL is a synchronous programming language for developing control-dominated reactive systems, and XEVE is a verification environment that analyzes circuit descriptions generated from ESTEREL programs. However, a circuit generated by the ESTEREL compiler from non-pure ESTEREL program often displays behaviors which may violate safety properties even when the source program does not. We introduce an automatic abstraction process for ESTEREL programs developed to tackle this problem. When the process is applied to a program augmented with observers to monitor the program's behavior, it results in a pure program that preserves the behavior of the source program, replacing value-carrying objects with pure signals. We have built a prototype tool that implements the abstraction and used it to purify control programs and robotic systems.**

***Index Terms*—Verification, Abstraction, Reactive Systems, ESTEREL.**

## I. INTRODUCTION

*Reactive systems* are computer systems that continuously react to their environment at a speed determined by the environment. Most industrial real-time systems are reactive [3]. ESTEREL[1] is an imperative concurrent language for the development of industrial-strength reactive systems, which is especially well-suited for control-dominated reactive systems such as real-time process control systems, embedded systems, communication protocols, peripheral drivers, human-machine interfaces, and others [4]. ESTEREL belongs to the family of *synchronous languages*— languages that are based on the synchrony hypothesis, which states that a program instantaneously reacts to its input. Control is assumed to takes no time and thus output is broadcast right when the input arrives. The notion of simultaneity is captured by the concept of *event*, which is a set of simultaneous occurrence of (possibly valued) signals [3]. ESTEREL offers significant advantages over traditional languages used in industrial settings [5], such as verification (due to the precise mathematical semantics and the existence of verification tools and techniques), reduction of testing (automatic verification covers safety requirements), high-level abstraction, and better code structuring. A full definition of the language, as of version 5.91, can be found in [4].

ESTEREL programs communicate with their environment by means of signals and sensors. Signals can be used both for input and output, and may convey values; sensors can only be input and always convey values. A signal can be either present or absent; no such concept exists for sensors. A signal that carries values is called a *valued signal*, and a signal that does not convey values is called a *pure signal*.

Reactive systems are often used to control safety-critical systems. Hence they require rigorous design methods, and formal verification must be considered [3]. A complete, consistent and precise specification is constructed, often employing formal language to avoid ambiguity. While this process is very potent in detecting errors already at the formal specification development phase [6], there still is a risk that there would be inconsistencies between the formal specification and the eventual product. Even while the specification is formally verified, the product itself may still be erroneous, and we want to verify that the program satisfies its safety properties.

*Verification by observers* [7] is an approach to verify code. *Observers* are program modules monitoring the program, testing that a property is satisfied and broadcasting specific signals when the property is violated. The observers are composed in parallel to the original program, and the resulting program is compiled using an ESTEREL compiler into a finite automaton. The properties of the automaton are verified using tools such as the X ESTEREL VERIFICATION ENVIRONMENT (XEVE) [7], reducing the verification problem to reachability problem in finite automata – finding if there exists an execution trace from the initial state to a state emitting one or more of those special observer signals.

The XEVE verification environment requires the program to be compiled into Berkeley Logic Interchange Format (BLIF), a logic-level hardware hierarchical circuit description in a textual form; however, ESTEREL compiler, as of version

This work has been partially supported by GIF (grant No. 1131-9.6/2011). The paper is based on the master's thesis of the first author. For the thesis and the prototype tool see [1]. This paper extends a very short paper we have published in [2].

[1] We refer to ESTEREL v5.92, which we use for teaching. The toolset and the documentation are available at http://www-sop.inria.fr/esterel.org/filesv5_92/.

5.91[2], can either compile pure ESTEREL programs into BLIF files without changing their semantics, or, using the `-soft` option, abstract the data and compile only the control aspect into BLIF [9]. Pure ESTEREL programs only handle pure signals, i.e., they involve no valued signals, types, constants, functions, procedures, tasks, or variables [9]. In this work we collectively refer to valued signals, sensors, and variables as *valued objects*. The problem is that XEVE may fail to verify properly circuits generated from observed programs in which data is involved in the control when using the `-soft` option. The reason is that the abstraction might add behaviors not displayed by the original program, possibly including behaviors in which observer signals are emitted. Based on the false observation, the user might reject a program which is actually correct. For example, consider the following program:

```
module SpuriousError
   output Error;
   var v := 1 : integer in
      if (v <> 1) then
         emit Error
      end if;
      pause
   end var
end module
```

This program declares a variable `v` and initializes it to 1. It is obvious that the signal `Error` is never emitted because the condition tested by the `if` statement is never fulfilled. We can compile this program into a BLIF file using the `-soft` option, yet XEVE suggests that `Error` is possibly emitted, probably since the role of the data in the control flow is ignored and every transition dependent on run-time values is always enabled.

Many control schemes, however, receive numerical inputs, conduct numerical calculations, and emit numerical outputs. Example for such schemes are signal processors and closed-loop feedback controllers. We *purify* such programs to allow automatic verification of their properties. This is a transformation of ESTEREL programs handling valued objects into pure ESTEREL programs. It abstracts an unbound, concrete system that handles data by replacing objects that take values from theoretically-infinite domains with pure signals to receive a finite system. The abstraction preserves the external observable behavior of the original program, i.e. there is a correspondence in terms of inputs, outputs, and timings between the two programs. The difference between the two programs is that the abstract program uses pure signals where the original one employs valued signals, sensors, and variables.

Our approach is largely based on *predicate abstraction* [10]. Predicate abstraction is an automatic mapping of an unbounded system (the *concrete system*) to a finite system (the *abstract system*). An abstract system is defined by a concrete system and a finite set of predicates. Its states correspond to truth assignments to these predicates. The predicates define the *abstraction function*, which maps the states of the concrete system to the states of the abstract system. A state of the abstract system is reachable if it is an abstraction of a reachable concrete state. A user who wants to prove certain invariants supplies them as part of the predicate set. If the predicates stating the invariants are true in all reachable abstract states, then it means that the invariants hold in any reachable state of the concrete system. Applying this idea to ESTEREL, we automatically derive predicates about current and previous values of valued objects, of the form "the value of object $x$ is in the range $I$". Each such predicate is translated to a pure signal. We can say that a state of the concrete system where such predicate is held is abstracted to a state of the purified program where the corresponding signal is present.

Usually, abstraction-based proof techniques are sound but not complete, since the abstraction is done such that every property proven to be satisfied by the abstract system has a concrete version which holds on the concrete system, yet the other way around is unnecessarily true [10]. However, verification using our abstraction technique is both sound and complete, since the abstract program adds no new behaviors to those displayed by the concrete program; in particular every pure signal is emitted by the observer-augmented concrete program if and only if it is emitted by its purified version. The abstraction alters the semantics of a few operations; however, these changes remain internal, i.e. the interaction with the environment is unaffected.

We suggest an algorithm that automatically purifies a certain group of programs and we characterize the class of programs verifiable using our technique. The main contribution of this work is extending the class of programs verifiable using observers with XEVE. We have implemented a prototype tool that purifies ESTEREL programs based on the algorithm described in Section II. This section also presents a running example of a *proportional controller*. We discuss the challenges we have faced when abstracting variables in Section III. The programs to which the method is applicable must comply with certain constraints, which are discussed in Section IV. Section V characterizes the class of programs to which the solution can be applied and provides test cases. We conclude and provide suggestions for future work in Section VI.

## II. ESTEREL PROGRAM PURIFICATION

The XEVE verification environment takes as input BLIF files. Compiling an ESTEREL program into this format preserves the program's behavior only if the source code is pure. Hence, we have to substitute valued objects with pure signals and to modify the statements controlling the flow of the program and manipulating these objects to use pure signals instead. In this section we explain how to transform a program that complies with certain necessary constraints, into a pure program preserving the behavior of the original program.

We describe only the fundamentals of the algorithm.[3] We use a running example to demonstrate our method. It shows the application of our method to a *proportional controller*. This

[2] We chose to focus on ESTEREL v5 since it is free, suitable for teaching, and can be verified using the free XEVE tool which is part of the ESTEREL v5_92 distribution, whereas ESTEREL v7 [8] is commercial. XEVE is used in the industry [7].

[3] Due to lack of space we omit some details. The full description can be found in [1].

is a closed-loop feedback controller whose control signal is proportional to the *error* – the difference between the *set point*, also known as the *reference* (the ideal point) and the measured quantity under control, i.e., the control signal is calculated by multiplying the *error signal* by a *gain* [11]. Following is the ESTEREL code before its purification:

```
module PropMotor:
   input SampleTime;
   sensor Speed : double;
   output MotorForce : double;
   output AC_ON, AC_OFF;
   loop
      % proportional controller for the motor force
      present SampleTime then
         emit MotorForce (3.0 * (100.0 - ?Speed))
      end present;
      await SampleTime
   end loop ||
   loop
      % bang-bang controller for the cooling sub-system
      present MotorForce then
         if ?MotorForce > 270.0 or
            ?MotorForce < -270.0 then
            emit AC_ON
         else
            if ?MotorForce > -30.0 and
               ?MotorForce < 30.0 then
               emit AC_OFF
            end if
         end if
      end present;
      await tick
   end loop
end module
```

In addition to the proportional controller, the example uses a *bang-bang controller* that is composed in parallel and controls a cooling system. Bang-bang controllers are feedback controllers that switch abruptly between two states [11]. The controller receives a measured quantity of interest, and outputs a certain value if that quantity is above a certain threshold, and a different value otherwise.

The example displays a control system for a motor. The set point is 100 km/h; the measured quantity is the motor's current speed, received by the `Speed` sensor; the gain is 3.0. The output signal, `MotorForce`, is the force that the motor should produce, and is obtained by multiplying the difference between 100 km/h and the current speed by the gain. The motor is cooled by an air conditioning unit when exerting more than a certain amount of force (270.0). The air conditioning stops when the amount of force exerted by the motor drops below a certain threshold (30.0) (the difference between the thresholds is deliberate and used for hysteresis[4]). To start the air conditioning, it emits the `AC_ON` signal, and to stop the air conditioning, it emits `AC_OFF`.

We want to verify several safety properties of the given program using observers that monitor the system's behavior and emit special signals once one of these properties is violated. The observers are assembled in threads parallel to the main program. We would be able to verify that these signals are never emitted by compiling the observer-augmented program into BLIF format; however, we cannot do so as long as the program handles data other than pure signals.

The algorithm removes from the program all variables, sensors, and valued signals. Instead, we represent their values using pure signals. Each Boolean-valued object needs only one signal – that we call a *value signal* – to represent its value (present when true, absent when false). Since a Boolean signal actually has three states (absent, present and true, present and false) it takes another signal, which we call a *presence signal*, to denote whether the simulated Boolean signal is considered present or absent.

As for numerical-valued objects whose values range over infinite domains, we partition their domains into non-overlapping intervals, which we hereby call *ranges*, in such way that two goals are achieved. The first one is being able to decide any condition containing an occurrence of some numerical valued object by knowing the range within which that object's value resides. For example, the condition $?s > 3.0$ for a sensor $s$ ($?s$ is the current value of $s$) can be decided if the information whether $?s \in (-\infty, 3.0]$ or $?s \in (3.0, +\infty)$ is available. The second goal is maintaining relationships between dependent objects. For instance, for the assignment $v := a * ?x + b$, where $v$ is a variable, $x$ is a valued signal and $a$ and $b$ are literal constants, we can determine the range within which the $v$'s value would reside following the assignment based on the range of $x$. E.g., for $v := 2 * ?x + 1$ where $?x$ is in (1,3], the value of $v$ is in $(2 \cdot 1 + 1, 2 \cdot 3 + 1] = (3, 7]$. Note that actually we calculate the ranges for $x$ given the partition for $v$.

Our prototype tool works in two stages that run in a sequence by a shell script. In the first stage a standalone tool parses the source ESTEREL program according to v5_91 grammar specification found in [4] and outputs an intermediate file containing a list of syntax rules it identifies. The second stage is another program that reads that file and constructs a parse tree representation of the program. The tree data structure holds the information about all objects and statements of the source program, such that it is possible to reconstruct the program entirely from the tree. During the construction of the tree items that are of interest to the abstraction process, such as valued objects or statements that manipulate or test data are identified and stored in collections from which the abstraction algorithm can efficiently access them later. The abstraction is performed on the tree. The tool automatically calculates the predicates and performs the abstraction, implementing the algorithm discussed in this work. At the final step, a pure, abstract program is written to a target file based on the transformed tree. For sake of readability, we have edited in the paper the code produced by the tool.

After the source program is parsed, the abstraction algorithm starts with a pre-processing step to simplify the program when needed. Currently it expands each `sustain` statement (a statement emitting a specified signal in every instant once started and remains active forever) for a valued signal by a loop in which that signal is emitted in every instant, such that we can handle these statements just like instantaneous signal emissions (`emit` statements). This is also the place to perform other pre-processing activities; for example, for the simplicity

[4] Hysteresis can be introduced by setting "dead zones" of no reaction around the set point in bang-bang controllers to avoid rapid on-off cycling [11].

of the algorithm, we require the user to guarantee a few preconditions the source program must fulfill (see Section IV) in order to apply the tool/algorithm to it; these assumptions can be obtained automatically by performing some pre-processing steps on the original program. This step can be expanded in future versions to include them as well.

We implement some interval-scalar arithmetic operations required by this algorithm. Let *I* be an interval, and *a* and *b* be scalars:

- Multiplication of an interval by a scalar: $a \cdot I$ (equivalent to scaling an interval), and
- Addition of a scalar to an interval: *I* + *b* (equivalent to translating an interval).

Additionally, we define a *product* of two partitions $P_1$ and $P_2$ as $\{I_1 \cap I_2 | I_1 \in P_1 \wedge I_2 \in P_2\}$. This is a "mutual refinement" of $P_1$ and $P_2$ by one another. We denote the partition of the domain of a numerical valued object *x* by PARTITION(*x*). The partitioning process starts from statements that assign a constant value to a valued object (i.e. *var* := *const* or emit *val_sig* (*const*) ). The object's domain is partitioned according to the assigned constant. Suppose, e.g. that at some stage of the partitioning process PARTITION(*v*) = {(–∞, 1), [1, 1], (1, +∞)} for a variable *v*, once considering an assignment *v* := 2.0, PARTITION(*v*) is refined to {(–∞, 1), [1, 1], (1, 2), [2, 2], (2, +∞)}. This refinement allows us to refer later to *v* = 2 (equivalent to *v* ∈ [2, 2]) as a predicate in our abstract system, which is true once the assignment takes place in the original program. Note that this is the product of {(–∞, 1), [1, 1], (1, +∞)} and {(–∞, 2), [2, 2], (2, +∞)}.

Next, we consider the Boolean data expression in which a valued object occurs. Let (*a**D*+*b*) *R c* be a Boolean data expression, where *a*, *b* and *c* are literals ($a \neq 0$), *D* is either the current or the previous (i.e. pre(?*x*)) value of *x*, all have the same data type, and *R* is a relational operator. We denote $k = \frac{c-b}{a}$. If *a* divides $c - b$ (i.e. $\lfloor k \rfloor = k$) or *x* is floating-point real, the partition of *x* is refined with the partition {(–∞, *k*), [*k*, *k*], (*k*, +∞)}; otherwise, *x*'s partition is refined with $\{(-\infty, \lfloor k \rfloor], [\lceil k \rceil, +\infty)\}$. During the partitioning process we distinguish integer objects from float and double objects. For integer objects, non-integer values are illegal, therefore – an interval that does not contain integers is irrelevant. Also, non-integer values can be excluded from ranges computed for integer objects. When we compute ranges for an object of type integer, if an interval (range) end is neither at $+\infty$ nor at $-\infty$, we round it to the nearest integer inside the interval, and close the end. For example, if we compute a range $(-10, 3.5]$ for an integer object, then this range contains exactly the same values as [–9, 3] in the integer domain, hence $(-10, 3.5]$ is replaced by [–9, 3].

The partitioning process continues based on dependency between valued objects. We consider a valued object *y* to be *dependent* on another valued object *x* if *x*'s current or previous value occurs in an assignment to, or an emission of, another valued object *y* (the data expression is of the form *a* * *D* + *b* where *a* and *b* are literals and *D* denotes *x* for a variable *x*, ?*x* for a sensor *x*, or either ?*x* or pre(?*x*) when *x* is a signal). Suppose *y*'s partition is $\{Y_1, Y_2, ..., Y_n\}$, then *x*'s partition is refined with $\{{}^1\!/_a \cdot Y_1 - {}^b\!/_a, {}^1\!/_a \cdot Y_2 - {}^b\!/_a, ..., {}^1\!/_a \cdot Y_n - {}^b\!/_a\}$. The order by which these dependencies are considered is detailed in Section IV.

There are two valued objects in the example: the sensor Speed and the output signal MotorForce. Since MotorForce is set with Speed's transformed value, MotorForce's ranges are computed before Speed's ranges. MotorForce is computed the following nine ranges: {(–∞, –270.0), [–270.0, –270.0], (–270.0, –30.0), [–30.0, –30.0], (–30.0, 30.0), [30.0, 30.0], (30.0, 270.0), [270.0, 270.0], (270.0, +∞)}. We denote them by $R_1^{\text{MotorForce}}$, ..., $R_9^{\text{MotorForce}}$, respectively. The only valued signal emission found in the entire program is: emit MotorForce (3.0 * (100.0 – ?Speed)). Let us denote this statement by $E_1$. Once executed, ?MotorForce = –3.0 * ?Speed + 300.0. Since Speed = 100.0 – MotorForce / 3, the partitioning provides the following ranges for Speed: {(–∞, 10.0), [10.0, 10.0], (10.0, 90.0), [90.0, 90.0], (90.0, 110.0), [110.0, 110.0], (110.0, 190.0), [190.0, 190.0], (190.0, +∞)}, denoted by $R_1^{\text{Speed}}$, ..., $R_9^{\text{Speed}}$. Note that when this emission is executed, MotorForce is in $R_1^{\text{MotorForce}}$ if and only if Speed is in $R_9^{\text{Speed}}$, MotorForce is in $R_2^{\text{MotorForce}}$ if and only if Speed is in $R_8^{\text{Speed}}$, and so forth.

Each range is matched with a pure *range signal*. For a valued object *x*, for which PARTITION(*x*) = $\{R_1^x, R_2^x, ..., R_n^x\}$, the ranges' corresponding signals are named R1_*x*, R2_*x*, etc. A configuration of the abstract, purified program in which R*i_x* is present corresponds to a configuration of the original, concrete program where the value of *x* is in $R_i^x$. In other words, an event in which R*i_x* is present in the abstract program stands for an event in which the value of *x* resides within range $R_i^x$ in the concrete program. In addition, if *x* in the source program is a signal, an occurrence of any of its range signals in the abstract program means that *x* is present in the corresponding configuration of the concrete program.

When we declare range signals for a numerical input signal or a numerical sensor (replacing the original signal or sensor definitions), we also declare an exclusion relation[5] among the range signals, such that the environment would not be able to provide more than one range for each original valued object at the same instant. As for sensors, since every sensor is ever-present and always carries a value, a thread is composed in parallel to the program, emitting the range signal for its first range when its other range signals are absent.

The declarations of sensor Speed and valued output signal MotorForce in the example are replaced with the following declarations:

```
input    R1_Speed, R2_Speed, R3_Speed, R4_Speed,
         R5_Speed, R6_Speed, R7_Speed, R8_Speed,
         R9_Speed;
relation R1_Speed # R2_Speed # R3_Speed # R4_Speed #
         R5_Speed # R6_Speed # R7_Speed # R8_Speed #
         R9_Speed;
output   R1_MotorForce, R2_MotorForce, R3_MotorForce,
```

[5] An *exclusion relation*, also known as *incompatibility*, is a declaration that asserts that no two signals listed by the relation declaration can be present simultaneously in the environment [4].

```
         R4_MotorForce, R5_MotorForce, R6_MotorForce,
         R7_MotorForce, R8_MotorForce, R9_MotorForce;
```

To implement Speed's "ever-presence" property, we compose the following code segment in a parallel thread:

```
loop
   present not ( R2_Speed or R3_Speed or R4_Speed or
                 R5_Speed or R6_Speed or R7_Speed or
                 R8_Speed or R9_Speed ) then
      emit R1_Speed
   end present;
   await tick
end loop
```

The original variable assignment and valued signal emission statements are replaced by emit statements, emitting pure local signals to denote the execution of their respective statements. The effect of the assignments and the emit statements from the program is simulated by parallel components, calculating the range signal that should be emitted every instant for each numerical valued object of the program, to represent its state in every instant. For a variable, the simulation code tests whether a signal denoting an assignment to that variable has occurred in the previous instant, and if so – it emits the range signal representing that variable's state following the assignment. If no "assignment" is performed, then the previous range signal is emitted once again. For full details regarding variable simulation using pure signals, see Section III. For a valued signal, the simulation code works similarly, with two differences: the previous range signal emitted need not be emitted in an instant if no "emission" takes place in that instant, and every emission operation can be handled independently.

We need to translate the emission of MotorForce in the example to range signal terms. The original emit statement is replaced with emit E1. A new code segment is composed in a thread running in parallel to the current module's body, emitting the correct range signal for MotorForce in response to E1's presence. The local signal E1 must be declared as well, and its scope must include both the transformed module's original threads and the new thread shown below.

```
... || [ loop
            present E1 then
               present R1_Speed then
                  emit R9_MotorForce
               end present;
               present R2_Speed then
                  emit R8_MotorForce
               end present;
               ...
               present R9_Speed then
                  emit R1_MotorForce
               end present;
               await tick
         end loop ] || ...
```

Variable abstraction also employs pure signals and resembles valued signal abstraction; however, it is different because a variable can store values for future instants and can change value multiple times in every instant. It is elaborated on in Section III.

We now discuss transforming expressions. ESTEREL features three types of expressions [4]: *data expressions*, which are built by combining basic objects using operators and function calls; *signal expressions*, which are Boolean expressions over signal statuses; and *delay expressions*[6], which are used by temporal statements such as await[7] and abort[8].

Signal expressions consist of: signal identifiers, which may appear within pre operators (meaning their status from the previous instant); parentheses; and logical operators (and, or and not). They are tested by present statements (conditional statements testing signal statuses instead of data) or occur within delay expressions used by temporal statements. These expressions are transformed as follows:

- An identifier of a signal *S* having a numerical data type, whose ranges are $R_1^S, R_2^S, ..., R_n^S$, is replaced by the subexpression (R1_*S* or R2_*S* or ... or R*n*_*S*).
- An identifier of a signal whose data type is boolean is replaced by the identifier of its respective presence signal.

In our example, there is a present statement testing whether MotorForce is present before testing its value. Accordingly, we replace the MotorForce identifier with a chain of or operators over the range signals representing MotorForce in the purified program: present (R1_MotorForce or R2_MotorForce or ... or R9_MotorForce) then...

Numerical data expressions whose values are assigned to variables or carried by signals are handled by the variable assignment and valued signal emission simulation mechanisms. Hence we would like to focus on Boolean data expressions, whose values are not only used in variable assignments and signal emissions, but also in if statements.

Boolean data expressions are recursively translated to signal expressions. They are tested at each instant in a loop composed in parallel to the program. If the expression is true, a special signal is broadcast and is used in the purified code:

1) To calculate the results of simulated assignments to boolean-typed variables and emission of boolean-typed signals; and
2) To simulate the evaluation of Boolean data expressions by if statements. An if statement testing a Boolean data expression $B_i$ is replaced with a present statement that tests its corresponding signal B*i* in the transformed program.

In the example there are two Boolean data expressions. We translate the first, ?MotorForce > 270.0 or ?MotorForce < -270.0, into the signals expression R9_MotorForce or R1_MotorForce and create a thread that runs in parallel to the module's original threads. The new thread contains a loop in which the signal B1 is emitted in every instant in which this signal expression is satisfied. Afterwards, if ?MotorForce > 270.0 or

[6] A delay expression is an expression defining a delay that begins when the temporal statement bearing it starts and elapses in some later instant, possibly in the same instant in which the delay starts (delays that may elapse immediately are called *immediate delays*, and they start with the immediate keyword). There are three types of delays (standard, immediate, and count delays), but all delay expressions use signal expressions.

[7] The await statement pauses program execution until a delay elapses [4].

[8] The abort statement kills its body when a delay elapses [4].

`?MotorForce < -270.0 then...` can be replaced by the statement `present B1 then...`.

Additional transformations include removing the original interface and local declarations for the valued objects and modifying input relations. There are two types of input relations offered by ESTEREL: *exclusion relations* and *implication relations*. Exclusion relations, shown earlier as means to ensure that no two range signals representing a numerical input signal, assert that no two signals listed by the relation declaration can be present in the environment simultaneously. If the original program contains an exclusion relation where a numerical valued signal appears, than the declaration is expanded to include all range signals representing it. Identifiers of valued Boolean signals are replaced by the identifiers of their respective presence signals. Implication relations, which are relations of the form `relation` *A* `=>` *B*, asserting that if *A* is present then *B* must be present, are harder to transform, as they require a different solution for each case. For example:

- If *A* is a valued signal and its type is numerical while *B* is pure, then the relation is broken-down to a series of declarations for each range signal representing *A*, stating that an occurrence of any range signal of *A* implies that *B* is present.
- If *A* is pure and *B* is valued and numerical, then we add a thread containing a loop, where in every instant we check if *A* is present, and if so we ensure that if no other range signal representing *B* is present, then `R1_`*B* is internally-emitted. The reason that we cannot use implication in this case is that we want to allow any range signal representing *B* to appear if *A* is present, and not necessarily one particular range signal.

The full list appears in [1].

We want to verify the following properties in the example:

1) The program never simultaneously emits the signals `AC_ON` and `AC_OFF`. Such a situation “confuses” the external controller of the air conditioner. When an occurrence of both signals in the same instant happens, the observer emits `ErrorACConflict`.
2) The air conditioning is not off when the force exerted by the motor is greater than 270.0 in either direction. In case of violation, the observer emits the signal `ErrorNoAC`.

Hereby is the observer code – expressed in terms of the abstract program:

```
signal AC_isOn, AC_isOff in
   loop
      % no air condition conflict: ON and OFF not emitted
      % at the same instant
      present AC_ON and AC_OFF then
           emit ErrorACConflict
      end present;
      % air condition is never off when motor force is
      % greater than 270 or lower than -270
      present AC_isOff and
              (R1_MotorForce or R9_MotorForce)
        then emit ErrorNoAC
      end present;
      await tick
   end loop ||
   [ abort
         sustain AC_isOff
      when immediate AC_ON;
      loop
         present AC_OFF then
            emit AC_isOff
         else
            present AC_ON then
               emit AC_isOn
            else
               present pre(AC_isOff) then emit AC_isOff
               end present;
               present pre(AC_isOn) then emit AC_isOn
               end present
            end present
         end present;
         await tick
      end loop ]
end signal
```

The declaration of the output signals `ErrorACConflict` and `ErrorNoAC` joins the interface of the module. This observer code consists of one parallel thread in charge of emitting the observer signals in every instant in case of a property violation, and another parallel thread calculating auxiliary signals. The auxiliary signals indicate whether the air conditioning is on or off, depending on the most recent control signal to the air conditioning subsystem emitted by the program. Composing the observer code in parallel with the main program's body, and running XEVE on the resulting program, neither `ErrorACConflict` nor `ErrorNoAC` are ever emitted.

When applying our technique to abstract programs that satisfy the technique's requirements, verification using observers is both sound and complete [1]. We proved it by showing that an abstract program has the same reactive behavior as the concrete program, up to communicating with the environment by means of pure signals instead of the original valued signals and sensors. There is a loss of information and the abstract program allows theoretically more behaviors than the concrete program, in the sense that the abstract program provides ranges of possible values for the outputs given ranges of possible values for the inputs. However, the two programs make the same control decisions under the same circumstances and emit corresponding outputs, in particular with respect to pure signals, including, but not limited to, the observer signals, hence the soundness and completeness of the proof technique as a whole. The proof starts by defining for every event of the concrete program a corresponding event of the abstract system. The proof that the two systems behave in the same way given corresponding timed input sequences is statement-wise. On the one hand, statements that do not handle valued objects remain unchanged by the abstraction. On the other hand, statements that handle them are altered or replaced, hence the proof focuses on them. We identify four groups of statements that concern valued objects: statements that declare local signals and variables; statements that test data, i.e. make control decisions based on the data (namely `if` statements); statements that test signals, i.e. make control decisions based on statuses (presence or absence) of signals; and statements that manipulate data. We investigate, for every statement, the behavior of the construct replacing it in the abstract program. The full details appear in [1].

## III. Variable Abstraction

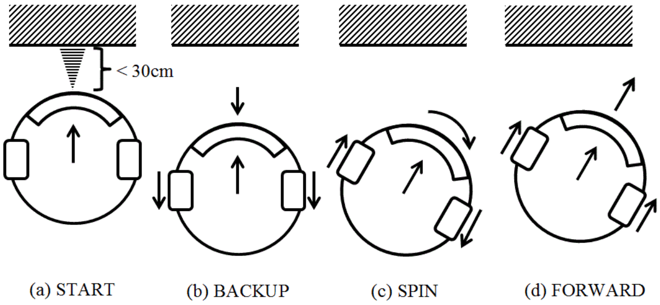
As we do for valued signals, we replace variables with local pure signals representing their current values. Employing pure signals for representing variables in the abstract program has a major drawback. ESTEREL supports multiple assignments to a variable in a single instant, as it does not contradict the synchrony hypothesis [4]; however, unlike variables, a pure signal may have only one status at each instant – it may either be present or absent, not both.

In the current version of the algorithm, we impose two restrictions on programs supported by the abstraction in addition to those listed in Section IV: (i) each variable is assigned at most once on every instant – after it is read[9] by all statements referencing it in that instant; and (ii) initializing a variable takes an instant (as a result of the previous constraint). Given these assumptions, we have implemented variable simulation[10] in the abstract program as follows:

- A local pure signal is declared for each assignment statement in the concrete program: `A1`, `A2`, `A3`, etc.
- Every assignment statement is replaced by an `emit` statement emitting the corresponding pure signal.
- For every variable $v$ a concurrent thread is added to the program in which there is a loop that emits in every instant the pure signal representing the current value of $v$.
- The code in this loop checks the previous status of signals indicating an assignment to $v$. If one of these signals has been emitted during the previous instant, it calculates the pure-signal representation of $v$'s value in the current instant based on the statuses of the pure signals representing the values of valued objects from the previous instant. If no signal indicating an assignment to $v$ has been present in the previous instant, the pure signals representing $v$'s previous value are re-emitted in the current instant; since no assignment to $v$ means that it retains its previous value.

This implementation with its implied restrictions is suitable for a certain group of programs, such as programs using a variable to manage a state machine: at the beginning of each reaction the current state is checked by reading the state variable, the reaction depends on the state and the input, and at the end of the reaction, if the program changes state or mode of operation, it assigns a new value to the state variable.

For example, in [1] we show a translation to ESTEREL of an *Escape* behavior for autonomous mobile robots based on a program taken from [11]. In this example, once a robot encounters an obstacle it backs off a predefined distance, rotates a predefined angle, and then returns to its previous occupation (e.g. cruising). We add a range detector such that the robot can escape an obstacle based on proximity without actually having to collide with it. The program uses a state variable, and in every reaction responds according to the current sensor reading and the current state. A reaction ends with assigning a new value to the state variable when transitioning to a new state. Figure 1 illustrates the implementation.



(a) START (b) BACKUP (c) SPIN (d) FORWARD

Figure 1: Robot Escape Behavior (with range detection).

To represent the behavior's state, we use an `integer` variable called `State`. The states are enumerated from 1 to 4. The partition of the integers domain to ranges for `State` is $\{(-\infty, 0], [1, 1], [2, 2], [3, 3], [4, 4], [5, +\infty)\}$ since `State` is assigned the values of 1, 2, 3 and 4, and is also tested to be equal to either one of them.

In the purification process we replace every statement assigning a value to `State` in the concrete program (see the source code in [1]) with a signal emission; e.g., the first assignment `State := 1` is replaced with `emit A1`; the second assignment `State := 2` is replaced with `emit A2`; and so forth. We compose in parallel to the original program a thread containing a loop that emits at every instant the range signal representing `State`'s current value:

```
|| loop
      % for each signal denoting an assignment to State
      % the next range signal for State is emitted
      present pre(A1) then
         emit R2_State
      else present (A2) then
         emit R3_State
      else ...
      else
         % if no assignment has occurred in the previous
         % instant then State's previous range signal
         % is emitted
         present pre (R1_State) then
            emit R1_State
         end present;
         present pre(R2_State) then
            emit R2_State
         end present;
         ...
```

The choice whether to postpone the effect of an assignment to the next reaction or to respond to it immediately affects only the constraint we have to impose, because either way our abstraction allows at most one assignment to a variable in every reaction. In the first option we allow one assignment at the end of the reaction after all statements use the value assigned to it in a previous instant, whereas the second option

[9] Reading a variable refers to an evaluation of a data expression containing an occurrence of that variable.

[10] By variable simulation we refer to the code that the abstraction adds to the program that simulates operations carried-out on a variable, i.e. assigning new values to it, in terms of the pure signals employed by the abstract program to represent the value of that variable.

allows one assignment at the beginning of the reaction and using that value henceforth, until the start of the next reaction where the variable is assigned again with a new value.

The difference between variable assignment in the concrete program, which is done immediately, and the variable assignment simulation in the abstract program, where the assignment is postponed to the next instant, has no externally-visible implications for programs that satisfy the two constraint mentioned above, since we require the concrete program not to use the new value of an assigned variable until the next instant. However, this approach severely confines the use of variables. One approach that we considered for allowing free use of variables was adding a delay immediately following an assignment, in order to "buy time" for the variable to update, i.e. giving the simulation code time to detect the signal notifying on the change and respond by emitting the new pure signals representing the variable's new value. However, this transformation, which in practice breaks-down seemingly one reaction into a series of reactions, can change the timing and semantics of the program significantly. In particular, it violates the assumption that every signal retains the same status throughout the entire reaction. For example, have a look at the following program segment:

```
emit O;
present O then
   v := 1
end present;
present O then
   v : = 2
end present;
```

Since `O` is present during the instant in which this code segment is executed, the two signal tests are satisfied and `v` is assigned consecutively the values of 1 and 2. Therefore, in the end of this program segment, `v` has the value of 2. Suppose we add a delay after assignment to `v` to comply with the assumptions above, e.g. by waiting for an arrival of some signal `DELAY` after each assignment statement. We get the following code:

```
emit O;
present O then
   v := 1;
   await DELAY
end present;
present O then
   v : = 2;
   await DELAY
end present;
```

In this program segment `v` is assigned the value of 1 since `O` is present before the first arrival of the `DELAY` signal. If we assume that `O` is not emitted by any parallel thread, then `v` is not assigned the value of 2 since `O` is absent after the arrival of `DELAY`.

We suggest an improvement to our algorithm, a workaround that enables us to assume that every variable is assigned at most once in every instant. This requires implementing a pre-processing step, combined with an alternative implementation of variable simulation that responds immediately to signals notifying on assignments to the original variable.

The pre-processing step will transform the concrete program such that each variable is assigned at most once in every reaction. We call the target form SINGLE ASSIGNMENT PER REACTION (SAPR). It is inspired by *Static Single Assignment* (SSA) form. An SSA form is a representation of a program involving separating each variable *v* in the program into several variables $v_i$ such that each variable is assigned only once [12]. In this work we refer to the variables $v_i$ as *instances* of *v*. The SAPR form is distinguished from SSA form in the sense that every variable is assigned at most once in every instant but not necessarily once in the entire program code. An additional difference is that SSA form employs Φ-functions: in a node in the control flow graphs having incoming edges from two other nodes where there are two different valid instances of some variable (for example, such node could be a statement following an `if` statement having two branches – `then` and `else`, each of which manipulates a certain variable `L` from the original program code, such that new instances are created in each branch for `L` – `L1` and `L2` respectively), the SSA form defines a new variable (e.g. `L3`) and assigns to it a Φ-functions standing for the value of valid variable instance at the entrance to that node (i.e. `L3` ← Φ(`L1`, `L2`)). In the SAPR transformation we need an equivalent of a Φ-function where the control converges back at terminations of branching statements and traps. We use special variable instances which are assigned the value of the valid variable instances at the ends of branches and before `exit` statements escaping traps.

For example, suppose that in the program investigated in Section II we wanted to limit the value of control signal to a maximum of 50 in either direction, and wanted to break-down the calculation of the control signal to several steps using variables, we could have used the following code:

```
var Error, CtrlSigVal : double in
   Error := 100.00 - ?Speed;
   CtrlSigVal := Error * 3.0;
   if CtrlSigVal > 50.0 then
      CtrlSigVal := 50.0
   end if;
   if CtrlSigVal < -50.0 then
      CtrlSigVal := -50.0
   end if;
   emit MotorForce(CtrlSigVal)
end var
```

This code segment employs two variables: `Error`, which contains the error, i.e. the difference between the set point (the target speed – 100) and the current speed; and `CtrlSigVal`, which contains the value of the control signal, obtained by multiplying the error (the value of `Error`) by the gain (3). If the product of the error and the gain is smaller than –50, then it is set to –50, and if it is larger than 50 it is set to 50. Below is the SAPR form, guaranteeing that in each instant each variable is assigned at most once:

```
var Error_0, Error_1,
    CtrlSigVal_0, CtrlSigVal_1,
    CtrlSigVal_2, CtrlSigVal_3,
    CtrlSigVal_4, CtrlSigVal_5 : double in
   Error_1 := 100.00 - ?Speed;
   CtrlSigVal_1 := Error_1 * 3.0;
   if CtrlSigVal_1 > 50.0 then
      [ CtrlSigVal_2 := 50.0 ] ;
      CtrlSigVal_3 := CtrlSigVal_2
   else
      CtrlSigVal_3 := CtrlSigVal_1
   end if;
   if CtrlSigVal_3 < -50.0 then
      [ CtrlSigVal_4 := -50.0 ] ;
      CtrlSigVal_5 := CtrlSigVal_4
    else
      CtrlSigVal_5 := CtrlSigVal_3
    end if;
    emit MotorCtrlSig(CtrlSigVal_5)
end var
```

In the transformed code segment we create a variable instance for every assignment statement. Once assigned, a variable instance substitutes the variable from the original program wherever it appears in data expressions from that point on, until the next assignment to the original variable, where another variable instance is used. `if` statements, which have no `else`-branch in the original program, are added with an `else`-branch. A new variable instance is created to hold the valid value upon leaving the `if` statement, and it is assigned at the end of each branch. This exemplifies the reason why `else`- and `then`-branches are introduced when missing – if a variable is manipulated in one branch and not in another then still at the end of the `if` statement there is one variable instance replacing the original variable and holding the right value. In [1] we elaborate on the transformation and provide the full example, including simulation results.

## IV. Discussion

Our method supports only programs that compile and execute without errors and that comply with several assumptions and constraints, which are described in this section. These constraints are required for maintaining relationships between valued objects and for replacing numerical and Boolean calculations with pure signal operations. However, our method still is useful for a large set of control and robotic programs. Various robot behaviors appearing in [11] can be implemented in Esterel, processing pure signals or requiring sufficiently simple calculations, thus our method can be applied to them.

In [1] we list assumptions about the input program that exist mainly for the simplicity of the solution. These assumptions usually concern Esterel syntactic sugaring instructions. However, they do not reduce the expressive power of Esterel, as each assumption can be attained by replacing the original construct by a semantically-equivalent construct, either manually or automatically. For example, we require that the program will consist of only one module. When the program consists of several modules, one module calls another using the `run` statement. The `run` statement instantiates one module within another module by syntactically replacing the `run` statement with the body of the instantiated sub-module, exporting the data declarations of the instantiated sub-module to the parent module, and discarding the interface declaration of the instantiated sub-module [4]. By replacing any `run` statement in a program having multiple modules with the instantiated sub-module's code, we can comply with this assumption.

Unfortunately, we have some real constraints that limit the family of programs to which our abstraction is applicable.

*a) We do not support external code, i.e., code written in the host language:* The Esterel v5 compiler translates the program into a program or circuit written in a host language chosen by the user, for example C. The programmer can declare abstractly and use various function, procedures, tasks, constants and data types to be implemented in the host language and linked to the code generated by the Esterel compiler [4]. Being outside the scope of the Esterel program, its behavior cannot be taken into consideration, as our method verifies the Esterel code.

*b) We do not support cyclic dependencies between values of numerical valued objects:* In order to maintain relationships between targets and sources of assignments (as well as valued signal emissions) of numerical valued objects, we restrict the numerical data expressions used to affine transformations of values of numerical valued objects (see constraint 3 in the list below in this section). We refine a valued object $x$'s partition of (−∞, +∞) to ranges for every assignment of $x$'s value to a valued object $y$, according to the data expression, defining an equation with $x$ and $y$ which is satisfied once that assignment is executed.

Since the order of partitioning is derived from the dependencies between valued objects, if there exist cyclic dependencies[11] between valued objects then the order of partitioning cannot be determined, since no strict ordering of the partitions calculated can reflect the dependencies between the objects.

Even when an object's value is transformed and set to itself directly (e.g. `v:=3*v+1` for a variable $v$), the partitioning process should theoretically refine the partition of that object's domain to ranges over and over again, infinitely many times. One solution we considered was to allow self-assignments, in the following form: Let $P = \{R_1, R_2, ..., R_n\}$ be a partition of $(-\infty, +\infty)$ for some valued object $x$. Suppose we do not take self-assignments into consideration when partitioning, but rather leave the current partition, and define assignment as "transitioning" $x$ from its previous value's bounding range(s) to a union of ranges bounding $x$'s value following the assignment. If $x$ occurs in a formula $F$ testing $x$'s value, and an assignment causes the bound of $x$'s value to contain both ranges satisfying $F$ and ranges not satisfying $F$, we cannot decide if $F$ is satisfied. For example, suppose we compute for an integer variable $v$ the ranges [1, 8], [9, 9] and [10, 12] (among others) and we have a conditional testing the expression $v$ = 9. Consider the assignment $v$ := $v$ + 1. If $v$ is in [1, 8], and the assignment is executed, then in the next instant it can be in either [1, 8] or [9, 9], thus we cannot decide if the condition $v$ = 9 is satisfied.

To determine the order by which the partitioning process inspects valued objects when partitioning by dependency, a directed graph of valued object dependencies $G = \langle V, E\rangle$ is created in the following manner. Our inputs include the set VALOBJ of valued objects in the source program, and the set of all variable assignments and valued signal emissions

[11] The term *cyclic dependency* in Esterel usually refers to a situation where there exists an instantaneous circular dependency between a signal and itself [4] (also known as a *causality cycle*). Esterel programs that contain an instantaneous dependency cycle, for which the number of solutions is not exactly one, are considered invalid. In this section we discuss a different kind of dependency, however: a dependency between the current value of an object and its previous one, from which it is calculated. We use the terms *causality cycle* and *causality problem* to refer explicitly to a causality cycle.

which appear in the original program. $V$ and $E$ are defined as follows. For each numerical valued object define a vertex: $V = \{x \mid x \in$ VALOBJ and $x$ is of type `integer`, `float`, or `double`$\}$. The set $E$ of edges in $G$ is calculated using the following procedure.

1. $E \leftarrow \emptyset$
2. For each two numerical valued objects $v$ and $u$ in $V$:
   - 2.1. If the current or previous value of a numerical valued object $u$ is assigned to $v$ (i.e., if $v$ is a variable and there exists an assignment $v$ := *exp* or if $v$ is a signal and there exists an emission `emit` $v$(*exp*), where *exp* is a data expression with an occurrence of $u$), stretch a directed edge from $u$ to $v$; that is, add $(u, v)$ to $E$.
   - 2.2. If $G$ contains cycles, halt the process with an error message.

The order by which the process partitions the domains of valued objects is a post-order of this graph. That is, a valued object $v$'s partition is computed after the partition is computed for all numerical objects to which $v$ is assigned.

Hereby is a summary of the constraints on the programs to which our method is applicable. These compromises are basically due to two limitations:

1) One way to allow automatic methods to check a non-trivial property is to reduce the power of the language or, equivalently, reduce the class of program verifiable using the method; and
2) Technical issues with aspects of our abstraction process conflicting with the synchrony hypothesis.

The constraints are:

1) All valued objects are of type `boolean`, `integer`, or floating-point real (i.e. strings and user-defined types are not allowed);
2) All numerical formulas used in Boolean data expressions are of the form $(a{\star}D{+}b)\ R\ k$ where $R$ is an relational operator, $D$ is the current or previous value of a numerical valued object (sensor, valued signal or variable of type `integer`, `float` or `double`) and $a$, $b$ and $k$ are some literal constants;[12]
3) All numerical data expressions used in assignments and valued signal emissions are of the form $a{\star}x{+}b$, where $x$ is an occurrence of a valued object (in its current or previous value), and $a$ and $b$ are constant literals, all of which of the same domain;
4) All delayed expressions in temporal statements may not be count delays;
5) No repeat loops;[13]
6) No combined signals and no valued traps;
7) No cyclic dependencies between numerical valued objects or self-assignments (except in loop-free program segments); and
8) During one instant, the program does not access more than one incarnation of a local valued signal or a variable.

Currently the algorithm and tool support accessing the values of valued signals only during reactions in which they are present; however, this can be handled by adding signals to represent the latest value of a signal from the last time it occurs. This solution will be implemented in future versions.

The set of programs to which our method is not applicable includes, but not limited to, programs employing counters (due to a dependency of the counter by its previous value); programs performing calculations which are more complicated than allowed by constraints 2 and 3 (e.g., containing operations such as dividing by the value of some object or comparing one numerical valued object to another); and programs requiring extension in the host language. By design, Esterel's data definition facilities are minimal, since data-handling is not the primary concern in control-dominated reactive programming [4]. A major advantage of this approach is high portability of the code. To program complex systems, the programmer has to put a significant effort in implementing data handling capabilities in the host language. Without supporting host language extensions, our technique is limited to simpler, control-centric programs.

A *Proportional-Integral-Derivative (PID) controller* [11] is an example of a controller to which our method is not applicable. This controller has three terms: proportional (proportional to the error signal), derivative (proportional to the derivative of the control signal, i.e. the rate of change in the error signal over time), and integral (proportional to the integral of the error signal, i.e. the summation of error over time). To calculate the integral term, the system needs to sum of the error over time, essentially creating a cyclic dependency. However, in the high level at which robot behaviors are programmed in behavior-based robotics, proportional controllers, which our method supports, are nearly always sufficient [11]. Moreover, PID controllers are not the usual application of Esterel. Esterel, as a *state-based formalism*[14], is better suited for problems where control flow is prevalent, e.g. systems that jump between different functioning modes [13]. Another style of synchronous programming, where the system's behavior is represented as a set of recurrent equations, characterizing languages such as Signal and Lustre, is well-adapted to problems where data-flow is prevalent [13], hence being more natural for implementing PID controllers where the transfer function is repeatedly calculated.

We currently investigate means to lift some of these constraints. Some research directions appear in Section VI.

[12] Currently we support only linear equations and inequalities in one variable, for future work see Section VI.

[13] Esterel v5 offers several loop constructs: `loop`, `loop-each` and `every` (temporal loops) and `repeat`. A `repeat` loop executes a finite number of times, unlike the others, which can loop forever [4]. We do not support only the `repeat` statement.

[14] A state-based formalism uses a state transition diagram where arrows are labeled with communication actions to represent the system's behavior. The diagram can be explicit in visual formalisms such as in Statecharts or implicit in imperative formalisms such as Esterel and CSML [13].

### *State-Space Complexity Evaluation*

The complexity of the number of pure signals that the abstraction adds to the program is $O\left(N_{Bool} + N_{num} \cdot (B_{atomic} + A_{const}) + A + E + B\right)$ when there is at most one relation between every two valued objects; $N_{Bool}$ is the number of Boolean valued objects, $N_{num}$ is the number of numerical valued objects, $B_{atomic}$ is the number of atomic Boolean data expressions having occurrences of numerical valued objects, $A_{const}$ is the number of assignments of constant values to numerical valued objects, and *A*, *E*, and *B* are the numbers of variable assignments, valued signal emissions, and Boolean data expressions respectively. For more complicated situations in which there can be two or more relations between two numerical valued objects we provide a recurrence relation in [1].

In general, an automaton generated from an ESTEREL program may suffer from size explosion [14]. Increasing the number of input and output signals may significantly enlarge the sets of states and transitions of the automaton produced from the purified program compared to the one generated from the original program. In two of the examples presented in [1, Chapter 5] the number of states remains the same following the purification and for one example the number of states increases from 6 to 16. However, one must remember that the abstraction is done for verification purposes only. Moreover, a typical ESTEREL application yields a fairly small number of states, usually between 10 and 100 [14].

## V. Case Studies

The constraints described in the previous section provide an accurate characterization of the family of programs to which the method can be applied. The example in Section II demonstrates the applicability of the method to two classes of control systems within this set: bang-bang controllers and proportional controllers. Another example demonstrating the application of the method to a bang-bang controller is a temperature controller system maintaining a constant temperature in a chamber using a thermometer, a timer, and two boilers [1]. We present there also the application of our method various robot programs:

- Border-following robots using one or two light sensors.
- Bang-bang implementation of the *Home* robot behavior: a robot homes on a destination marked by a beacon using differential sensing. It is based on a program appearing in [11]. We use our technique to verify that the robot does not run over the beacon, assuming that when the robot is too close then the light intensity picked by the sensor is greater than a given threshold.
- *Escape* behavior: once colliding with an obstacle, a robot goes back a predefined distance, rotates a predefined angle and then continues moving forward. It is also based on a program that is described in [11] (though modified to use a range detector and escape once detecting a close obstacle). We verify that the robot never reaches a non-reactive state, in which it fails to find a reaction and emits an error signal.

The last two examples (Home and Escape behaviors) demonstrate the application of the method to mobile robot programming following the *reactive control paradigm*. This paradigm, based on animal models of intelligence, decomposes the overall action of the robot by behavior, allowing handling multiple goals and multiple sensors, increasing robustness and extensibility [15]. By combining various behaviors and control algorithms, complicated control systems to which our method is applicable can be derived. The full code for the examples, including original program code, abstract program code, and observer code are provided in [1].

The abstraction we propose has an advantage over complete data abstraction performed when providing the `-soft` flag to the ESTEREL compiler, since it avoids adding behaviors not displayed by the original program. For example, consider the following code portion switching on and off an actuator based on a numerical input through a sensor `S`:

```
if ?S < 90.0 then emit On end if;
if ?S > 110.0 then emit Off end if
```

The ESTEREL compiler generates with the `-soft` flag a circuit in which both conditions can be satisfied at the same time, therefore it can emit both `On` and `Off` at the same reaction. However, using our abstraction, in the purified program no two range signals representing `S` can occur simultaneously, therefore both conditions cannot be satisfied at the same time.

## VI. Conclusion

Combining the verification power of XEVE with the transformation of non-Pure ESTEREL programs into Pure ESTEREL programs, we can verify safety properties of a larger family of programs. We provided examples for various categories of programs that can be verified.

There still is a large set of programs to which we cannot apply our method, e.g. those involving complicated calculations and counters. By giving up completeness, the full potential of the technique developed is realized. Applying the technique to parts of the program that fulfill the constraints while letting the ESTEREL compiler remove the rest of the data when compiling the program into a circuit can produce a more precise over-approximation than total control-based abstraction. This is especially useful when the system consists of several subsystems, some of which fulfilling the constraints while others not. In [1] we provide an example of a system comprised of a PID controller and a limit switch. The limit switch shuts the process down by cutting off the PID controller's output once an undesired limit is reached. After the measured value drops back to the safe zone, the switch can be manually reset in order to reactivate the control system. An observer helps to verify safety properties of the system by emitting a special signal whenever the program emits the output signal while the measured value is higher than some critical threshold. Not only that the calculations performed by the PID controller are not supported by our technique, but also the program takes the set point, clock interval, and gains from constants defined in the host language. The abstraction performed by the ESTEREL compiler alone produces a circuit which XEVE reports to possibly emit the observer signal. However, using our technique to abstract the safety limit switch and the observer, which comply with the requirements of our techniques, and letting the compiler abstract the rest of the program creates a circuit which never emits the observer signal, as XEVE guarantees.

The current version of the algorithm supports only linear equations and inequalities in formulas occurring in Boolean data expressions. We plan to extend the class of supported formulas, for example to handle also univariate polynomial equations and inequalities whose the roots can found analytically, i.e. polynomials of second, third, and forth degrees; the roots can be used to partition the domain of the valued object (which serves as the variable of the formula) to intervals.

One alternative approach to perform the analysis and the abstraction is based on the constructive semantics of ESTEREL [16]. ESTEREL statements are divided into two groups: *kernel statements*, forming a primitive core of the language, and *derived statements*, which are definable as combinations of kernel statements, whose purpose is to make programming more convenient. The semantics of an ESTEREL program is obtained by structural induction on the statements which it consists of. Considering the constructive semantics of ESTEREL, it is essentially enough to define the abstraction process and prove its correctness for kernel statements to obtain a process and a proof that applies to the entire ESTEREL syntax. Yet it would require to implement a much more complicated abstraction technique, that not only finds and transforms items of interest, but also breaks-down the program to kernel statements. Our approach focuses on the occurrences of valued objects in the program, leaving the rest of the program untouched.

The TEMPEST [17] toolset provides a compiler for linear temporal logic formulas representing safety properties to observers in ESTEREL language. The automaton compiled from the observer-augmented program can be verified using other tools from the TEMPEST package. When the control flow of the program to be verified is independent of valued objects (i.e., there are no conditionals testing run-time values of valued objects), the control structure can be fully determined at compile-time, therefore the verification is both sound and complete. However, when the program's control structure depends on run-time values, the technique is sound but not complete, since the ESTEREL compiler considers all paths, including those that are unreachable due to data values [18, Section 4].

We plan to incorporate the power of TEMPEST with our method to extend both the class of verifiable properties and the class of programs to which those techniques can be applied. TEMPEST should be able to successfully verify programs where the control structure is independent on the run-time values of valued objects; therefore, TEMPEST should be directly applied to programs not complying with the constraints which our method requires, while actually not querying values of objects in conditionals. Hybrid versions may be suggested, such as purifying only valued objects on which the control-flow depends directly or indirectly, while leaving the rest of the valued objects untouched (as long as the valued objects on which the control flow depends satisfy our constraints). This, however, cannot be verified using XEVE. An additional possible research direction is extending the language used in TEMPEST to express safety properties querying the run-time values of valued objects, such as "variable $V$ can never be greater than 20". In certain cases, this might be obtained by purifying together the parallel composition of the main program and the observer component.

## REFERENCES

[1] N. Koblenc, "Purification of Esterel Programs," Master's thesis, Dept. Mathematics and Computer Science, Open Univ. of Israel, Ra'anana, Israel, June 2015, the thesis and prototype tool are available at http://www.cs.tau.ac.il/~tyshbe/NIR/nirThesis.html.

[2] N. Koblenc and S. Tyszberowicz, "Purification of Esterel Programs," in *Preproceedings of the Brazilian Symp. on Formal Methods (SBMF)*, C. Braga and N. Martì-Oliet, Eds., 2014, pp. 183–188, available at: http://www2.ic.uff.br/~cbraga/sbmf14/sbmf14-preproceedings.pdf (visited September 2015).

[3] N. Halbwachs, *Synchronous Programming of Reactive Systems*, Stankovic, J. A. (Consulting Editor), Ed. Kluwer, 1993. [Online]. Available: http://dx.doi.org//10.1007/978-1-4757-2231-4

[4] G. Berry, "The Esterel v5 Language Primer, Version v5_91," Centre de Mathématiques Appliquées – Ecole des Mines and INRIA, 06565 Sophia-Antipolis, 2000.

[5] L. J. Jagadeesan, C. Puchol, and J. E. von Olnhausen, "A formal approach to reactive systems software: a telecommunications application in Esterel," *Formal Methods in System Design*, vol. 8, no. 2, pp. 123–151, 1996. [Online]. Available: http://dx.doi.org/10.1007/BF00122418

[6] I. Sommerville, *Software Engineering*, 8th ed. Addison-Wesley, 2007.

[7] A. Bouali, "XEVE, an Esterel verification environment," in *Computer Aided Verification (CAV)*, ser. LNCS, A. J. Hu and M. Y. Vardi, Eds., vol. 1427. Springer-Verlag, 1998, pp. 500–504. [Online]. Available: http://dx.doi.org/10.1007/BFb0028770

[8] "The Esterel v7 Reference Manual, Version v7_30 – initial IEEE standardization proposal," Esterel Technologies, 2005.

[9] G. Berry and the Esterel Team, *The Esterel v5_91 System Manual*, Centre de Mathématiques Appliquées – Ecole des Mines de Paris / INRIA, Sophia-Antipolis, 2000.

[10] S. Das, D. L. Dill, and S. Park, "Experience with predicate abstraction," in *Computer Aided Verification (CAV)*, ser. LNCS, N. Halbwachs and D. Peled, Eds., vol. 1633. Springer, 1999, pp. 160–171. [Online]. Available: http://dx.doi.org/10.1007/3-540-48683-6_16

[11] J. L. Jones and D. Roth, *Robot Programming: A Practical Guide to Behavior-Based Robotics*. McGraw-Hill, 2003.

[12] B. Alpern, M. N. Wegman, and F. K. Zadeck, "Detecting equality of variables in progress," in *Principles of Programming Languages (POPL 1988)*. ACM, 1988, pp. 1–11. [Online]. Available: http://dx.doi.org/10.1145/73560.73561

[13] A. Benveniste and G. Berry, "The synchronous approach to reactive and real-time systems," *Proceedings of the IEEE*, vol. 79, no. 9, pp. 1270–1282, September 1991. [Online]. Available: http://dx.doi.org/10.1109/5.97297

[14] G. Berry and G. Gonthier, "The Esterel synchronous programming language: design, semantics, implementation," *Science of Computer Programming*, vol. 19, no. 2, pp. 87–152, 1992. [Online]. Available: http://dx.doi.org/10.1016/0167-6423(92)90005-V

[15] G. Dudek and M. Jenkin, *Computational Principles of Mobile Robotics*. Cambridge University Press, 2000.

[16] G. Berry, *The Constructive Semantics of Pure Esterel*, 2002, draft book, version 3, available at: http://www-sop.inria.fr/members/Gerard.Berry/Papers/EsterelConstructiveBook.pdf (visited September 2015).

[17] C. Puchol, *The TempEst Program Verification Toolset*, AT&T Bell Laboratories and the University of Texas at Austin, product documentation.

[18] L. J. Jagadeesan, C. Puchol, and J. E. Von Olnhausen, "Safety property verification of Esterel programs and applications to telecommunications software," in *Computer Aided Verification (CAV)*, ser. LNCS, P. Wolper, Ed., vol. 939. Springer-Verlag, 1995, pp. 127–140. [Online]. Available: http://dx.doi.org/10.1007/3-540-60045-0_45

DOI: 10.15439/2015F45
ACSIS, Vol. 5

# Real-Time Cyber-Physical Systems Transatlantic Engineering Curricula Framework

Wojciech Grega
Department of Automatics and Biomedical Engineering
AGH University of Science and Technology
Kraków, Poland
wgr@agh.edu

Andrew J. Kornecki
Electrical, Computer, Software and System Engineering
Embry Riddle Aeronautical University
Daytona Beach, FL, USA
kornecka@erau.edu

[1]*Abstract*—**How to educate future engineers, so that they acquired new skills and competences to become developers of Cyber Physical Systems (CPS)? The paper demonstrates a curriculum framework that was developed and successfully implemented some years ago, as an outcome of two international projects undertaken by a consortium of the European and American universities. The projects had focused on special category of CPS - Real-Time CPS. Examples of laboratory environment supporting education in the field of Real Time CPS are given in the paper.**

CYBER Physical System is an example of advanced complex technological system interfacing with physical plant and with computing environment running software intensive and most often time critical application. CPS's are often safety-critical. In case of missed time deadlines or component failures CPS may result in violating safety constraints and event life-threatening consequences. We refer to such systems as Real-Time Cyber-Physical Systems (R-TCPS).

Car airbag system is a good example of such system. When a cyber-physical safety system in a car detects a crash, the airbag must inflate in tens of milliseconds to avoid injuries of the driver. The success of the airbag system relies upon the crash sensors working accurately and extremely quickly but also upon algorithms processing in real time. Such a time-critical and safety critical functionality requires specific engineering solutions.

Typically, the challenges of R-TCPS are viewed as separate problems of data measurement, transfer, distributed control, real time operating systems, etc. Domain experts contribute individual components of a design, but they often lack cross-disciplinary knowledge of how these components interact and what impact their behavior may have on other components. As the prevalence and sophistication of cyber-physical systems increases, so does the need for developers with sufficient cross-disciplinary education. In summary, designing of R-TCPS require a multi-disciplinary and integrating knowledge, covering a broad area, from control theory to real-time computing.

The problem addressed in this paper is education: how to educate future developers of R-TCPS that would often work in multinational companies interacting with fellow engineers from different countries?

The following ABET Engineering Criteria Program Educational Outcomes were formulated for accreditation process of CPS education [1]:

- Ability to apply mathematical models of physical systems, cyber systems, and their composition.
- Ability to design and conduct simulations and tests of a cyber-physical system and to analyze the results.
- Ability to apply good engineering practices in the design of a system that mixes cyber and physical components subject to constraints including safety, security, cost, and dependability.
- Ability to function effectively on multi-disciplinary teams spanning cyber and physical domains.
- Ability to identify, formulate and solve engineering problems that have both cyber and physical aspects.
- Understanding of the professional and ethical responsibilities of the design of life- and safety-critical systems.
- Ability to communicate effectively across cyber and physical domains.
- Understanding of how design decisions in the cyber domain may affect the physical domain and visa-versa.
- Recognition of the need for, and ability to engage in life-long learning.
- Knowledge of contemporary issues with cyber-physical systems.
- Ability to select and use appropriate techniques, skills, and modern engineering tools that span the cyber and physical domains.

Most of the ABET criteria were addresses by two-year long ILERT project: International Learning Environment for Real-Time Software Intensive Control Systems (RSIC) started in 2006 [2]. The project was sponsored by the Fund for Improvement of Postsecondary Education (FIPSE) of the U.S. Department of Education and the European Commission. The project was implemented in a consortium

[1] The work presented in this paper was supported by the AGH University of Science and Technology, grants number 11.11.120.768, 11.11.120.396 and by the U.S. Department of Education FIPSE grants P116J060005 and P116J080025.

of one American (ERAU - Embry Riddle Aeronautical University, Daytona Beach, FL) and three European universities (AGH - AGH University of Science and Technology, Krakow, Poland; UJF - Université Joseph Fourier, Grenoble, France; and BUT - Brno University of Technology, Brno, Czech Republic). The academic departments leading the project represented computing, control, and telecommunication disciplines.

Based on own research and the industry surveys, the consortium faculty confirmed that modern computer control systems are heavily software-centric, implementing reactive and time critical software, where safety is the issue and the margin for error is narrow. Examples include aircraft avionics, air traffic control, space shuttle control, medical and automotive equipment, power stations and more. It is vital for future software developers to understand the basic real-time applications concepts such as the issues of timing, concurrency, inter-process communication, resource sharing, interrupts, and handling external physical devices. Generally, challenge is the situation where the physical systems operating in Newtonian (absolute) time must interface with computational systems and networks evolving in cyber time.

The purpose of ILERT project was defined as follows [2]: … {four university partners} *propose a study leading to establishment of a RSIC international, multidisciplinary curriculum framework focusing on important aspect of the computer/system/control/software engineering education. We plan to explore the mechanism for involving students from multilingual geographically separated institutions in a coordinated educational experience exposing them to the problems, methods, solution techniques, infrastructure, technologies, regulatory issues, and tools in the domain of dependable real-time safety-critical software-intensive control systems*.

The consortium members were deeply convinced that engineering curricula require continuous modifications to prepare students for the technological challenges of the modern workplace. Rapid progress of computing technologies is the major reason that the academic programs like electronics, computer and software engineering, robotics, and control engineering need continuous updates and integration of knowledge.

An additional issue was the internationalization and globalization of complex systems development. Several large companies, specifically in the aerospace industry, engage international teams working in geographically diverse locations often using diverse standards, guidelines, and processes. It is advantageous for future engineers to understand the implications of international collaboration and to appreciate cultural differences.

Successful ILERT project was subsequently continued by Dependable Systems International Research and Education Experience (DeSIRE2) project which started in 2008 [3]. Two American universities joined the consortium but only one continued and contributed to the project (UCF – University of Central Florida, Orlando, FL). The objective of this implementation project was to establish a platform for a sustained and consistent mobility exchange of students engaged in R-TCPS oriented programs. The target of the programs was to deliver on the labor market graduates, capable of working efficiently in multidisciplinary teams and participating in international collaboration on industrial R-TCPS projects, which require conformance to specific standards mandated by regulatory authorities.

This paper summarizes the results of both projects, as an example of a Real-Time Cyber-Physical Systems education framework.

## I. First Project: Curriculum Design

The first project (ILERT) started with a research phase, including the analysis of industry requirements related to graduates in the proposed domain. The collected data were analyzed and the results were used to help identify academic program learning objectives and outcomes, thus preparing a base for creation of a new curriculum framework [3].

The subsequent steps included:

- Defining learning objectives and outcomes, identifying the curriculum framework, exploring the partners' programs common features, laboratory infrastructure, identification of the curricula contents, and analysis of the educational process assessment. One needs to note, that the existence of common characteristics did not imply automatic commonality of the ways how individual institutions pursued common educational objectives. Universities often find their own methods to let shared content shape the development of procedure and instructional delivery.
- Defining the roles of the project partners and focus of interest. Considering the four universities strengths in various areas such as control, software engineering, telecommunication, and embedded systems, the consortium defined as their point of interest Real-time Software-intensive Control (RSIC) - a subset of RT-TCPS domain.
- Curriculum development included two complementary groups of activities. The first one, more general, had focused on methodology and process of creation of a unique and curriculum framework in the area of real-time software-intensive control, which combines elements of control, software, and computer engineering programs. Such curriculum was in high demand by industry as shown in earlier industry surveys and interviews. The second was a practical case study adapting a selected RSIC curriculum unit acceptable for engineering programs in four organizations. The curricula were reviewed to prioritize and integrate the various elements in order to meet the learning objectives and demands of interdisciplinary specialization in Real-Time Software-Intensive Control serving RSIC. An appropriate sequence of courses was proposed.

- Credit transfer and accreditation issues. The development of new curriculum framework in information technologies may require new approaches to validate and accredit learning. Existing and emerging structures for accreditation, quality control and credit transfer (such as the European Credit Transfer and Accumulation Scheme) were analyzed and coordinated. The proposed curriculum units, focusing on the objectives and outcomes of the educational activity, were developed according to the U.S. Accreditation Board of Engineering and Technology (ABET) standards as well as the applicable standards of Ministry of Higher Education in the European countries.
- Students mobility. Based on the developed curricula, students' mobility between partners institutions was proposed, opening possibility of collaborating and enrolling in the same course offered concurrently in four partner sites.

The project identified six RSIC Framework areas [4]: Software Engineering, Digital Systems, Computer Control, Real-Time Systems, Networking, and Systems Engineering.

1. Software Engineering: software engineering concepts and practices, lifecycle models, project management, processes, software modeling and formal representation; software requirements; software architectural and module design; software construction methods and practices, testing and quality assurance; software maintenance, notations and tools.
2. Digital Systems: methods, techniques, and tools used to support the design of combinational and sequential digital circuits and the design of fault tolerant and advanced networked hardware components.
3. Computer Control: concepts of feedback control, continuous vs. discrete models of dynamic systems, simulation, controller design, implementation of control algorithms in real-time, integrated control design and implementation (hardware-in-the-loop), security aspects, implementation for physical processes.
4. Real-Time Systems: timing and dependability properties of software intensive systems, RTOS concepts and applications, concurrency, synchronization and communication, scheduling, reliability and safety aspects.
5. Networking: data communication, network topology, analysis and design, information security, algorithms, encryption, bus architectures, wireless, distributed control and monitoring.
6. System Engineering: system engineering concepts, principles, and practices; system engineering processes (technical and management), system requirements, design, integration, and testing; special emphasis on the development of a RSIC system and the integration of RSIC system elements.

The basic organizational unit for this framework was RSIC "component". A RSIC component was defined as a curriculum unit covering theory, knowledge, and practice - supporting the RSIC objectives and outcomes. The proposed RSIC Curriculum Framework did not specify the way in which component topics might be formed into modules or courses reflecting the RSIC concept. Component topics were focused in one or two courses, or spread among several courses, along with other non-RSIC topics. For each component more detailed didactic aspects were defined such as: the prerequisite knowledge, component learning objectives, information about required facilities and equipment, and guidelines and suggestions for course design and delivery.

In addition to the above six RSIC Framework areas, additional institutional, regional, or national requirements in areas of "general education" were required to be added to the curricula as, for example, requirements for oral and written communication, for arts and humanities, or for the social sciences. These areas also supported the RSIC curriculum objectives and outcomes related to ethical and professional responsibility, effective communications skills, ability to work as part of a team, and lifelong learning. Each ILERT partner analyzed how the framework could be applied to their program. The challenge was to maintain the program integrity and at the same time include all necessary elements of the RSIC Framework.

## II. Second Project: Student Exchanges

Subsequently, another project received support from both FIPSE and European Commission under umbrella of Atlantis Program designed to support trans-Atlantic student mobility exchanges between Europe and USA. The objective of Dependable Systems International Research and Educational Experience project (DeSIRE2) [4] was to train engineering students to achieve global competency and to provide exposure for the students to work in international settings with foreign partners, to get opportunity for exploring the host country language and customs, and to make available instruction the students in the areas not available in the home school.

The learning objectives and outcomes of DeSIRE2 project were to prepare specialists to solve problems in a multidisciplinary way utilizing wide range of CPS domains. The expected graduate student competencies provided by completion of the DeSIRE2 program were as follows:

- Demonstrate professionalism in work and grow professionally through continued learning.
- Demonstrate understanding of analysis and design to implement software-intensive systems.
- Demonstrate understanding of analysis and design to implement control systems.
- Apply advanced software engineering techniques to implement real-time concepts.
- Implement a rigorous quality assurance process.
- Implement hardware/ software integration.

- Demonstrate knowledge of the principles and techniques needed for the analysis and design of a system form dependability perspective.
- Assume a variety of roles in teams of diverse membership.

According to the initial plans, the student mobility was to be distributed between the original six consortium partner universities: three from the EU in cooperation with three universities from the US. During the first year of the project a multilateral Memorandum of Understanding (MoU) was signed by partners' schools. The memorandum stipulated the conditions of: mobility exchange period, tuition and enrollment, student eligibility, student and faculty mobility, and student selection. Due to legal issues, the third American university left consortium in the first year not contributing to the mobility exchanges.

Each consortium university partners had a dedicated unit in their organizational structure handling foreign students' exchanges and supporting the recruitment. Students received support in both pre- and post-mobility activities. The project contact faculty at each university served as the focal point to advise, guide, and help both local and the visiting students.

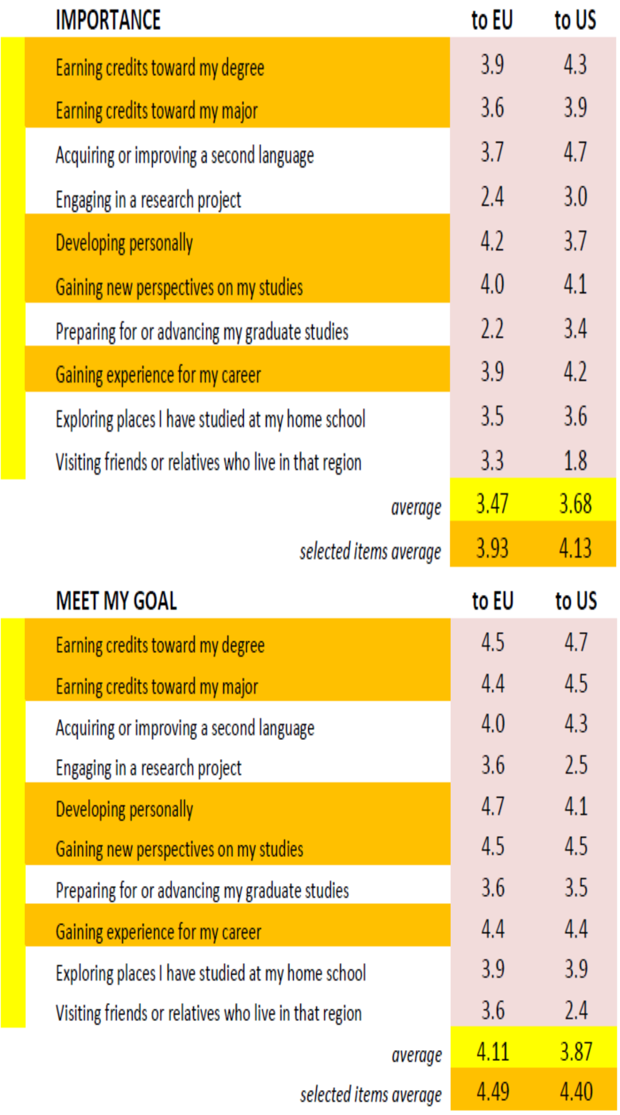
Individual program of study was developed for each candidate. The reason was that students from partner universities have had different capabilities and knowledge base due to the differences of their local academic programs. Therefore, it was necessary to shift from courses forcing the "one size fits all" approach to courses accommodating individual learning paths. Such paradigm shift was supported by the "component" approach described in the Section 2.

Due to the individual nature of each exchange and necessity to find appropriate selection of courses, the task of creating an individual learning path for a single exchange was rather complicated. The logistic of the mobility exchange required a candidate student to register for course work at the home institution in coordination with the local mobility exchange project contact faculty, who helped the student to find an appropriate set of courses in one of the overseas partners' schools. The main objective was to focus the individual plan of study in the area related to DeSIRE2 project specialization. The students then took the course overseas, with the grade passed back to the home school international office and recorded with the school registration office (or a dean office) for inclusion in the student's official record. The administrative issues such as course equivalency, satisfying the student's program core requirements, prerequisite requirements etc., were usually approved by the Faculty Dean and appropriate graduate program coordinator in advance before commencement of the mobility exchange. The approved courses were counted towards a degree at the home institution. The mechanism of ECTS/US credits equivalents, as defined by the MoU, was applied.

The consortium partners exchanged 36 students (nineteen from US studying in Europe and seventeen EU students studying in the US) achieving nearly balanced mobility exchanges participation. It needs to be noted that the consortium faced significant challenge recruiting American students. The major obstacles were constraints of academic calendar, missing opportunities of earning while abroad and thus related financial issues, and difficulties with finding appropriate set of courses to be credited to the student transcript [5].

The project requested students to complete post-mobility survey. The sample of students engaged in the mobility is still statistically insignificant, but the post-mobility surveys they filed provide interesting observations. At the project completion we collected total of 30 surveys (83% return rate): 15 from European students studying in the USA and 15 from the American students studying in Europe.

On a scale of 1 (negative) to 5 (positive), the level of satisfaction related to the program merit, in terms of importance and meeting the goals, are for American students in Europe 3.93 and 4.49 respectively, while for European students in the USA they are 4.13 and 4.40 (Figure 1).

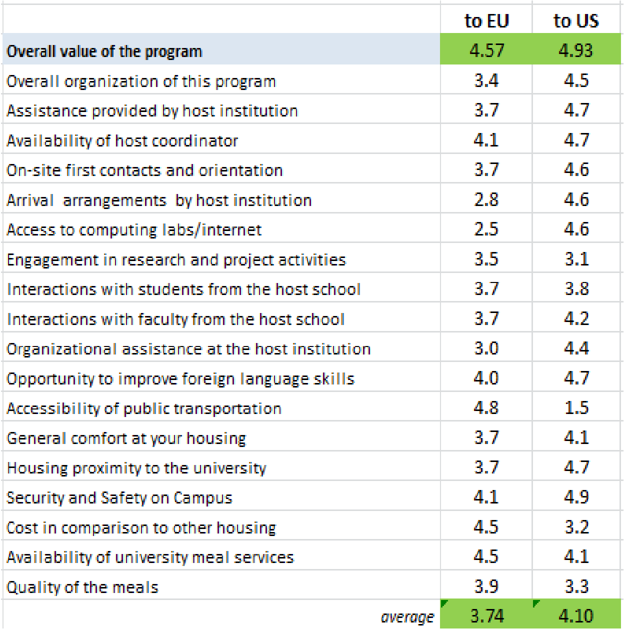


| IMPORTANCE | to EU | to US |
|---|---|---|
| Earning credits toward my degree | 3.9 | 4.3 |
| Earning credits toward my major | 3.6 | 3.9 |
| Acquiring or improving a second language | 3.7 | 4.7 |
| Engaging in a research project | 2.4 | 3.0 |
| Developing personally | 4.2 | 3.7 |
| Gaining new perspectives on my studies | 4.0 | 4.1 |
| Preparing for or advancing my graduate studies | 2.2 | 3.4 |
| Gaining experience for my career | 3.9 | 4.2 |
| Exploring places I have studied at my home school | 3.5 | 3.6 |
| Visiting friends or relatives who live in that region | 3.3 | 1.8 |
| *average* | 3.47 | 3.68 |
| *selected items average* | 3.93 | 4.13 |

| MEET MY GOAL | to EU | to US |
|---|---|---|
| Earning credits toward my degree | 4.5 | 4.7 |
| Earning credits toward my major | 4.4 | 4.5 |
| Acquiring or improving a second language | 4.0 | 4.3 |
| Engaging in a research project | 3.6 | 2.5 |
| Developing personally | 4.7 | 4.1 |
| Gaining new perspectives on my studies | 4.5 | 4.5 |
| Preparing for or advancing my graduate studies | 3.6 | 3.5 |
| Gaining experience for my career | 4.4 | 4.4 |
| Exploring places I have studied at my home school | 3.9 | 3.9 |
| Visiting friends or relatives who live in that region | 3.6 | 2.4 |
| *average* | 4.11 | 3.87 |
| *selected items average* | 4.49 | 4.40 |

Fig.1. DeSIRE mobility exchanges: importance and meeting goals

The satisfaction survey results show that the overall value of the program was nearly perfect 4.93 for European students visiting USA and 4.57 for Americans in Europe. The averages of the relevant survey items are: 4.10 for the European students in USA and slightly lower 3.74 for the Americans in Europe (Figure 2). Statistical t-test analysis allows us to state with 88% confidence that the experiences of European students in USA seem to be more positive that their counterpart in Europe.



| | to EU | to US |
|---|---|---|
| **Overall value of the program** | 4.57 | 4.93 |
| Overall organization of this program | 3.4 | 4.5 |
| Assistance provided by host institution | 3.7 | 4.7 |
| Availability of host coordinator | 4.1 | 4.7 |
| On-site first contacts and orientation | 3.7 | 4.6 |
| Arrival arrangements by host institution | 2.8 | 4.6 |
| Access to computing labs/internet | 2.5 | 4.6 |
| Engagement in research and project activities | 3.5 | 3.1 |
| Interactions with students from the host school | 3.7 | 3.8 |
| Interactions with faculty from the host school | 3.7 | 4.2 |
| Organizational assistance at the host institution | 3.0 | 4.4 |
| Opportunity to improve foreign language skills | 4.0 | 4.7 |
| Accessibility of public transportation | 4.8 | 1.5 |
| General comfort at your housing | 3.7 | 4.1 |
| Housing proximity to the university | 3.7 | 4.7 |
| Security and Safety on Campus | 4.1 | 4.9 |
| Cost in comparison to other housing | 4.5 | 3.2 |
| Availability of university meal services | 4.5 | 4.1 |
| Quality of the meals | 3.9 | 3.3 |
| *average* | 3.74 | 4.10 |

Fig.2. DeSIRE mobility exchanges: satisfaction survey

Due to the demanding nature of engineering courses conducted in English in three European partner universities, the issue of the language instruction was not a priority. However, the exchange students engaged in language and/or cultural instruction classes, participated in a variety of activities to learn the host country, interacted with their peers from the host school, and in general were satisfied with the opportunity to travel and explore different cultures. The native Slavic languages in two European partner universities were rather difficult to learn for the American students – except those with the specific country heritage that may already know the language (being typically an additional initiative to apply for the mobility exchange)

Additionally, faculty mobility exchanges were conducted to support: lectures delivery at the host institutions related to the objectives of the project, pre-mobility orientations and program promotion at host institutions, face-to-face meetings with mobility candidates and active mobility students, and planning/co-ordination meetings with the host university faculty and administration.

A dedicated website supporting the student exchange was maintained and systematically upgraded, containing up-to date information about the project objectives, activities, outcomes, progress and deliverables. One of the requirements of the mobility exchange completion was for the students to create personal blogs reflecting their experiences. The blogs were linked from the project website (http://www.desire.agh.edu.pl/).

All students engaged in the program had very positive impression of their mobility stay and appreciated the opportunity provided by the project funding. In general the students expressed positive feedback related to the general study abroad experience. Only few comments from post-mobility surveys addressed the actual classroom experience.

- To Europe:
  - *...academic system is a little harder than US because of the lack of assignments and the dominant influence of the exams (midterm and final) on your grade for the class* (ERAU->UJF).
  - *...enjoyed very much the curriculum and the hands on approach of the courses, especially the opportunity to practice our newly gained knowledge through the academic project. I believe that this experience thought me a lot on more than just academic knowledge but it also allowed me to grow on a more personal level* (UCF->UJF).
- To USA:
  - *...engineering and software departments are also very good here. There is lot of interesting projects, classes. Another great thing is that the teachers are mostly people that are working or worked in the field they are now teaching, i.e. I had Software Project Management class with the instructor from Boeing Company – a real software project manager* (AGH->ERAU)
  - *... this was a real experience for me abroad. It brought me a lot of knowledge in certain class and a new method of working* (UJF->ERAU).

## III. Implementation Examples

Hands-on learning approach was proposed for the majority of RSIC components. This approach involves learning experience which enhances student's ability to think critically. The student must plan an experiment to test a hypothesis, put the experiment into motion, see the experiment to completion, and then be able to explain and interpret the experiment's results. Another hands-on experience has been engagement in a team project requiring students to analyze the problem, specify system requirements, find appropriate solution, resulting in designing and implementing a small scale cyber-physical system. Two selected examples are provided below.

### *A. Example 1: CPS Laboratory Experiment*

The example of lab environment supporting education in Real-Time Cyber-Physical Systems is provided below [6].

The goal of this educational module (lectures and practical training) is to familiarize students with industrial Ethernet IP technology studying a model of an industrial Allen Bradley Networked Programmable Logic Controller (PLC) connected through different gateways to the physical system (model of air lift). The system uses distance measurements and implements control signals in a similar way to typical industrial process. Students became familiar with essential characteristics of data transmission networks and with the influence of the network parameters on control quality of the physical system. Students also work with several standard testing methods and items of industrial equipment.

The Ethernet IP laboratory, shown in Figure 4 consists of six data transmission nodes: CompactLogix L35E Programmable Logic Controller, POINT_IO: 1734-AENT, PowerFlex 40 inverter, WAGO 750-341 Coupler, Internet Camera (WebCam), and PanelView 600 Plus – Touch Panel.

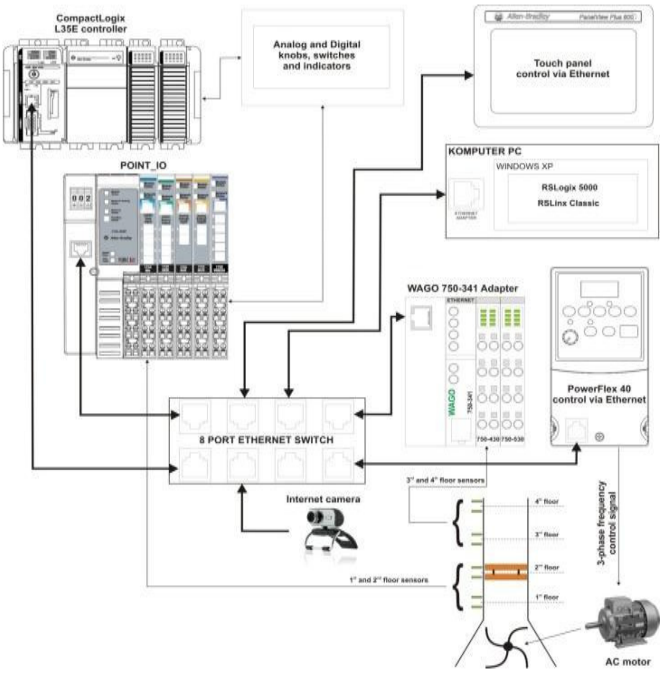


Fig.3. Laboratory Setup

A standard personal computer is used as a development and Ethernet monitoring platform and the web camera generates noise imitating the network data traffic.

The major issues for consideration by the students were: (a) analyze the Ethernet/IP packets and Ethernet network parameters (throughput, round trip time etc.) by utilizing the Wireshark application, (b) diagnose and analyze the jitter for a several scan times and the RPI parameter (Request Packet Interval), and (c) find the effects of load Ethernet network on stabilization of the position of the Aero-lift.

The physical system used in the module is Aero-lift. The test-rig contains the blower based on three-phase drive controlled by the inverter, a pipe for vertical motion of the cart and eight discrete sensors to measure the cart position (see Figure 4). The Aero-lift is an example of the dynamic system sensitive for sampling time and latency in data transmission. With this set-up a number of consequences caused by the inaccurate settings for Ethernet modules and I/O blocks can be observed.

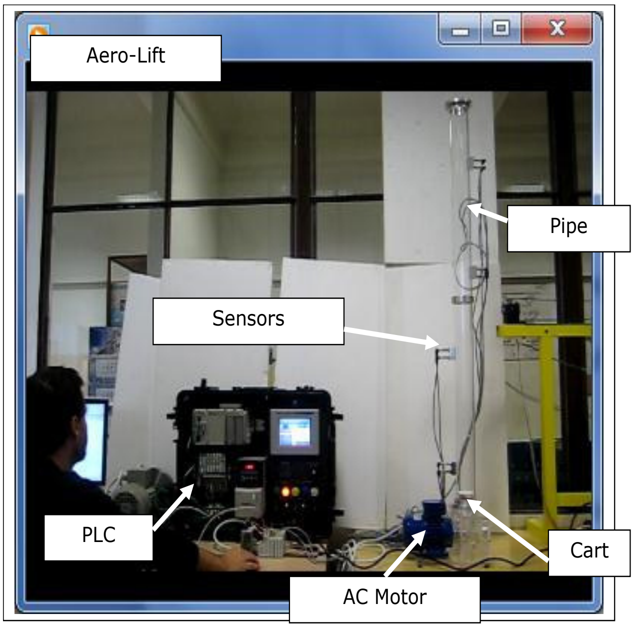


Fig.4. Physical system used in the laboratory setup

The offered laboratory demonstrates all the features and includes all components expected for Cyber Physical Systems: industrial equipment reflecting real physical plant, computing environment, time critical software, and distance data transfer. The laboratory shows that traditional real-time performance guarantees are insufficient for CPS when system components are spatially distributed and connected via data transmission networks.

## *B. Example 2: Teamwork Development of a CPS*

The consortium members recognized the demand for efficient development of quality RSIC/ CPS systems and need for an international learning environment. An innovative learning approach was implemented involving students from multilingual, geographically separated institutions in a coordinated educational experience in laboratory CPS development [7]. The project was realized as a component of the ILERT engaging on-site students and faculty of universities from four countries (USA, Poland, France, and Czech Republic).

The technical objective of the project was to develop real-time software packages for two robots (physical system) that cooperate in searching for a target in a simple maze. The target search in the maze was designed to simulate a practical example a rescue problem.

The two robots were considered to work within the limited operating area. The robots were assigned different roles: The "Leader" Robot starts in one of the four corners and, according the implemented algorithm, it is searching for the black square that is positioned in the center of the maze. The "Follower" Robot starts in one of the four corners (which may be different that the starting corner of the

Leader Robot) and according the information received from the Leader robot it attempts to find the black corner is a shortest possible time. The use of rigorous software development process was emphasized and the comprehensive documentation was expected as an integral part of the project's outcomes. Team cooperation and related intensive inter- and intra-team communication was also assumed. Each team was responsible for their developed components from the initial phases of analysis and design through component implementation to integration and testing.

The two major Project Learning Objectives were focusing on the Process and Product:

- Process - effective work in multicultural and multidisciplinary teams including such aspects as team organization and planning, team communication, and project management.
- Product - development of an RSIC system resulting in developing real-time embedded software, hardware and software system integration, implementation of digital control in real time, design/implement communication protocols.

To meet the above objectives, international students' teams have worked on robotic design and control experiments with LEGO MINDSTORMS NXT kits.

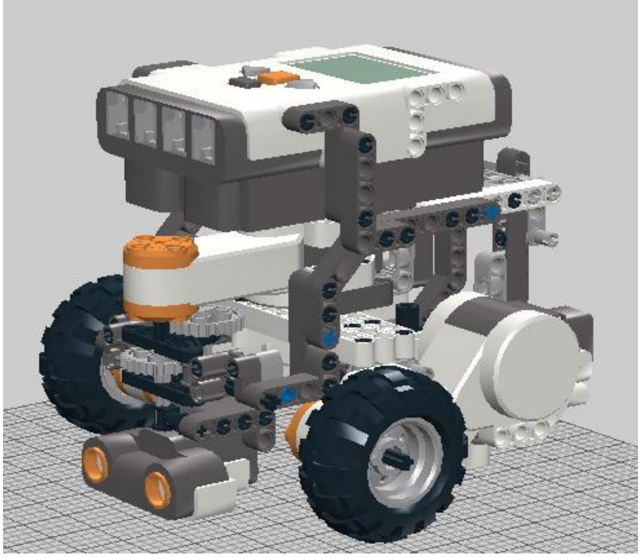


Fig.5. Design of three wheels robot construction

The basic problem was an accurate and precise motor control. The error in the motor control can affect the correct navigation and robot movements. From the educational viewpoint it was not so important that the robot moves, but how precise it operated and what was the accuracy of the controlled devices.

LEGO-MINDSTROMS NXT kit (Fig. 5) was selected as a common platform for students' teams. The used LEGO platform proved flexible enough to achieve both technical learning objectives. A Web-based Project Management System (WebPMS) was used to improve communication and test Internet based tools for inter-university collaboration.

The programming solutions related to LEGO NXT were implemented in open source NXC (Not Exactly C) - a text-based language for MINDSTORMS® NXT robot microcontroller. The language supports multitasking on standard LEGO firmware. With freely available BrixCC Integrated Development Environment (IDE) it was possible to create quite complex and advanced data acquisition and control algorithms operating concurrently and in real-time, including file management and communication features. A custom Bluetooth communication program was designed and implemented so the robot(s) movements could be monitored and displayed on the host.

## IV. Conclusion

Cyber-Physical Systems are a novel category of engineering systems, involving closely interacting physical and computational components. In the CPS, control algorithms, embedded hardware, real-time software, and communication are closely intertwined while the digital cyber components are connected interacting with the physical real-world processes. Due to interaction with the environment and people, the issues of dependability and performance while meeting the functional requirements are of primary concern. It is critical that the future CPS developers are proficient in these multidisciplinary projects while being able to work with professionals representing different skills and organizational affiliations.

The laboratory environment delivers required features and includes all components expected for Cyber Physical Systems: physical plant, computing environment, software intensive and time critical, distance data transition.
CPS technology challenges included: CPS data acquisition, data precision, time dependencies, the critical elements of modern CPS, quality assurance and integration; data and information analytics, control algorithms and rigorous software development process. In particular, assures compliance with an important CPS education criterion formulated by ABET: "ability to function effectively on multi-disciplinary teams spanning cyber and physical domains".

The international consortium of universities, with support of the U.S. Department of Education FIPSE and the European Commission projects, has been engaged in the CPS related educational activities at graduate levels since 2007. The objective of these educational activities was to educate students in the principles and design issues in CPS and to develop student competency to be comfortable working in an international setting collaborating across the nations' boundaries. It was found that it has been easier to achieve these multi-disciplinary and integration-focused knowledge and skills required for the CPS education when engaging a group of universities, rather than when attempting to achieve them within a single university. Students' feedback confirms that they truly enjoyed the program, and they expressed great interest in future participation in the CPS related activities.

To continue potential future cooperation after project completion and expiration of the multilateral agreement, the participating universities signed bi-lateral agreements stipulating continuing cooperation in the area of the CPS, with both faculty and student exchanges.

## REFERENCES

[1] ABET Engineering Criteria Program Educational Outcomes, available http://www.foundationcoalition.org/home/keycomponents/assessment_eval/ec_outcomes_summaries.html, [Accessed March 15, 2015].

[2] W. Grega, A. Kornecki, M. Sveda, and J. Thiriet, "Developing Interdisciplinary and Multinational Software Engineering Curriculum", in *Proceedings of the ICEE'07*, Coimbra, Portugal, Sep. 3-7, 2007 http://www.desire.agh.edu.pl , [Accessed March 15, 2015].

[3] A.J. Kornecki, W. Grega,, T. Hilburn, J-M. Thiriet, M. Sveda, O. Rysavy, A. Pilat,"Transatlantic Engineering Programs: An Experience in International Cooperation", "Engineering the Computer Science and IT", Chapter 5, *In-Tech Publishing*, Croatia, 2009, ISBN 978-953-307-012-4, pp. 65-84.

[4] M. Sveda, W. Grega, T. Hilburn, A. Kornecki, O. Rysavy, J-M. Thiriet, (2009), "Real-Time Software-Intensive Systems Engineering in an International Perspective", in *Proceedings of 20th EAEEIE Annual Conference*, Valencia, Spain, June 22-24, 2009.

[5] A.J. Kornecki, W. Grega, A. Gonzalez, "Europe/USA Mobility Exchange in Information Engineering: Why Is It Less Attractive to the American Students?", in *Proceedings of EIEAEE*, June 2010 Palanga, Lithuania.

[6] CoNeT Mobile Lab, Ethernet IP – Section 5, CML User Manual, Rev. 1.0, http://www.ipnet.agh.edu.pl/Materials2/Module5/Module5-UserManual.pdf, [Accessed March15, 2015].

[7] A. Pilat, A.J. Kornecki, J-M. Thiriet, W. Grega, O. Rysavy , (2009), "Inter-university Project Based on LEGO NXT" , in *Proceedings of 3rd IEEE Multi-conference on Systems and Control*, Saint Petersburg, Russia, July 8-10, 2009, ISBN: 978-1-4244-4602-5, ISSN: 1085-1992.

DOI: 10.15439/2015F69
ACSIS, Vol. 5

# Qualitative and Quantitative Evaluation of Stochastic Time Petri Nets

Franco Cicirelli, Christian Nigro, Libero Nigro
Laboratorio di Ingegneria del Software
Dipartimento di Ingegneria Informatica Modellistica Elettronica e Sistemistica
Università della Calabria, Italy
{f.cicirelli@dimes.unical.it, christian.nigro@tiscali.it, l.nigro@unical.it}.

□

*Abstract*— **Time Petri Nets (TPN) are a well-known formalism for modelling time-dependent systems with timing constraints. This paper proposes an approach based on a stochastic extension of TPN (sTPN), which enables both qualitative assessment of feasible temporal behaviors through model checking, and quantitative evaluation of a probability measure of a given behavior, by statistical model checking. The experimental work rests on the use of the latest version of the UPPAAL toolbox which supports both exhaustive non deterministic analysis and statistical model checking of system properties. The approach is demonstrated through an example.**

## I. Introduction

The development of safety-critical software systems is challenged by the needs of addressing both functional and temporal correctness issues. Violation of timing constraints can have important consequences in the practical domain (e.g., economy, medicine, cyber physical control systems, etc.). Therefore, it is highly recommended the use of formal tools for modelling and analysis of concurrent and timed software.

From the point of view of analysis, both qualitative verification and quantitative evaluation of system properties are nowadays advocated by engineers and developers. Whereas qualitative verification of a system model tries to identify feasible behaviors, quantitative evaluation aims to associate them a measure of occurrence probability.

Qualitative verification is often based on the exhaustive enumeration of all the possible execution states of a model, organized in the so called model *state graph*, and on checking desired properties through efficient traversal algorithms on the state graph. To avoid an infinite growth of the state graph, each state node is implemented as a couple: a *discrete data part*, and *a dense time part*. The time component typically stores in a compact way (*zone*) all the clock (timer) inequalities which hold in the state. As a consequence, a state is in reality a *state class* which subsumes an infinite number of states which could be reached by only changing the clock values (firing times of transitions).

Despite the use of state classes, depending on the system model, the state graph can suffer of state explosion problems. In addition, the construction and navigation of the state graph can become undecidable when a complex combination of modelling factors (non-deterministic time of actions, sporadicity of process arrival, message passing, stochastic aspects etc.) occurs. In these cases, the study of system properties can only be approximated, e.g., through simulation.

Qualitative non deterministic verification based on model checking [1] has been demonstrated by Time Petri Nets [2]-[5] and Timed Automata [6] based tools, e.g. [7]-[12]. Quantitative evaluation of system behavior is supported by recent extensions to the ORIS tool [8] and by latest versions of UPPAAL [9]-[11] which include a statistical model checker (SMC) [13].

This paper proposes an original approach to qualitative and quantitative evaluation of systems with timing constraints which is based on the Time Petri Net (TPN) formalism which is very often used for modelling real-time and embedded systems, communication protocols etc. To permit both non-deterministic analysis and stochastic analysis of system properties, a stochastic extension of TPN (sTPN) [14] is also considered. The contribution of the paper consists in a mapping of TPN/sTPN onto UPPAAL so as to exploit, on a same model, both the exhaustive model checker and the stochastic model checker.

Whereas the support of sTPN in the ORIS tool is based on the concept of *stochastic state class* and *stochastic state graph*, i.e., a density probability distribution function is attached to each state class which characterizes the possible stochastic evolutions from it, i.e., estimating the probability of the outgoing state transitions, the use of sTPN in UPPAAL rests on batches of simulation runs and statistics inference of desired results from these runs. As a consequence, ORIS can provide a greater resolution on the probability measures. However, this paper argues that the proposed approach based on UPPAAL has the following strengths: (1) it is based on a popular and efficient toolbox, (2) it does not incur in an infinite stochastic state graph nor suffer of stochastic state explosion problems as discussed in [14] (3) it in any case can provide quantitative measures of probability which are valuable from the engineering point of view.

The paper is structured as follows. Section II provides basic definitions of TPN and sTPN formalisms. Section III

□ This work was not supported by any organization

describes a modelling example. Section IV gives an overview to non-deterministic analysis and stochastic analysis enabled by UPPAAL. Section V discusses the developed structural translation from TPN/sTPN to UPPAAL. Section VI illustrates the application of the proposed approach to a thorough property assessment of the model described in Section III, by detailing qualitative and quantitative analysis. Section VII concludes the paper by indicating directions of on-going and future work.

## II. Time Petri Nets Definitions

A basic TPN is a tuple $(P, T, B, F, M_0, I_{nh}, EFT^s, LFT^s)$ where:

- $P$ is a finite nonempty set of places;
- $T$ is a finite nonempty set of transitions;
- $P \cap T = \emptyset$;
- B is the backward incidence function, $B: P \times T \to \mathbb{N}$, where $\mathbb{N}$ denotes the set of natural numbers;
- $F$ is the forward incidence function, $F: P \times T \to \mathbb{N}$;
- $I_{nh}$ is the set of inhibitor arcs, $I_{nh} \subset P \times T$ where $(p,t) \in I_{nh} \Rightarrow B(p,t) = 0$;
- $M_0$ is the initial marking function, $M_0: P \to \mathbb{N}$, which associates with each place a number of tokens;
- $EFT^s: T \to R^+$ is a function which associates each transition with a (finite) earliest static firing time. $R^+$ denotes the set of non-negative real numbers;
- $LFT^s: T \to R^+ \cup \{\infty\}$ is a function which associates each transition with a (possibly infinite) latest static firing time. In any case it must be $LFT^s \geq EFT^s$.

Differently from [13], in this work TPNs admit both inhibitor arcs and non-unitary arc weights.

The state of a TPN is a pair $s =< m, \tau >$ where $m: P \to N$ is the net marking, and $\tau: T \to R^+$ associates each transition with a (dynamic) *time-to-fire* (clock or timer). The state evolves according to the *firability* and *firing* clauses.

A transition $t$ is *enabled*, as in classic Petri nets, if each of its input places contains sufficient tokens, i.e., iff

$$\forall\, p \in P, (p,t) \in I_{nh} \Rightarrow M(p) = 0\ \wedge \\ B(p,t) > 0 \Rightarrow M(p) \geq B(p,t)$$

Transition $t$ is *firable* if it is enabled and its time-to-fire $\tau(t)$ is not higher than that of any other enabled transition.

When $t$ fires, the state $s =< m, \tau >$ is replaced by a new states $s' =< m', \tau' >$ where marking $m'$ is derived from $m$ by the withdrawal of tokens from the input places and the deposit of tokens in the output places. More precisely, the firing process consists of the two (atomic) phases:

$$m_{int}(p) = m(p) - B(p,t) \text{ (withdraw phase)} \\ m'(p) = m_{int}(p) + F(p,t) \text{ (deposit phase)}$$

Transitions which are enabled in the intermediate marking $m_{int}$ (and then also in $m$) *and* in the final marking $m'$ are said *persistent* to the $t$ firing. Transitions which are enabled in $m'$ but not in $m_{int}$ are said *newly enabled*.

A transition which is multiple enabled in a state $s$ is supposed to consume its enablings one at a time (*single server* semantics). Therefore, after its own firing, would $t$ be still enabled, it is regarded as a newly enabled one.

For any transition $t_p$ which is persistent to the firing of $t$, its time-to-fire is reduced, in the new state $s'$, as follows:

$$\tau'(t_p) = \tau(t_p) - \tau(t)$$

For any newly enabled transition $t_{ne}$ its time-to-fire is constrained to occur not deterministically in its static time interval:

$$EFT^s(t_{ne}) \leq \tau'(t_{ne}) \leq LFT^s(t_{ne})$$

### A. Stochastic extensions

An sTPN [14] specializes a basic TPN as follows. The set of transitions is partitioned into two subsets: $T = T_1 \cup T_2$, where $T_1 \subseteq T$ is the subset of *timed transitions*, $T_2 \subset T$ is the set of *immediate transitions*. Besides its static time interval, a timed transition is attached a probability density distribution function $PDF: T_1 \to R^+$, which is constrained in the static time interval $[EFT^s, LFT^s]$ of the transition. It is also said the static time interval is the *support* of the pdf. An immediate transition is attached a real positive weight $\pi: T_2 \to R^+$.

The semantics interpretation of an sTPN is as follows. Immediate transitions (as in Generalized Stochastic Petri Nets –GSPN- [15]) always fire *before* any timed transition, and consumes no time. The set of simultaneously enabled immediate transitions in the current state constitutes a *random switch*, i.e., each immediate transition $t_i$ is firable with probability

$$Prob(t_i) = \frac{\pi(t_i)}{\sum_{t_j \in T_2 \text{ and } t_j \text{ is enabled}} \pi(t_j)}$$

The time-to-fire $\tau(t)$ of a timed transition $t$ is stochastically defined, at its enabling instant, by sampling the $PDF(t)$ with the constraint:

$$EFT^s(t) \leq PDF(t) \leq LFT^s(t)$$

As an example, the $PDF$ of a timed transition can be (default) a uniform distribution function which picks up a value in the static time interval of the transition, or a negative exponential distribution function. However, in the general case, a timed transition can follow a generally distributed function constrained in the support time interval.

Firing of a timed transition follows the same rules as in basic TPNs.

## III. A Modelling Example

Fig. 1 depicts a TPN model [14] made up of two (almost identical) production cells, identified by suffixes 1 and 2, which operate sequentially and cyclically.

Each cell admits two parallel activities named JobA (t1, t7) and JobB (t2, t8). JobA requires the use of a resource res (places p2 and p9) which may fail during the usage. In the case of failure, JobA is not completed, and a recovery action recA (t5 and t11) is instead executed which replaces the normal behavior provided by JobA. A failed resource is repaired by a repair transition (see t3 and t9). The execution of a production cell is started by the start transition (t0, t6). A cell logically terminates its current production phase when a token is generated in the couple of termination places (p5, p6) or (p12, p13).

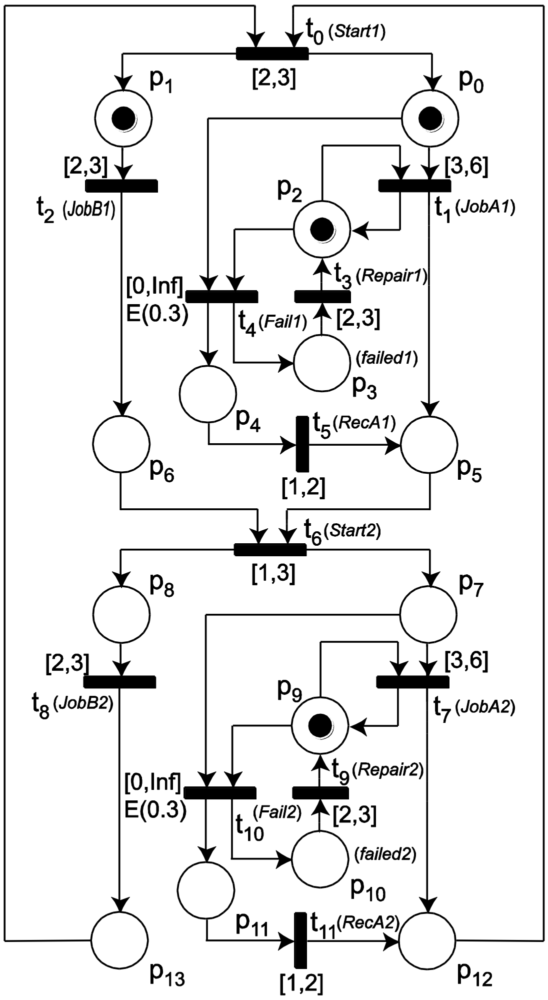


Fig. 1 A TPN/sTPN model with two production cells

All the transitions in Fig. 1 are supposed, in the stochastic interpretation, to be served by a uniform distribution in the associated static time interval, with the exception of the failure transitions (t4, t10) which have a support interval of [0,∞] and an exponential distribution function whose rate is $\lambda = 0.3$, as witnessed by the notation E(0.3) attached to the transitions.

The use of [0,∞] as the time interval of a failure transition is noteworthy. In both non-deterministic and stochastic interpretations the failure cannot occur later than 6 time units measured from the time instant in which the failure transition gets enabled. This happens because of the [3,6] time interval of the JobA transition, which forbids the failure to occur later than 6. In the stochastic interpretation, however, something subtle occurs. Due to the exponential distribution E(0.3), the failure is expected to happen with a very low probability, as the sample chosen from the E(0.3) pdf can be much greater than 6 and thus later with respect to the completion time of the JobA activity. All of this reflects the different concerns of non-deterministic analysis and stochastic analysis (see later in this paper).

## IV. Uppaal Concepts

The popular and efficient Uppaal toolbox [9]-[10] allows modelling and verification of time-dependent systems. An Uppaal model consists of a network of timed automata (TA) [6]. TA are designed as *template processes*, which can have parameters, can be instantiated, and consist of *atomic actions*.

TA are extended with local or global integer (and boolean) variables and arrays of integers, clocks and channels. In latest versions of the toolbox, C-like functions and structures are also permitted. Time is dense and can be controlled by means of *clock* variables. Clocks can only be reset and compared against nonnegative integer constants.

All the clocks of a model increase automatically with the same rate of advancement of the hidden system time. TA synchronize to one another by CSP-like channels (*rendezvous*) which carry no data values.

Asynchronous communication is provided by broadcast channels where a single sender can synchronize with a (possibly empty) group of receivers. The sender of a broadcast signal in no case is blocked. Locations (states) of an automaton are linked by *edges* (transitions).

Every edge can be annotated by a *command* with three (optional) elements: (i) a *guard*, (ii) a *synchronization* (? for input and ! for output) on a channel, and (iii) an *update* consisting of a set of clock resets and a list of variable assignments. Channel synchronization implies the commands of the sender and of the receiver(s) are jointly executed. However, the update of an output command is executed *before* that of a matching input command.

An atomic action consumes no time and refers either to the execution of an internal command of one automaton or to a joint execution of the multiple commands during a synchronization among two or more TA.

A clock *invariant* can be attached to a location as a *progress* condition. The timed automaton can remain into the location as long as its invariant holds. Uppaal supports also *committed* and *urgent* locations which must be exited immediately (i.e., without passage of time), and *urgent channels* whose synchronizations must be fired without passage of time. Committed locations, among them, can be interleaved. Similarly, urgent locations, among them, can be interleaved. However, committed locations have priority with respect to urgent locations.

The symbolic model checker of UPPAAL handles the parallel composition of the TA of a model. Parallel composition means generating all the possible action interleavings of the component concurrent processes.

UPPAAL consists of a graphical editor, a simulator and a verifier (the symbolic model checker verifyta). For exhaustive property assessment, the verifier tries to build the reachability graph of the model, where each state node holds a *data part* (variable values and location of each automaton) and a *firing domain* (time zone or clock inequalities system). The time zone implies each state graph node actually represents a *class* of equivalent states which fulfill the clock inequalities. The simulator executes a system model and visually documents the reached execution state by following a particular path in the model state graph. The simulator is useful for model debugging and to examine a diagnostic trace (counterexample) created by the verifier, e.g., when a property is not satisfied. Safety, absence of deadlocks, and bounded liveness (e.g., an end-to-end time constraint) properties can be verified by reachability analysis upon the state graph, using a subset of TCTL temporal logic formulas [10] as shown in the following:

- $E <> \varphi$ (Possibly $\psi$, i.e., a state exists where $\psi$ holds)
- $A[]\ \psi$ (Invariantly $\psi$, equivalent to: $not\ E <> not\ \psi$)
- $E[]\ \psi$ (Potentially Always $\psi$, i.e. a state path exists over which $\psi$ always holds)
- $A <> \psi$ (Always eventually $\psi$, equivalent to: $not\ E[]\ not\ \psi$)
- $\psi --> \xi$ ($\psi$ always *leads-to* $\xi$, equivalent to: $A[]\ (\psi\ imply\ A <> \xi)$)

where $\psi$ and $\xi$ are state properties (formulas), e.g., clock constraints or boolean expressions over predicates on locations.

Although min/max determinations, e.g., of a clock, can be achieved by using the above described logic queries, a shorthand notation is available as in the following:

sup{ state-predicate } : list-of-expressions
inf{ state-predicate } : list-of-expressions

These queries evaluate the superior/inferior value of the list of expressions, only in the states of the state-graph which satisfy the state predicate specified within { and }.

### A. UPPAAL Statistical Model Checker

The problem with symbolic model checking is that it could not be practically applied to realistic complex systems which generate an enormous (possible infinite) state graph, or it becomes undecidable for systems which combine in a complex way continuous time with stochastic behavior.

Property checking in these cases can only be approximated or estimated. In recent years the UPPAAL toolbox was extended to support stochastic model checking (SMC) [9]. UPPAAL SMC [11] avoids the construction of the state graph and checks properties by performing a certain number of simulation runs, e.g., in parallel on a modern multi-core machine. After that some statistics techniques are used to infer results from the simulation runs.

SMC refines and extends basic UPPAAL. Only broadcast synchronizations are allowed among stochastic TA (STA). In addition, either an invariant or the rate of an exponential distribution can be attached to a location. Stopwatches, i.e. clocks whose automatic advancement can be temporarily stopped (their first derivative is put equal to zero as an invariant of a location) can be exploited. A stopwatch resumes its advancement as soon as the automaton exits the location in which it was stopped.

UPPAAL SMC also provides floating point (double) variables which, e.g., can be assigned the value of a clock. Virtually, the symbolic model checker can be applied to a stochastic model too, in which case all doubles, exponential distribution rates etc. are simply ignored. However, on a stochastic TA model can be issued the following specific query types. Bold symbols are meta-symbols used to describe the SMC query language.

1. simulate N [ **(**clock|#|void**)** <=bound ] { Expression1, …, Expressionk }
2. Pr[ **(**clock|#|void**)** <=bound ] ( **(**<>|[]**)** Expression )
3. Pr[ **(**clock|#|void**)** <=bound ] ( **(**<>|[]**)** Expression) **(**<=|>=**)** PROB
4. Pr[ **(**clock|#|void**)** <=bound ] ( **(**<>|[]**)** Expression) **(**<=|>=**)** Pr[ **(**clock|#|void**)** <=bound ] ( **(**<>|[]**)** Expression)
5. E[ **(**clock|#|void**)** <=bound; N ] ( **(**min:|max:**)** Expression )

Expressions are state predicates without side-effects. They can specify an automaton to be in a certain location, or some constraints on data variables or clocks etc. All the queries are evaluated according to a bound which can be related to (implicit) global time or to a clock or to a number of simulation steps (#).

Query 1 makes N simulation experiments and collects information about the listed expressions. Query 2 evaluates the probability the given expression holds within the assigned bound (<>) or always holds within the bound ([]) with a confidence interval (the default 95% confidence degree can be customized by the user). Query 3 checks if the estimated probability is less/greater than a given probability value. Query 4 compares two probabilities. Query 5 estimates the minimum or the maximum value of an expression.

Responding to queries implies a certain number of simulation runs are carried out, either explicitly requested (see the parameter N in the queries 1 and 5) or implicitly defined by the query. Quantitative estimation of a query of type 2 rests on Monte Carlo-like simulations. Qualitative queries of the type 3 and 4 use sequential hypothesis testing. An important feature provided by UPPAAL SMC is *visualization* of simulation results. Following a satisfied query, the modeler can right click on the executed query and choose an available diagram (histogram, probability distribution etc.) to be plotted. At the time of this writing, UPPAAL SMC is supported by the development version 4.1.19.

## V. MAPPING TPN/sTPN ONTO UPPAAL

A TPN/sTPN model is translated into UPPAAL by associating each transition with a suitable template process and by introducing some global data and helper functions. A similar approach was adopted by authors in [16] which provides a formal correctness approach exploitable also in current work. For brevity, though, in the following only an informal semantics will be given.

### *A. TPN issues*

The structural translation adapts itself to the needs of both non deterministic analysis (TPN model checking) and the stochastic analysis (sTPN SMC). During the exhaustive verification of a TPN model, all the transitions are homogeneously modelled as timed transitions (see the tTransition automaton in Fig. 2). Immediate transitions, in particular, are expressed as timed transitions with a [0,0] static time interval. Random switches are thus replaced by non-deterministic selection (*race*) among transitions firable at the same time.

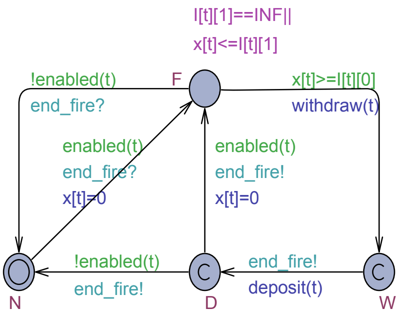


Fig. 2 The tTransition template automaton

The tTransition automaton in Fig. 2, with the sole parameter const tid t, models the generic timed transition of a TPN model. The t transition id is used to select the clock x[t] which captures the time-to-fire of the transition. The global functions enabled(t), withdraw(t) and deposit(t) assist the automaton evolution.

A timed transition starts into the N (Not enabled) location. As soon as it finds itself enabled, it resets its clock x[t] and moves to the F (Firable) location. The invariant attached to the F location states that the transition can remain in F until the latest static firing time which can be infinite. On the other hand, transition t can fire as soon as it happens its clock x[t] reaches the earliest static firing time. Bounds of the static time interval of the transition t, constrained to be positive integers (an upper bound infinite is denoted by the constant INF), are held into the global array $I: T \times 2 \rightarrow int$. $I[t][0]$ is the $EFT^s(t)$, $I[t][1]$ is the $LFT^s(t)$.

The firing process is accomplished with the help of the end_fire broadcast channel. Two broadcast synchronizations are actually used, separately triggering the withdraw-phase and the deposit-phase. Each end_fire synchronization is capable of influencing all the other transitions which are forced to re-examine their enabling status, and thus commanding a return from F to N when a firable transition detects it is no longer enabled, or moving from N to F when a not enabled transition finds itself enabled. It is worth noting that the two location W (Withdraw) and D (Deposit) are committed locations, thus ensuring the firing process is instantaneous and atomic. In the D location, a just fired transition which is still enabled, comes again into the F location and resets its clock x[t] (single server semantics). If the transition is not enabled, it reaches the default N location.

To bootstrap a TPN model, a Starter automaton (see Fig. 3) is exploited which launches an initial (fictitious) end_fire synchronization to force transitions which are enabled in the initial marking of the model, to reach the F location.

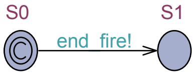


Fig. 3 The Starter automaton

Behind the design of the tTransition automaton, the following global declarations are introduced. Model topology is captured by the constants P (number of places), PRE (maximum number of input places of a transition), POST (maximum number of output places of a transition), T (number of transitions), B and F (input/output constant incidence matrices), M (marking vector), I (time intervals). Each element of the matrices B and F, purposely implemented as TxPRE and TxPOST respectively, holds the index of a place, and the weight of an input and, respectively, an output arc. Transitions are numbered from 0 to T-1. The subtypes tid and pid describe respectively the sub range types of possible transition or place ids.

### *B. sTPN issues*

The support of sTPN shares the same basic global declarations described in the previous sub-section plus some specific declarations required for statistical model checking.

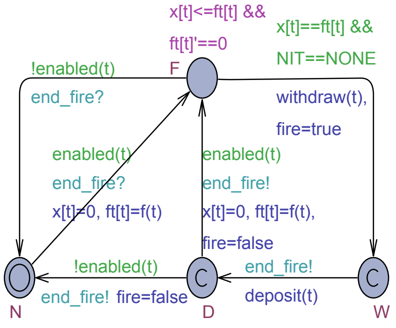


Fig. 4 The sTransition automaton

An sTPN model rests on two new timed automata shown in Fig. 4 and Fig. 5, respectively associated to a stochastic timed transition (sTransition) and to an immediate transition (iTransition).

As a matter of convention, first are numbered the stochastic transitions, from 0 to the TT-1 (TT is a model constant), then are numbered the immediate transitions (from TT to T–1). Of course, T=TT+IT. Correspondingly are defined the subtypes ttid and itid describing respectively the subranges of the stochastic and immediate transitions. An sTransition has the only the t (id) parameter of type ttid. Similarly, iTransition has the sole t parameter of type itid. All of this ensures the transition automata of a TPN/sTPN model can be implicitly instantiated at system configuration time.

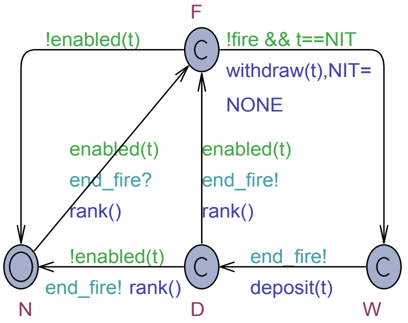


Fig. 5 The iTransition automaton

The sTransition uses two clocks: x[t] and ft[t]. The first one serves the same purposes discussed for the tTransition automaton (see Fig. 2). The second one actually stores the sampled fire time achieved (through integration of) the density distribution function (denoted by f(t)) associated to the transition. As one can see from Fig. 4, an sTransition actually fires (exiting the F location) when its time-to-fire clock x[t] equals the fire time held in the ft[t] clock. As part of the invariant attached to F, the first derivative of the clock ft[t] is put to 0 to avoid its advancement. Other details in Fig. 4 are as for the tTransition automaton.

The iTransition automaton in Fig. 5 does not use any clock. When it finds itself enabled, it moves to the committed location F. The rank() function implements the random switch and assigns to the global NIT variable (Next Immediate Transition) the id of the immediate transition probabilistically selected through the transition weights. The immediate transition that detects it is the selected NIT, instantly fires and resets NIT to NONE in preparation of a next execution of the random switch. As soon as a firable immediate transition discovers it is no longer enabled (for a conflict) it comes back to N.

Due to the committed location F in Fig. 5, and the (sub) guard NIT==NONE in Fig. 4, it is guaranteed that immediate transitions get fired before timed transitions. Moreover, the global bool fire variable in Fig. 4 and Fig. 5 ensures the fire of an immediate transition cannot actually start if the firing in progress of a timed transition is not completed.

## VI. Analysis of a TPN/sTPN Model

The structural translation from TPN/sTPN to UPPAAL described in the previous section, was applied in a case to the model presented in Section III.

### *A. Non deterministic analysis*

Here the concern was to evaluate qualitative properties of the model. Towards this timed transitions and immediate transitions are both modelled with the tTransition automaton in Fig. 2, and the exhaustive model checker of UPPAAL was used along with the construction of the (hopefully finite) state graph. The following clarifies the system configuration of the TPN model:

```
system Starter, tTransition;
```

where implicit instantiation is exploited. Due to the design of the template processes, only one instance of Starter is created, whereas T instances of tTransition (as demanded by the tid sub range type) are generated.

As a preliminary assessment of system behavior, the model was checked for the absence of deadlocks. The query was:

```
A[] !deadlock                              satisfied
```

where deadlock is a reserved keyword of UPPAAL. As a part of this first query, it was also implicitly assessed that the model is *safe*, i.e., 1-bounded. This was simply achieved by having declared the marking vector thus:

```
int[0,1] M[P]={ 1,1,1,0,0,0,0,0,0,1,0,0,0,0 };
```

and never observing an assignment out of bounds to M.

After that the model was checked for the existence of *regenerative states* [14]. A regenerative state is one where previous history of the system can be disregarded because it does not influence the future behavior.

First it was checked the property stated in [14], page 715, that repair2 cannot be persistent at each firing of start1, thus (also activating the generation of a diagnostic trace):

```
A[] (tTransition(0).W||tTransition(0).D) imply !tTransition(9).F
```

which asks if invariantly in both the intermediate marking (determined by the withdraw phase) and the final marking (determined by the deposit phase) of the firing of t0 (start1), the transition t9 (repair2) cannot be in its firable location F.

The property was found to be false as witnessed by the counterexample proposed by UPPAAL that shows effectively repair2 can be persistent to start1.

In reality, the two production cells could admit regenerative states when a *complete* processing of the second cell certainly finds the first cell in its home marking, and vice versa: when the first cell finishes its processing, the second cell *is* in its home marking. The following queries were issued to the model checker (recall places are uniquely identified by their subscripts, i.e., the numbers from 0 to P−1):

```
A[] M[5]==1 && M[6]==1 imply (M[9]==1 &&
   forall(i:int[7,13]) i!=9 imply M[i]==0)                satisfied

A[] M[12]==1 && M[13]==1 imply (M[2]==1 &&
   forall(i:int[0,6]) i!=2 imply M[i]==0)                 satisfied
```

which verify if, invariantly, each time one cell finishes its activities, i.e., a token is generated in the terminal places p5 and p6 or p12 and p13 in Fig. 1, the marking of the partner cell is surely its initial marking, meaning that no previous behavior is still in progress in the partner cell. As it was confirmed previously, e.g., at the processing end of the second cell the only possibly still pending sub activity in the second cell is the repairing of the failed res2 resource. However, satisfaction of latter queries guarantee that at each termination of the first cell, the second one always founds itself in its home marking, thus no previous behavior exists in the second cell which could interfere with new behavior.

The existence of regenerative states was a key for assessing specific behaviors of the model. For example, it was first checked about the possibility for the two resources to be failed simultaneously:

```
E<> M[3]==1 && M[10]==1                                   satisfied
```

Having found a double failed state exists, the next step was to identify *all* the possible sequences of transition firings which bring the system into the double failed state. Towards this an array s of transition ids (type tid) was introduced, together with a top variable, to memorize in the occurrence order a transition firing sequence.

The problem, obviously, was the proper dimensioning of the array s. Here it was fundamental the previous study on the regenerative states. In the light of the previous results, it was possible to create the array s with T elements, one per transition. In fact, when e.g. the second cell finishes its processing, there is no need to store again transition firings occurring in the first cell. In particular, it were also detected the instants when the array s can be safely reset in order to help model checking by forcing in s a default value. For instance, when one studies the sequence *first-cell-to-second-cell*, at the time of firing of the start1 (i.e., t0) the array s can be reset. Similarly, when the behavior *second-cell-to-first-cell* is observed, at the subsequent firing of start2 (i.e., t6) the array s can be reset.

Together with the array s, a fired(t) boolean function was introduced to check specifically for the firing of a certain transition. As an example, the existence of the double failed state was also assessed by the query:

```
E<> M[3]==1 && M[10]==1 && !fired(0)          satisfied
```

launched with the initial marking shown in Fig. 1. Now, a state with the two resources failed was checked without an intervening firing of start1, i.e., of the t0 transition, which would begin a new system execution.

Moreover, before launching the previous query, it was also asked to the model checker to generate a diagnostic trace. This way, with the property verified, it was inspected in the simulator both the sequence of events which brings the model in the double failed state, and the values of top and the first top values of s to identify the firing sequence. Therefore, as a further benefit of the previous existential query, it emerged that the sequence: t2-t4-t5-t6-t10 is capable of installing the double failed state.

The analysis was also enriched by the introduction of a global *decoration clock* y which is reset initially and at each firing of start1 (t0), and its value checked as soon as the double failed state is reached. Such a clock permits to measure the time needed, in the best and worst case, to reach the goal state. The conditional reset operation is simply added to the deposit(tid) function. An additional global decoration clock z was introduced to measure the *sojourn time* in the double failed state. The clock z is reset when the goal state is reached, and reset again as soon as the goal state is abandoned. Finding the minimal/maximal values of z provided the sojourn time. As a side benefit, watching the data and constraints (clock values) in the simulator, at the conclusion of the previous query, suggested the interval [4,9] for the possible duration of the sequence t2-t4-t5-t6-t10.

The following query was issued to find, if there are any, a new sequence:

```
E<> M[3]==1 && M[10]==1 && !fired(0) &&
!(s[0]==2 && s[1]==4 && s[2]==5 && s[3]==6 &&s[4]==10)
                                                  satisfied
```

It emerged the new sequence: t4-t2-t5-t6-t10 with a duration of [4,6] time units. The query:

```
E<> M[failed1]==1 && M[failed2]==1 && !fired(t0) &&
!(s[0]==2 && s[1]==4 && s[2]==5 && s[3]==6 &&s[4]==10) &&
!(s[0]==4 && s[1]==2 && s[2]==5 && s[3]==6 && s[4]==10)
                                                  satisfied
```

allowed to discover the third sequence t4-t5-t2-t6-t10 with duration [3,5]. Continuing with updating the query, it emerged that no more sequences exist. The sojourn time for the three firing sequences was found to be in interval [0,1].

By changing the initial marking in Fig. 1 so as to start the execution from the second cell towards the first, and repeating the work of firing sequence detection, it emerged that only other three sequences exist for this new scenario (although in [14] the existence of 7 firing paths in this second scenario was cited but not detailed). Results of the two scenarios are summarized in the Table 1 and Table 2.

As one can see from Fig. 1, a slight temporal difference exists between the first and the second cell, and is concerned with the lower bound of the time intervals of start1 and start2 which are respectively [2,3] and [1,3]. In the case both cells have the same [1,3] start interval, as expected, the paths in the two scenarios will have an identical temporal behavior. In [14] it is shown that the use of identical time intervals for start1 and start2 implies an infinite stochastic state graph which prevents the analysis. Such problems do not arise in the use of UPPAAL.

TABLE 1 FIRING SEQUENCES TO DOUBLE FAILED STATE FROM FIRST-TO-SECOND CELL

| | Firing sequence | Duration | Sojourn time |
|---|---|---|---|
| ρ0 | t2-t4-t5-t6-t10 | [4,9] | [0,1] |
| ρ1 | t4-t2-t5-t6-t10 | [3,6] | |
| ρ2 | t4-t5-t2-t6-t10 | [3,5] | |

TABLE 2 FIRING SEQUENCES TO DOUBLE FAILED STATE FROM SECOND-TO-FIRST CELL

| | Firing sequence | Duration | Sojourn time |
|---|---|---|---|
| ρ3 | t8-t10-t11-t0-t4 | [5,9] | [0,0] |
| ρ4 | t10-t8-t11-t0-t4 | [4,6] | |
| ρ5 | t10-t11-t8-t0-t4 | [4,5] | |

A further investigation was finally devoted to finding the *end-to-end* delay of a whole operation of the two production cells, i.e., measuring the min/max time required to generate the tokens in the places p12 and p13. The clock y was specialized to this new purpose: it is now reset initially and at each firing of start1 (t0) and checked when p12 and p13 have a token. The two queries:

```
inf{ M[12]==1 && M[13]==1 } : y
sup{ M[12]==1 && M[13]==1 } : y
```

furnished the time window [5,22] for the end-to-end delay.

The model checking work confirmed results achieved in [14] using the ORIS tool, but also suggested some new details.

Qualitative analysis of TPN models is rather efficient. Its scalability ultimately depends on the model topology (number of places and number of timed transitions and associated clocks) and the degree of the necessary model boundedness. As discussed in [16], what is really critical is the number of *active clocks* (i.e., the number of simultaneously fireable timed transitions) mirroring the parallelism degree of the model. An active clock is one which is growing and whose value is checked in a subsequent invariant or edge guard.

### *A. Stochastic analysis*

Whereas the non-deterministic analysis based on model checking says something can happen in the system, i.e., it predicates about qualitative concerns (properties) of the system, the stochastic analysis aims to determine a quantitative measure of the probability with which a given property can actually occur in the system.

As a concrete example, UPPAAL SMC was applied to the model in Fig. 1. To this end the system model was configured by using the sTransition automaton for the timed transitions, and the iTransition automaton for the immediate transitions (which are not used in Fig. 1).

As a preliminary step, it was checked the time limit to use for simulations. As an example, transitions start1 (t0) and start2 (t6) were monitored by cumulating their service time samples and by counting their number of firings. Then the following query was issued:

```
simulate 1 [<=100000] { t0mst, t6mst }
```

where t0mst and t6mst are decoration variables (of type double) added to the model for SMC, reflecting the monitored service time means of the two selected uniform distributions (expected values 2.5 and 2). Fig. 6 confirms that after about $3*10^4$ the mean values are reached indicating an end of transient behavior.

The following query was used to check the probability of occurrence of a single failure in the resources (time is a decoration clock mirroring system time advancement):

```
Pr[<=100000] (<>time>=30000 && (M[3]==1 || M[10]==1) )
```

UPPAAL SMC determined for the event, using 36 simulation runs, a confidence interval (CI) of [0.902606,1] with a confidence degree of 95%. Moreover, it proposed a time span like that depicted in Fig. 7 to highlight a probability distribution for the event:

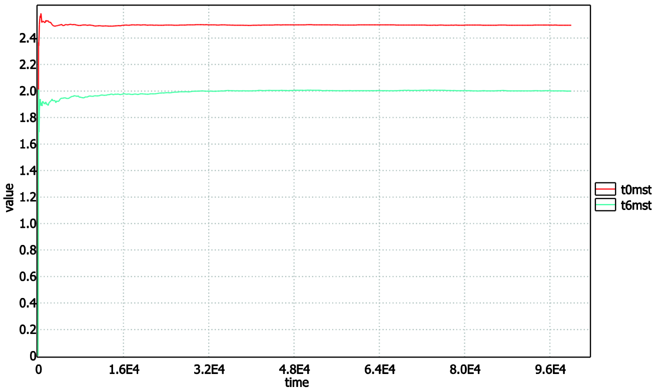
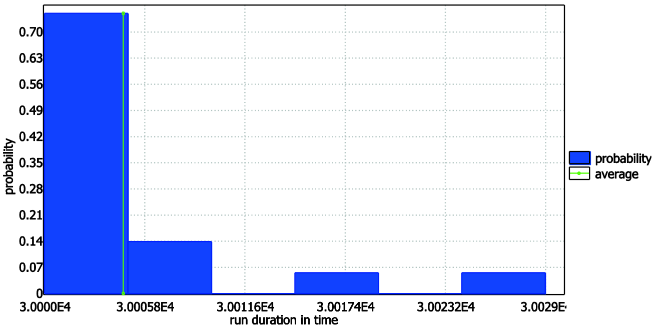


Fig. 6 Monitored values of t0 and t4 mean service time



```
Parameters: α=0.05, ε=0.05, bucket width=4.831, bucket count=6
Runs: 36 in total, 36 (100%) displayed, 0 (0%) remaining
Span of displayed sample: [30000, 30028.9858340878]
Mean of displayed sample: 30004.5594488342 ± 2.39910895990846 (95% CI)
```

Fig. 7 A probability distribution for the single failure

In a similar way it was checked the probability of occurrence of a simultaneous double failure of resources:

Pr[<=100000] (<>time>=30000 && (M[3]==1 && M[10]==1))

The event has still a CI of [0.902606,1] 95%, with a suggested probability distribution depicted in Fig. 8.

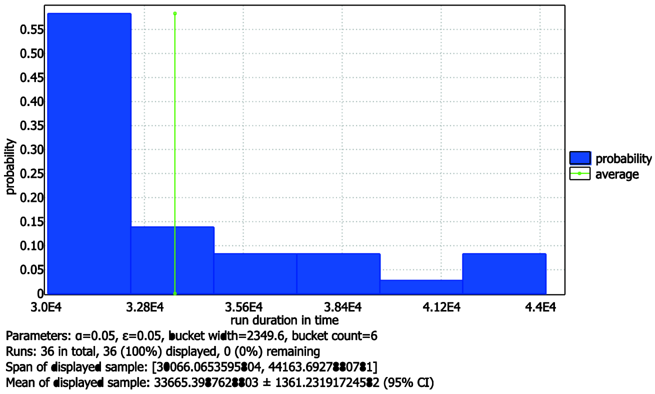


Fig. 8 A probability distribution for the double failure

All the firing sequences in Table 1 and Table 2 have a very low probability of occurrence. Each event was checked as exemplified by the following query:

Pr[<=100000] (<>time>=30000 && s[0]==2 && s[1]==4 && s[2]==5 && s[3]==6 && s[4]==10)

For all the sequence paths a CI [0,0.0973938] 95% was indicated by UPPAAL SMC.

Some experiments were specifically devoted to quantify the probability of achieving a given end-to-end delay (EED), i.e., the elapsed time from a firing of start1 (t0) to the generation of a token in both places p12 and p13 of Fig. 1. The next query was used to assess the probability of having an EED<=10:

Pr[<=100000] (<>time>=30000 && (M[12]==1 && M[13]==1) && y<=10)

Clock y is reset at each firing of t0 and measured at the end of the execution of the second cell. UPPAAL SMC proposes a CI of [0.902606,1] 95% and a cumulative probability distribution as portrayed in Fig. 9.

The probability of having an EED<=10 was checked against 80% as follows:

Pr[<=100000] (<>time>=30000 && (M[12]==1 && M[13]==1) && y<=10) >= 0.80

UPPAAL responds that such a probability is >=0.81 with 95% of confidence.

Monitored values during simulation of the end-to-end delay, collected in Fig. 10, were achieved by the query:

simulate 1 [<=100000] { EED }

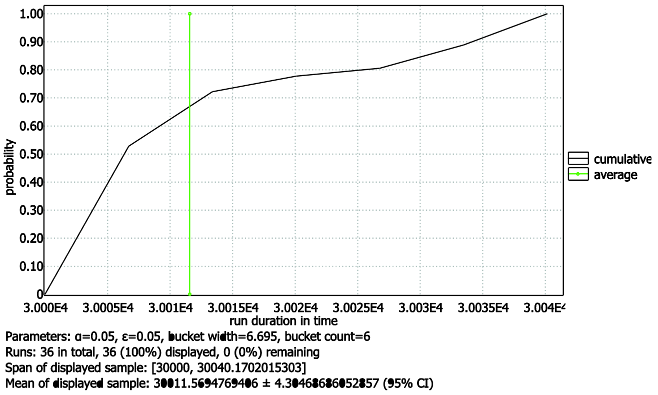


Fig. 9 A cumulative probability distribution for an EED<=10

EED is a decoration variable which receives the value of the clock y at each end of the second cell execution. From Fig. 10 it emerged an average value for the EED of about 10 tu.

Further property estimation was based on MITL formulas [11] which can provide more tight answers. The following query checks the probability that starting from the marking M[0]==1 && M[2]==1 it can follow within 3 time units a failure of res1:

Pr( M[0]==1 && M[2]==1 U[0,3] M[3]==1 )

Using 738 runs, [0,0.05] was returned as a 95% CI, indicating that the event happens with a very low probability.

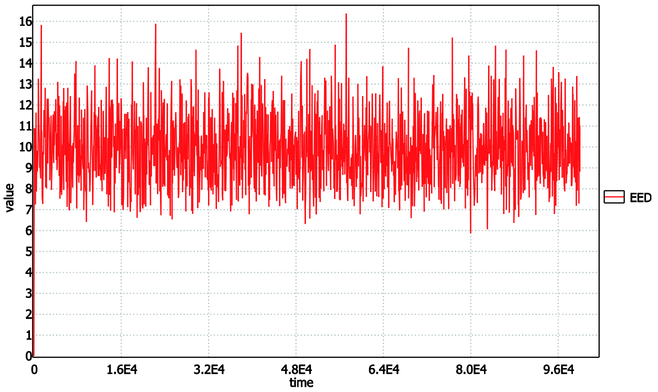


Fig. 10 Monitored EED value

The probability that within [0,1000] tu the double failure state can occur was re-checked with the query:

Pr( <>[0,1000] M[3]==1 && M[10]==1 )

which returned a [0.142412,0.242412] 95% CI. The query:

Pr( <>[0,10] M[12]==1 && M[13]==1 )

was used to evaluate the probability that within 10 tu the double cell terminates its execution. The [0.605827,0.705827] 95% CI was observed.

To check specifically the termination event within [10,22] tu it was issued the query:

```
Pr( <>[10,22] M[12]==1 && M[13]==1 )
```

which returned a [0.849729,0.949729] 95% CI to witness a very high probability of occurrence.

The occurrence probability of the firing sequences leading to double failure, was re-checked by the query:

```
Pr( <>[0,1000] top==5 && s[0]==2 && s[1]==4 && s[2]==5 &&
s[3]==6 && s[4]==10 )
```

for which (using 738 runs) a [0,0.054065] 95% CI was proposed. Similarly, for the second and third path of Table 1, it was estimated respectively a CI of [0,0.05] and [0,0.05271] 95%.

Since stochastic analysis depends on batches of simulation runs, that is UPPAAL SMC renounces to build the model state graph, there are no scalability problems when the model admits more clocks, variables etc. All of this was exploited in the case study by introducing a tailored decoration which is simply ignored when the same model is interpreted for non-deterministic analysis.

## VII. CONCLUSIONS

This paper proposes an approach to modelling and analysis of Time Petri Nets (TPN) [2]-[5] also in the presence of stochastic features (sTPN) [14].

The approach is centered on a structural translation of TPN/sTPN onto the latest version of UPPAAL which enables both qualitative non-deterministic analysis based on exhaustive model checking, and quantitative evaluation of system properties by exploiting the statistical model checker.

The UPPAAL translation makes it possible, during analysis, to reason directly in the terms of TPN/sTPN vocabulary (e.g., marking reachability and transition firings), thus simplifying its practical usage.

Prosecution of the research is directed at:

- Extending the sTPN modelling through the support of generally distributed probability functions attached to transitions.
- Improving the automatic translation of TPN/sTPN models in the context of the TPN Designer toolbox [20].
- Applying the approach, e.g., to complex maintenance procedures, phased-mission systems [17], time-constrained workflow systems.
- Experimenting the methodology with the PRISM Probabilistic Model Checker [12].
- Specializing the approach to Preemptive Time Petri Nets [18], i.e., towards the schedulability analysis of real-time tasking sets under general conditions, when the schedulability problem becomes undecidable. A preliminary framework was prototyped in [19] using UPPAAL stopwatches and over-approximation. The goal is to allow non-deterministic qualitative analysis (when possible) and quantitative analysis of task response times using the statistical model checker.

## REFERENCES

[1] E.M. Clarke, O. Grumberg, D.A. Peled, *Model checking*, MIT Press, 2000.

[2] P.M. Merlin, D.J. Farber, "Recoverability of communication protocols: implications of a theoretical study", *IEEE Trans. Commun.*, **24**(9):1036-1043, 1976.

[3] B. Berthomieu, M. Diaz, "Modeling and verification of time dependent systems using Time Petri Nets," *IEEE Trans. Soft. Eng.*, Vol. 17, No. 3, pp. 259-273, Mar. 1991.

[4] B. Berthomieu, M. Menasche, "An enumerative approach for analyzing Time Petri Nets," Information Processing: Proc. IFIP Congress 1983, R.E.A. Mason, ed., vol. 9, pp. 41-46, 1983.

[5] E. Vicario, "Static analysis and dynamic steering of time dependent systems using Time Petri Nets," *IEEE Trans. Soft. Eng.*, vol. 27, no. 8, pp. 728-748, Aug. 2001.

[6] R. Alur, D.L. Dill, "A theory of timed automata", *Theoretical Computer Science*, Vol. 126, pp. 183-235, 1994.

[7] B. Berthomieu, P.-O. Ribet, and F. Vernadat, "The Tool TINA—Construction of abstract state spaces for Petri Nets and Time Petri Nets," *Int. J. Production Research*, Vol. 42, No. 14, 2004.

[8] G. Bucci, L. Carnevali, L. Ridi, E. Vicario, "ORIS: a tool for modeling, verification and evaluation of real-time systems", *Int. J. on Software Tools for Technology Transfer*, Springer (2010) 12:391–403, DOI 10.1007/s10009-010-0156-8.

[9] UPPAAL on-line, www.UPPAAL.org.

[10] G. Behrmann, A. David, K.G. Larsen, "A tutorial on UPPAAL", In: *Formal Methods for the Design of Real-Time Systems*, M. Bernardo and F. Corradini Eds., Lecture Notes in Computer Science, Vol. 3185, Springer-Verlag, pp. 200-236, 2004.

[11] A. David, K.G. Larsen, A. Legay, M. Mikucionis, D.B. Poulsen, UPPAAL SMS Tutorial, *Int. J. on Software Tools for Technology Transfer*, Springer, **17**:1-19, 06.01.2015, DOI 10.1007/s10009-014-0361-y, 2015.

[12] M.Z. Kwiatkowska, G. Norman, D. Parker, "PRISM 4.0: Verification of Probabilistic Real-Time Systems". In *Proc. of CAV 2011*, pp. 585-591, 2011.

[13] H.L.S. Younes, "*Verification and planning for stochastic processes with asynchronous events*", PhD Thesis, Carneige Mellon, 2005.

[14] E. Vicario, L. Sassoli, L. Carnevali, "Using stochastic state classes in quantitative evaluation of dense-time reactive systems", *IEEE Trans. on Soft. Eng.*, Vol 35, No. 5, pp. 703-719, 2009.

[15] M.A. Marsan, G. Balbo, G. Conte, S. Donatelli, G. Franceschinis, *Modelling with Generalized Stochastic Petri Nets*, John Wiley and Sons, 2004.

[16] F. Cicirelli, A. Furfaro, L. Nigro, "Model checking time-dependent system specifications using time stream Petri nets and UPPAAL", *Appl. Math. Comp.*, Vol. 218, pp. 8160-8186, 2012.

[17] L. Carnevali, M. Paolieri, K. Tadano, E. Vicario, "Towards the quantitative evaluation of phased maintenance procedures using non-Markovian regenerative analysis, *Lecture Notes in Computer Science*, Springer, Vol. 8168, pp 176-190, 2013.

[18] G. Bucci, A. Fedeli, L. Sassoli, and E. Vicario, "Timed state space analysis of real time preemptive systems," *IEEE Trans. Soft. Eng.*, vol. 30, no. 2, pp. 97-111, Feb. 2004.

[19] F. Cicirelli, A. Furfaro, L. Nigro, F. Pupo, "Development of a schedulability analysis framework based on pTPN and UPPAAL with stopwatches". In *Proc. of the 16th IEEE/ACM Int. Symp. on Distributed Simulation and Real Time Applications (DS-RT)*, pp. 57-64, 2012.

[20] L. Carullo, A. Furfaro, L. Nigro, F. Pupo. "Modelling and Simulation of Complex Systems using TPN Designer". *Simulation Modelling Practice and Theory*. **11**/7-8, pp. 503-532, 2003.

DOI: 10.15439/2015F32
ACSIS, Vol. 5

# Modelling and Verification of Starvation-Free Mutual Exclusion Algorithms based on Weak Semaphores

Franco Cicirelli, Libero Nigro
Laboratorio di Ingegneria del Software
Dipartimento di Ingegneria Informatica Modellistica Elettronica e Sistemistica
Università della Calabria, Italy
{f.cicirelli@dimes.unical.it, l.nigro@unical.it}.

*Abstract—* **This paper proposes an original framework for modelling and verification (M&V) of starvation-free mutual exclusion algorithms based on weak semaphores, that are without a built-in waiting-process queue. The goal is to support the implementation of light-weight starvation-free semaphores useful in general concurrent systems including cyber physical systems. The M&V approach depends on UPPAAL. First known weak semaphores are modelled. Then they are exploited for model checking classic algorithms. Known properties are retrieved but subtle new ones are discovered. As part of the developed approach, a new algorithm is proposed which uses two semaphores of the weakest type, N bits (N being the number of processes) and a counter. This algorithm too is proved to be correct.**

## I. INTRODUCTION

MUTUAL exclusion is the well-known problem of synchronizing a group of concurrent processes (or threads) sharing some data variables, so as to avoid interferences on shared data. The problem is to ensure that processes can enter their critical section (i.e., a block of instructions accessing/modifying the shared data) one at a time. To be acceptable, though, a mutual exclusion algorithm should be also starvation-free, that is a process waiting to enter its own critical section should experiment a bounded waiting time. This in turn favors process fairness.

Commonly, mutual exclusion can be based on semaphores or on monitor locks. This paper focuses on starvation-free mutual exclusion algorithms based on *weak semaphores*, i.e., semaphores without an in-built process waiting queue ensuring a first-in-first-out awakening policy.

The results of this paper can be exploited for implementing light-weight starvation-free semaphores, useful in general concurrent applications and cyber physical systems (e.g., [1]), and also in distributed shared memory systems where it is challenging to build a classical queue-based semaphore when the processes belong to distinct physical computation nodes (address spaces).

The goal is to propose an original approach based on Timed Automata (TA) in the context of the UPPAAL toolbox [2], for modelling and verification through model checking [3] of any mutual exclusion algorithm designed on top of weak semaphores. The goal is similar to that described in [4] where an approach based on the use of the PVS theorem prover was developed. This paper argues that the use of a toolbox like UPPAAL can be preferable as a proving framework because it avoids the mathematical formalization necessary to specify and check properties of an algorithm. In addition the approach permits to disclose subtle aspects of a modelled algorithm, e.g., related to timing, which are normally out of reach of a theorem prover.

The modelling and verification (M&V) approach is applied to known classic algorithms, e.g. [5]-[8], of which are confirmed known properties. Nevertheless, some new properties (e.g., the existence of a zeno-cycle and of a time-sensitive behavior which affects the kind of the weak semaphores which can be used) are disclosed which were not previously documented in the literature. As a part of the accomplished work a novel algorithm based on the Morris one [5] is proposed which rests only on two weak semaphores, one counter and N bits. This algorithm too is proved to be starvation-free.

The developed proving framework can provide some new arguments on the E. Dijkstra conjecture [9] about the impossibility to build a starvation-free semaphore using only weak semaphores.

The paper is structured as follows. First an overview of the UPPAAL M&V concepts is furnished. Then the three known types of weak semaphores are introduced and modelled into UPPAAL. After that, a common vision [4] of classic starvation-free mutual exclusion algorithms is discussed. Then the developed M&V approach is applied to classic algorithms as well as to a new one proposed in this paper. A comparison of the algorithms properties is finally presented. Some indications about on-going and future work are given in the conclusions.

## II. UPPAAL CONCEPTS

UPPAAL [2] is a popular and efficient toolbox based on Timed Automata (TA) [10] suited for modelling and verification of real-time systems. A timed automaton is a finite automaton augmented with a set $C$ of real-valued variables named *clocks*. Clocks model the time elapsing and are assumed to grow synchronously at the same pace of the hidden system time. Constraints, of the form $x \sim k$ or

This work was not supported by any organization

978-83-60810-66-8/$25.00©2015, IEEE

$x - y \sim k$ where $x$ and $y$ are clocks, $k$ is a non-negative integer and $\sim \in \{\leq, <, =, >, \geq\}$, are called *clock constraints* and can be introduced to restrict the behavior of the automaton. A set of clock constraints used to label an edge it is called a *guard* and determines the possible values which can be assumed by the involved clocks for the corresponding state transition to be allowed. Clock constraints of the type $x \sim k$ can also be used to label locations and are called *invariants*. An automaton can stay in a location as long as the clocks satisfy the location invariant attached to the location. Additionally, edges can be labeled by a set of clocks, which are reset as the corresponding transition is taken, and by an *action* label.

TA can be composed to form a network of concurrent TA whose semantics is based on interleaving of actions as well as hand-shake synchronizations. UPPAAL adopts the notion of a *channel* for input and output action synchronization and uses a CSP-like notation. The edge of automaton labeled with `ch!` (output action), where `ch` is a channel, matches with an edge of another automaton labeled with `ch?` (input action). At a given time it may exist more than one pair of enabled and matched edges in which case a choice is made non-deterministically. Taking a transition (edge) in an automaton denotes an *atomic action* in the TA concurrent model. Moreover, the update of a sender is executed *before* that of a receiver.

The UPPAAL model-checker generates *on-the-fly* the state graph of a network of TA (see, e.g., [11]) for checking formulas as in the following:

- $E <> \emptyset$ (Possibly $\emptyset$, i.e., a state exists where $\emptyset$ holds)
- $A[]\ \emptyset$ (Invariantly $\emptyset$, equivalent to: $not\ E <> not\ \emptyset$)
- $E[]\ \emptyset$ (Potentially Always $\emptyset$, i.e. a state path exists over which $\emptyset$ always holds)
- $A <> \emptyset$ (Always eventually $\emptyset$, equivalent to: $not\ E[]\ not\ \emptyset$)
- $\emptyset --> \psi$ ($\emptyset$ always *leads-to* $\psi$, equivalent to: $A[]\ (\emptyset\ imply\ A <> \psi)$)

where $\emptyset$ and $\psi$ are state properties, e.g., clock constraints or boolean expressions over predicates on locations.

In addition to the support for classic TA, UPPAAL provides integer variables with a bounded set of values, arrays and structs, and a notion of automata *templates* which can be instantiated multiple times by specifying different values for their parameters.

Locations can also be labeled as being *committed* (C) or *urgent* (U) both of which must be abandoned with no time passing. Committed locations have precedence over urgent locations. UPPAAL provides also broadcast synchronizations. Channels can be declared to be urgent. An enabled synchronization on a urgent channel is required to occur without time passage.

Finally, it is worth mentioning the possibility of building a counterexample (i.e., diagnostic trace) of a not satisfied property, which can be analyzed in the simulator of the toolbox.

## III. MODELLING WEAK SEMAPHORES

A semaphore is an abstract object which hides an integer variable which can only be modified by the two atomic operations `P` and `V`. Fig. 1 shows an UPPAAL model of a basic *plain* binary semaphore, whose value can be `0` or `1`. The `P/V` operations are modelled through a matrix of channels. The first index `s` specifies the semaphore id. The second one carries the id of the requesting process. The model in Fig. 1 makes a non-deterministic selection of the requesting process at a `P` or `V` synchronization. A not chosen process rests blocked on the requesting `P!` or `V!` operation.

Of course the model in Fig. 1 is unfair: at the time of a `V`, if more processes are blocked on the `P!` synchronization on the same semaphore, one of them is chosen not-deterministically to proceed. It is also possible for the `V`-executor process to compete and reacquire immediately the semaphore. Therefore, every process can experiment an unbounded waiting. It is known that to turn an unfair semaphore into a fair one a queue can be added to the semaphore where processes which find the semaphore red (0) at the time of a `P` are stored in first-in-first-out order and then awaken in the same order at a subsequent `V`.

A challenging research goal in the literature is to achieve a fair semaphore using only a minimum number of unfair semaphores. In the case when the queue is replaced by a set, the semaphore becomes unfair and it is often referred to as a *buffered* [4] or *blocked-set* semaphore [12]. A third version of weak semaphore is the so called *polite* semaphore proposed in [4]. A *polite* semaphore is similar to a *plain* semaphore but forbids the process which executes a `V` operation to reacquire immediately the semaphore.

The three types of weak semaphores can be modelled in UPPAAL as shown in the Fig. from 2 to 4 where for generality a `P` operation is supposed to be immediately followed by a `GO` synchronization to unblock the `P`-requester process. A `V` operation, being not blocking, never requires to be followed by a `GO` synchronization.

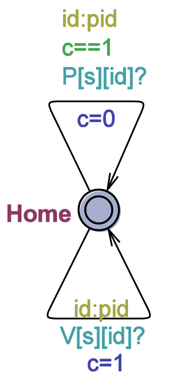


Fig. 1. A plain binary semaphore automaton

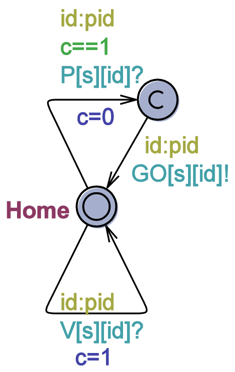


Fig. 2. Adopted `PlainBinary-Semaphore` automaton

Models in the Fig. 2 to 4 represent formal definitions of basic weak semaphores. Their correctness can be checked as follows. In the *polite* model in Fig. 3, the local variable `last` stores the id of the `V`-executor. The default value of `last` is `NONE`. At the time of a `P`, if the semaphore is green (`1`) and the process id is different from `last`, atomically the semaphore is turned to red (`0`) and the process (held in the `this` local variable) immediately receives the `GO` signal. Would the semaphore be red, or its id coincides with the `last` value, the process partially executes the `P` operation by incrementing the (local) counter `cnt` of blocked processes. A blocked process can be awaken by a `GO` synchronization raised in the right edge of Fig. 3, which completes its `P` operation by turning red the semaphore, decrementing the `cnt` variable and by assigning `NONE` to `last`. A subsequent `V` operation can (non-deterministically, as for a *plain* semaphore) unblock a waiting process, provided its id is different from `last`, or permit to a newly arrived process different from `last` to acquire the semaphore through a `P` operation.

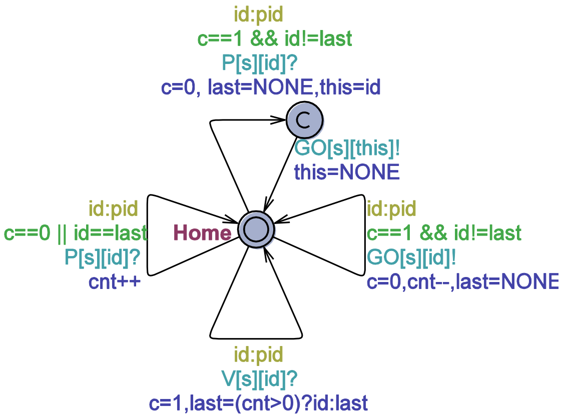


Fig. 3. `PoliteBinarySemaphore` automaton

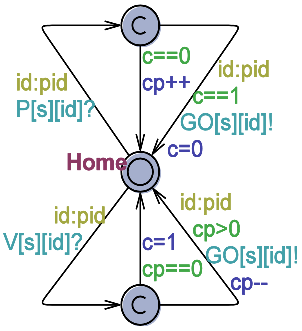


Fig. 4 `BufferedBinarySemaphore` automaton

If `s` is a *polite* binary semaphore, the following invariants hold:

1) `A[] s.last!=NONE imply s.cnt>0`
2) `A[] s.last==NONE || s.cnt>0`

Due to the invariant 1) it was omitted in the right edge with a `GO` synchronization of Fig. 3 the test `cnt>0` in the edge guard.

In the *buffered* semaphore model in Fig. 4 the buffer was purposely achieved implicitly in UPPAAL by the waiting locations of the set of processes which have just executed the `P!` operation and found red the semaphore. The number of such blocked processes is stored in the variable cp. At the time of a `V!`, if there are blocked processes, the semaphore is not turned to 1 and one of such processes is chosen non deterministically and receives the synchronization on the corresponding `GO` channel. It should be noted that both *polite* and *buffered* models exclude the `V`-executor to reacquire immediately the semaphore. But the *buffered* semaphore is stronger than the *polite* because at the time of a `V` operation, only one already blocked process can receive the semaphore pass. Fig. 5 summarizes global declarations of an UPPAAL model which makes use of any semaphore model in the Fig. from 2 to 4.

```
const int N=…;//number of processes
const int SEM=…;//number of weak semaphores
const int NONE=-1;

typedef int[0,N-1] pid;
typedef int[0,SEM-1] sid;

//semaphore IDs
…
//semaphore channels
urgent chan P[sid][pid];
urgent chan GO[sid][pid];
urgent chan V[sid][pid];
```

Fig. 5 Common global declarations for weak semaphores

Since a location without a clock invariant can disrupt liveness of an UPPAAL model being possible to remain in the location an arbitrary (potentially infinite) time, semaphore channels were declared as urgent. This way, without hurting model non-determinism, when a given operation is enabled it will be allowed to occur without time passage. This measure was adopted as a way to realize in UPPAAL the *finite delay property* [12] or *weak fairness* [4] of processes in a concurrent model, which requires a continuously enabled action eventually happens.

As a final remark, all the models in the Fig. from 2 to 4 can be implemented in a concurrent programming language (e.g., Java) using busy-waiting.

## IV. Model Checking Starvation-free Mutual Exclusion Algorithms based on Weak Semaphores

If `S` is a fair binary semaphore initialized to 1, the usual pattern for achieving mutual exclusion among `N` (>2) competing processes accessing some shared data is the following:

**process**(p)=**loop** NCS; P(S); CS; V(S); **endloop**.

A problem which has received the attention of many researchers consists in the possibility of building a sound fair semaphore using only a minimal number of weak semaphores. Starting from late seventies some starvation-free mutual exclusion algorithms were proposed, though without an adequate proof of their correctness. Recently, in [4] a very interesting proof system based on the PVS theorem prover was defined and used to establish the correctness of three fundamental algorithms proposed in [5]-[7].

This paper claims that the approach and the results described in [4] are still unsatisfactory and that some properties exist to be discovered about those and other algorithms. A fundamental step in [4] was the identification of an abstract algorithm which facilitates the interpretation and analysis of the three mentioned algorithms.

The abstract algorithm is founded on the elevator metaphor: "While there are interested processes they enter the elevator at the first floor. When there are no processes arriving anymore, the elevator goes to the second floor and lets its occupants into CS, one by one. When the elevator is empty, it goes down again. When the elevator is not at the first floor, arriving processes have to wait. After CS, the processes go down by stairs."

The abstract algorithm uses 4 integer variables: ne, nm, se and sm. The first two variables model respectively the number of processes at the first floor waiting for the elevator, and the number of occupants within the elevator. The last two variables model respectively the doors at the first and second floor. Initially all variables are set to 0 except for the se which is set to 1. The abstract algorithm is reproduced in Fig. 6 where atomic actions are enclosed within <…>.

```
process(p)=
loop
      NCS;
      <ne++>
      <await se greater-than 0; nm++; ne--;
         if ne==0 then sm++; se--; endif;>
      <await sm greater-than 0; sm--; nm--;>
      CS;
      <if nm greater-than 0 then sm++; else se++; endif;>
endloop.
```

Fig. 6 Mutual exclusion abstract algorithm

Correctness of the abstract algorithm can be verified in UPPAAL by deriving a corresponding native model like that shown in Fig. 7. Such a model only depends on the concurrency model of UPPAAL and in particular on the atomic actions labelling the various edges. The model consists of two TA: the Process(const pid p) (see Fig. 7) and the Synch(ronizer (see Fig. 8). The Process automaton embodies the mutex algorithm and is instantiated N (e.g., N=4) times.

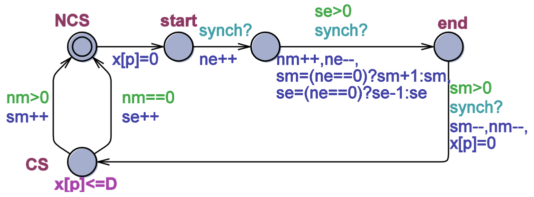


Fig. 7 UPPAAL Process automaton

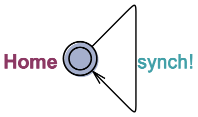


Fig. 8 The Synch automaton

All these instances share the algorithm variables, declared globally. Only one instance instead exists for the synchronizer. Each process instance p uses a clock x[p] to measure the waiting time before entering the critical section and the duration of the critical section.

The synchronizer is always ready to send a signal over the urgent unicast channel synch. Such a signal is a key to ensure progress to the model in Fig. 7 where some normal locations without clock invariants like start and end, are used.

Also the NCS location is without clock invariant, to mirror the fact that the non-critical section lasts an arbitrary number (also 0) of time units.

The entry protocol of the mutex algorithm is played from the start to the end location. The exit protocol is coded on the arcs outgoing the CS location.

The following queries were used for property checking of the abstract algorithm.

```
1)A[] !deadlock                          satisfied

2)A[] forall(i:pid) forall(j:pid)
  Process(i).CS && Process(j).CS
  imply j==i                             satisfied

3)Process(0).start --> Process(0).CS
                                     not satisfied

4)A[] forall(i:pid) Process(i).end
  imply x[i]<=2*(N-1)*D                  satisfied

5)A[] forall(i:pid) Process(i).end
  imply x[i]<=2*(N-1)*D-1            not satisfied
```

Queries 1) and 2) check *safety properties*. Query 3) verifies a *liveness property*. Queries 4) and 5) check a *bounded liveness property*.

Satisfaction of query 1) guarantees the model has no deadlock (predefined keyword in UPPAAL). Query 2) ensures only one process at a time can be in the critical section. Query 3) checks if any process which finds itself in the start location eventually reaches the CS (critical section) location. Noteworthy, this property is not satisfied. Queries 4) and 5) check about the waiting time of each process before entering CS (whose duration is supposed to be at most D time units).

Note that clock x[p] is reset on entering start and on exiting end. It is confirmed that every process p has an *overtaking factor* of 2, i.e., its blocking time is determined by all the other processes which enter two times their CS before p can enter its CS.

Absence of liveness (query 3) is a direct consequence of the fact that the model has a *zeno-cycle*, i.e., it is possible for any process to (re)enter the CS an infinite number of times before any other process can enter its CS, by consuming 0 time. The zeno-cycle mirrors the fact that the critical section as well as the non-critical section can have a 0 duration, and that nothing forbids (non-determinism) the same process to always get the synch signals.

The zeno-cycle can be eliminated by guaranteeing the critical section necessarily consumes a finite (although very small) duration (the guard x[p]>0 can be added to both edges exiting from the CS location). However, the existence of the zeno-cycle does not prevent the model checker to determine the worst-case waiting time of processes, in which case UPPAAL considers scenarios (behaviors) on the state graph where time is really advancing.

It should be noted that the presence of a zeno-cycle naturally expresses an intrinsic feature of the algorithm/model design. A different design can be without any zeno-cycle, independently from any consideration about timing.

The following invariants also hold for the model of Fig. 7:

```
A[] se==1 imply sm==0,
A[] sm==1 imply se==0,
```

which express functionality concerns of the abstract algorithm.

In [4] it is shown as some classic starvation-free mutual exclusion algorithms based on weak semaphores can be regarded as different interpretations of the abstract algorithm, where atomic operations are achieved by using a few unfair binary semaphores.

For brevity, in the following only the Morris algorithm [5] will be detailed. For the other studied algorithms, though, the experimental analysis will be synthetically reported.

### *A. Morris Algorithm*

Fig. 9 recapitulates the Morris algorithm which can be viewed as a concrete instance of the abstract algorithm of Fig. 6. The Morris algorithm uses three weak semaphores: sb, protecting specifically the ne counter holding the number of processes awaiting the elevator at the first floor, se and sm respectively controlling the door at the first and the second floor. They act as a split binary semaphore. Initially, sb and se are set to 1, sm to 0.

The initial values of the other variables are as in the abstract algorithm.

```
process(p)=
  loop
    NCS;
    P(sb); ne++; V(sb);
    P(se); nm++; P(sb); ne--;
    if ne>0 then V(sb); V(se);
    else V(sb); V(sm); endif;
    P(sm); nm--; CS;
    if nm>0 then V(sm); else V(se); endif;
  endloop.
```

Fig. 9 The Morris mutual exclusion algorithm

In Fig. 10 it is depicted an UPPAAL Process automaton corresponding to the algorithm in Fig. 9.

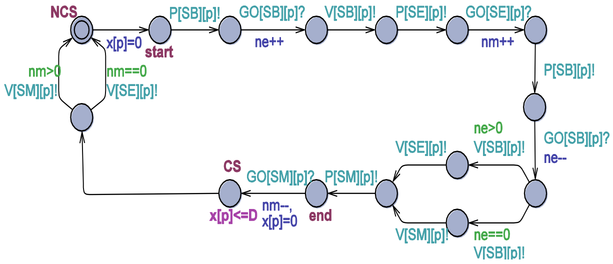


Fig. 10 UPPAAL Process model corresponding to the Morris algorithm of Fig. 9

In this case, the use of urgent semaphore channels avoids the recourse to other channels like synch of Fig. 8 in order to guarantee model progress. The model in Fig. 10 was model checked using the same queries 1) to 5) previously discussed for the abstract algorithm, and using for the se and sm semaphores the PlainBinarySemaphore template (Fig. 2) and separately checking the model behavior when the sb semaphore is implemented respectively as a *plain*, *polite*, or *buffered* binary semaphore, i.e., passing from the weakest to the strongest unfair semaphore.

In [4] a proof system was built to demonstrate that the Morris algorithm is correct, i.e., it is without deadlock, it ensures mutual exclusion and it guarantees a bounded waiting time (with an overtaking factor of 2) for the blocked processes, for the sole case sb-*buffered* semaphore, se, sm-*plain* semaphores. From our analysis based on model checking it emerged that the Morris algorithm, even with sb being a *buffered* semaphore, always has a zeno-cycle which means, under the hypothesis of zero time duration of any action in the algorithm, that the overtaking factor for a blocked process is unbounded. Only when the critical section is supposed to consume even a very small time duration, the zeno-cycle disappears. Moreover, in the presence of timing of the critical section, the overtaking factor is effectively 2 as for the abstract algorithm but for any implementation of the sb semaphore. In other words, model checking the Morris algorithm, in the presence of timing, confirmed that the algorithm is correct with three *plain* binary semaphores, contrary to what is stated in [4] and [12].

The study of the Morris algorithm suggested to us the design of a simple variation of the algorithm based on two *plain* semaphores (the se and sm semaphores of the Morris algorithm), N bits and the nm counter. The algorithm is proposed in Fig. 11 and modelled in UPPAAL as in Fig. 12. It avoids the sb semaphore and uses instead an array e of N booleans, each element being associated to a distinct process.

```
process(p)=
  loop
    NCS;
    e[p]=true;
    P(se); nm++; e[p]=false;
    if ne() then V(se);
    else V(sm); endif;
    P(sm); nm--; CS;
    if nm>0 then V(sm); else V(se); endif;
  endloop.
```

Fig. 11 Proposed variation of the Morris algorithm

The array e replaces the ne counter of the abstract algorithm. Each process p sets e[p] to true when it starts waiting for the elevator at the first floor, and resets it to false when it enters the elevator, at which time the nm counter is incremented. Since each process manages its own element in the array e, no interference can ever occur on e. The test about the existence of other processes which want to enter the elevator at the first floor, previously based on the counter ne, it is now based on checking if there are some true elements in the array e (the check is actually delegated to a function ne() which returns true if some element in the array is true, false otherwise). Of course, a true value in e can be found in the current test or it will be sensed the next time.

Model checking the model in Fig. 12 confirmed that all the five queries proposed for the abstract algorithm are now satisfied. Also the liveness property (query 3) now holds, i.e., the new algorithm is without any zeno-cycle, and correctly operates even when timing is ignored.

### B. Algorithms Comparison

During the development of the modelling and verification approach described in this paper, besides the Morris algorithm, other starvation-free mutual exclusion algorithms based on weak semaphores were studied. Model checking results summarized in the Table 1 confirm known results in the literature and in some cases are more detailed. In the column of the semaphore types, the weakest admissible types for the algorithm are shown. More stronger versions could, but unnecessarily, be used. For instance, the sb (as in the Morris algorithm) semaphore of the Udding algorithm must be *buffered*. The other two semaphores can be *plain*. The Udding algorithm is no longer starvation-free if sb is implemented with a *polite* semaphore.

Analysis results concerning the Martin & Burch algorithm [6] coincide with those formally identified in [4]. The Haldar & Subramanian algorithm which relies on two semaphores and 2 bits [8] was also investigated in [13]. The weak semaphore type the authors assumed corresponds to a *buffered* one and the overtaking factor was indicated as

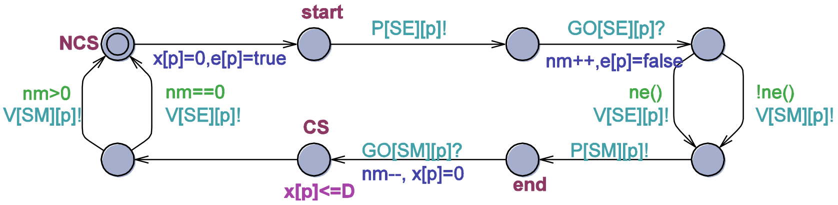


Fig. 12 Variation of the Morris algorithm based on two plain semaphores, N bits and the sole nm counter

TABLE I.
MODEL CHECKING RESULTS OF MUTUAL EXCLUSION ALGORITHMS.

| Algorithm | No of weak semaphores | Semaphore types | Zeno-cycle | Overtaking factor |
|---|---|---|---|---|
| Morris | 3 | 3 plain | yes | 2 |
| Morris variation proposed in this paper | 2 | 2 plain | no | 2 |
| Udding | 3 | 1 buffered – 2 plain | yes | 2 |
| Martin & Burch | 2 | 1 polite – 1 plain | yes | 2 |
| Haldar & Subramanian | 2 | 2 polite | yes | lesser than 2 |

being 2. However, the model checking approach developed in this work has shown that two *polite* semaphores suffice and that the waiting time of a process interested in entering its critical section is exactly 2*(N-1)*D-D, i.e., one critical section lesser than the other algorithms.

As it emerges from Table 1, the Morris variation algorithm proposed in this paper outperforms classic known algorithms. With respect to the Haldar & Subramanian algorithm, our algorithm uses only 2 binary semaphores of the weakest type (*plain*) although it uses some more memory (N bits plus the nm counter vs. 2 bits of the Haldar & Subramanian algorithm). Moreover, the proposed algorithm is the only one which is without zeno-cycles.

## V. Conclusions

The Dijkstra conjecture [9] about the impossibility of building a fair semaphore using a few weak semaphores was confuted by the development of algorithms proposed by Morris [5], Martin & Burch [6], Udding [7] etc. However the correctness proof of such algorithms was only partially provided, also considering the methodological approach proposed in [12] or the proof framework developed in [4] which does not allow a full analysis of mutual exclusion algorithms in the presence of the timing dimension.

In this paper an original proving framework based on timed automata (TA) and the Uppaal toolbox is proposed which permits modelling and full verification of the properties of starvation-free mutual exclusion algorithms based on weak semaphores, also in the presence of the timing dimension.

The approach models the three known types of weak semaphores: *plain* (the weakest type), *polite* and *buffered* (the strongest type). It is worthy of note that in its description, the Dijkstra conjecture implicitly refers to the use of *buffered* semaphores.

A key factor of the proposed approach is its modelling and analysis flexibility, being it possible to transparently replace a semaphore type with another one thus enabling a thorough study of a given algorithm.

The application of the approach confirms known properties of classic algorithms, but has the potential to discover subtle features of the considered algorithms such as the existence of a zeno-cycle or of a time-sensitive behavior which influences the kind of weak semaphores which can be actually used. All known algorithms suffer of a zeno-cycle, in the light of which the overtaking factor of a waiting process is (theorically) unbounded. However, when the critical section consumes an even infinitesimal time, the bounded waiting time and overtaking factor of the classic algorithms is effectively guaranteed. In this hypothesis the Morris algorithm is correct with three *plain* semaphores.

As part of this work, a variation of the Morris algorithm was designed which intrinsically eliminates any zeno-cycle, rests only on two *plain* semaphores and replaces a counter of the Morris algorithm with N bits. This new algorithm too was proved to be correct.

The paper contribution enables the implementation of light-weight starvation-free semaphores which can be exploited in general concurrent systems including cyber physical systems.

Prosecution of the research is geared at:

- Modelling and analysis of other mutual exclusion algorithms designed in terms of weak semaphores.
- Implementing weak semaphores and fair semaphores corresponding to mutual exclusion algorithms, in a concurrent programming language, e.g., Java.
- Experimenting with the use of weak semaphores in practical systems programming and in the development of cyber physical systems.
- Exploiting light-weight starvation-free semaphores in distributed shared memory systems, e.g., based on Java and the Terracotta middleware [14] which provides the vision of a "network heap" where shared data can be accessed by threads belonging to distributed JVMs.

## References

[1] S. Srbljic, D. Skvorc, M. Popovic, "Programming languages for end-user personalization of Cyber-Physical Systems", *Automatika*, Vol. 53, No. 3, pp. 294-310, 2012.

[2] G. Behrmann, A. David, K.G. Larsen, "A tutorial on Uppaal", In: *Formal Methods for the Design of Real-Time Systems*, M. Bernardo and F. Corradini Eds., Lecture Notes in Computer Science, Vol. 3185, Springer-Verlag, pp. 200-236, 2004.

[3] E.M. Clarke, O. Grumberg, D.A. Peled, *Model checking*, MIT Press, 2000.

[4] W.H. Hesselink, M. IJbema, M. "Starvation-free mutual exclusion with semaphores", *Formal Aspects of Computing*, DOI 10.1007/s00165-011-0219-y, 2011.

[5] J.M. Morris, "A starvation-free solution to the mutual exclusion problem", *Inf. Proc. Lett.*, Vol. 8, pp. 76-80, 1979.

[6] A.J. Martin, J.R. Burch, "Fair mutual exclusion with unfair P and V operations", *Inf. Proc. Lett.*, Vol. 21, pp. 97-100, 1985.

[7] J.T. Udding, "Absence of individual starvation using weak semaphores", *Inf. Proc. Lett.*, Vol. 23, pp. 159-162, 1986.

[8] S. Haldar, D.K. Subramanian, "An efficient solution of the mutual exclusion problem using unfair and weak semaphores", *ACM SIGOPS Operating Systems Review*, Vol. 22, pp. 60-66, 1988.

[9] E.W. Dijkstra, "A strong P/V-implementation of conditional critical regions", Tech. Rep., Tech. Univ. Eindhoven, EWD 651, www.cs.utexas.edu/users/EWD, 1977.

[10] R. Alur, D.L. Dill, "A theory of timed automata", *Theoretical Computer Science*, Vol. 126, pp. 183-235, 1994.

[11] F. Cicirelli, A. Furfaro, L. Nigro, "Model checking time-dependent system specifications using time stream Petri nets and Uppaal", *Appl. Math. Comp.*, Vol. 218, pp. 8160-8186, 2012.

[12] E.W. Stark, "Semaphore primitives and starvation-free mutual exclusion", *J. of ACM*, Vol. 29, pp. 1049-1072, 1982.

[13] H.P. Hofstee, K.R. Leino, L.A. van de Snepscheut, "Proof of mutual exclusion algorithm by Haldar and Subramanian", HPH11-0, California Institute of Technology, 18 December 1991.

[14] F. Cicirelli, A. Furfaro, A. Giordano, L. Nigro, "Performance of a multi-agent system over a multi-core cluster managed by Terracotta", In *Proc. of Symp. on Theory of Modeling & Simulation: DEVS Integrative M&S Symp.*, pp. 125-133, 2011.

DOI: 10.15439/2015F46
ACSIS, Vol. 5

# Using Domain Specific Languages to Improve the Development of a Power Control Unit

Mathijs Schuts
Philips HealthTech,
Best,
The Netherlands
Email: mathijs.schuts@philips.com

Jozef Hooman
Radboud University & TNO,
Nijmegen & Eindhoven,
The Netherlands
Email: jozef.hooman@tno.nl

***Abstract*—To improve the design of a power control unit at Philips, two Domain Specific Languages (DSLs) have been used. The first DSL provides a concise and readable notation for the essential state transitions. It is used to generate both configuration files and analysis models. In addition, we also generate instances of a second DSL which represents test traces. This second DSL is used to generate test cases for the power control unit. The use of DSLs not only improved productivity, but also the quality of the configuration files and the test set.**

## I. Introduction

THIS PAPER discusses industrial experience with the use of Domain Specific Languages (DSLs) at Philips HealthTech. We have used DSLs to develop power control units of systems for image guide therapy. These systems are used during minimally invasive medical treatments, such as the treatment of cardiology and vascular diseases. An example is the interventional X-ray system shown in Fig. 1, where X-ray images support minimally-invasive medical procedures such as placing a stent via a catheter.

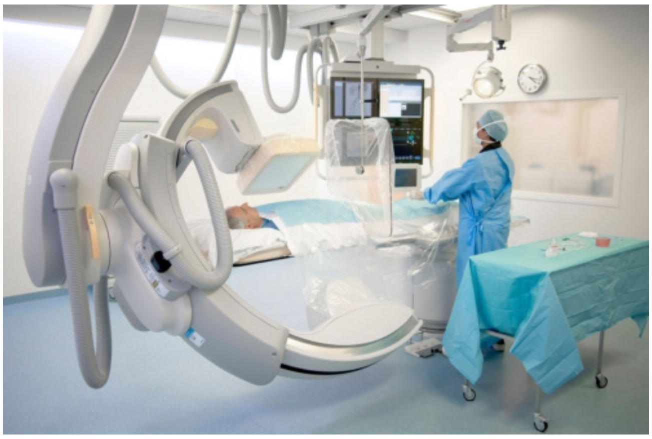


Fig. 1. Interventional X-ray System

Given the long history of these systems and the frequent need for changes to support new medial procedures, it is important to keep the software architecture and its components flexible and extensible. Hence, legacy components have to be renewed to keep them prepared for the future.

An example of such a legacy component is the power control unit. It uses low-level configuration files which are hard to read and difficult to maintain. Given the increasing number of configurations and new third-party components that have to be supported, this starts to become a potential bottleneck. In addition, only a limited set of regression tests is available.

Already long ago, DSLs have been suggested as a way to raise the level of abstraction, to deal with variability, and to improve productivity and maintainability. An early overview of terminology, techniques and applications can be found in [1]. As one of the disadvantages of DSLs this paper mentions the costs of designing, implementing and maintaining a DSL. Since then, however, large improvements have been achieved in the area of language workbenches. Such tools facilitate the efficient construction of languages, editors, and transformations [2], [3]. Examples of workbenches are MetaEdit+ [4], Rascal [5], Spoofax [6], EMFText [7], and Xtext [8].

There are a number of relevant applications in the domain of embedded systems. For instance, there is an interesting laboratory experiment of the application of MetaEdit+ to heart rate monitors of Polar [9], showing a large increase in productivity. Xtext has been used to define a DSL which models the hardware configuration of the complex lithography machines of ASML [10]. From this DSL, a simulation of hardware behaviour which enables software in the loop simulation has been generated. In [11], a DSL based on Xtext has been developed to generate code for real-time large-scale distributed data processing. By means of the MPS approach [12], an impressive extension of the C language has been constructed [13], [14].-

The aim of our work is to investigate whether DSLs could provide a solution to improve the maintainability and testability of our power control unit. We would like to get an answer to the following questions:

- How much time is needed to learn the tools and techniques?
- How much effort is needed to migrate the current legacy component to a component which is defined by a high-level human-usable DSL?
- Does the DSL approach support the combination with analysis techniques such as simulation tools and formal

This research was supported by the Dutch national program COMMIT and carried out as part of the Allegio project.

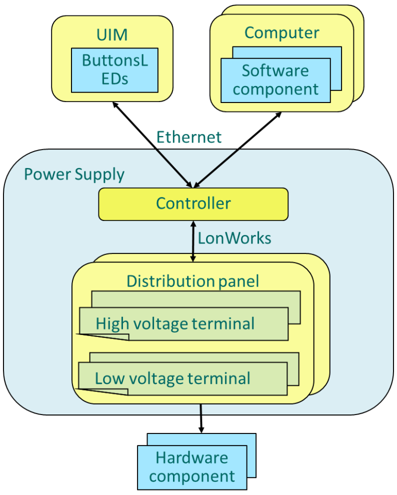


Fig. 2. Overview Power Control Unit

model checkers?

- What are the benefits of introducing these new techniques compared to the current way of working?

The paper is organized as follows. Sect. II describes our industrial case. The developed DSLs are presented in Sect. III and Sect. IV. Sect. V contains an overview of the results and an evaluation of the approach.

## II. Case: Power Control Unit

The interventional X-ray systems of Philips use a distributed architecture with a large number of hardware and software components. The system is highly configurable, i.e., customers can select a particular combination of X-ray stands, monitors, image processing capabilities, etc. Powering the hardware components, and starting up and shutting down the software components are the responsibility of the power control unit. This unit is installed in a technical room together with a number of cabinets which contain the required hardware components. The power control unit consists of a controller that has three interfaces, as shown in Fig. 2:

- An interface to a User Interface Module (UIM) that has On and Off buttons, and two LEDs for user feedback.
- An interface with software components running on computers.
- An interface with power distribution panels that are placed in the cabinets to power the hardware components installed within the same cabinet. Each power distribution panel has a number of individually switchable high and low voltage terminals that are managed by the controller.

The controller and all distribution panels have a 16 bit microcontroller running an embedded application. They communicate with each other via LonWorks [15] using a master-slave topology. LonWorks creates a communication channel superimposed on the power line with which the controller powers the distribution panels. The controller implements the system start-up/shut-down behaviour using a state machine with two parts:

- A high-level state machine that is part of the application running inside the controller.
- A configuration file that describes the low-level state behaviour of the power control unit.

The configuration file is used by the high-level state machine to perform the configuration-specific transitions. This high-level state machine is implemented with VisualState [16] and describes the main states and the associated LED behaviour when transitioning between these main states. These main states are:

- Off: the power control unit is not powered;
- Init: represents the start-up of the power control unit, in this state a Power On Self Test (POST) is executed;
- Standby: the system is off for the user, but power control unit is standby and some continuous power terminals are powered;
- Operational: the system is on for the user, typically all terminals provide power in this state;
- Emergency Power Off (EPO): the controller cuts off the power of the distribution panels immediately and thereby also all the terminals loose power (only the controller stays powered) - used when the user presses a red safety button;
- Stop: a terminal state which is entered when critical parts of the power control unit are detected to be faulty during the POST; in this state only the controller is powered to be able to diagnose the problem.

The low-level state machine for the power control unit defines the so-called recalls and the transitions between these recalls. Each recall denotes a required setting of the high and low voltage terminals, i.e., whether an individual terminal needs to provide power or not. These settings are described in a separate configuration file (not described in this paper).

To realize a particular recall, the controller compares the current status of the low- and high voltage terminals, which it has stored in volatile memory, with the desired status of the low and high voltage terminals. If the current status is different from the desired one, the controller starts communicating with the distribution panels to change the status. The transition from one recall to another may take a considerable amount of time, because of the inherently slow LonWorks communication. Depending on the chosen hardware components by the customer, there are two or three cabinets and it takes between 10 and 30 seconds to address all distribution panels.

Transitions between recalls are not atomic, that is, during such a transition a stimulus might lead to another required recall. To represent the state of these transitions, each main state consists of three substates:

- Entry: the controller compares the current status of the low and high voltage terminals with the desired recall. If they are different the next substate is Transitioning, otherwise it is Stable, except for the first recall where it stays in Entry.

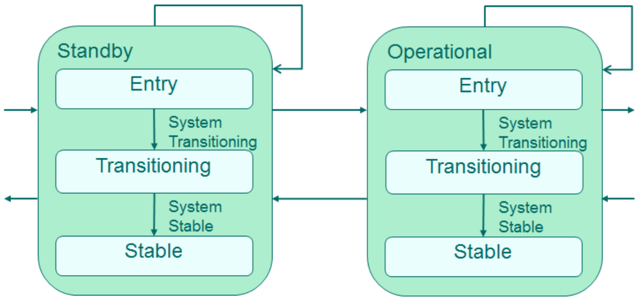


Fig. 3. Two main states and their substates

TABLE I
LINE OF A CONFIGURATION FILE FOR A LOW-LEVEL STATE MACHINE

| | |
|---|---|
| 2 2 0 00000000 00000000 112 4 2 | # < OPERATIONAL > recall 2<br># exit out of forced off |

- Transitioning: the controller is busy changing the state of the low and high voltage terminals.
- Stable: all distribution panels have reached the desired state for the low and high voltage terminals.

Fig 3 shows part of the high level VisualState state machine with two main states and their substates. The main states and substates are fixed, whereas the number of recalls is variable and defined in the configuration file. The low-level state machine and the recalls are different for every system release. The configuration file describes for each recall and stimulus, possibly with a given guard, what the next recall is and between which main states it has to transition. This is all coded in numbers. The main states are numbered, e.g., Standby = 2 and Operational = 3 and similar for the substates: Entry = 0, Transitioning = 1, and Stable = 2. Also all stimuli and all transitions between the main states have a fixed number. The recalls have a configurable number. The guard of a transition consists of two values: the relevant values of a status register and a mask. Table I shows an example of a line in the configuration file. Everything after a # is a comment.
The first three columns of Table I describe the state or -1 if it does not care.

1) The first column is the main state which is the source of the transition (in this example, state 2 denotes Standby).
2) The second column the substate which is the source of the transition (here 2 denotes substate Stable).
3) The third column the source recall (here 0 denoting that all terminals are off).

The fourth and fifth column describe the guard.

4) The fourth column describes the bits of the status register.
5) The fifth column the mask that will be applied.

The other columns have the following meaning:

6) The sixth column, describes the stimulus number (in this example, 112 denotes pushing the on button for 3 seconds).
7) The seventh column is the number of the specified transition between two main states (it might be a self-transition).
8) The eighth column describes the required recall (recall 2 in this example). By default, the substate will be Entry.

TABLE II
PART OF A TEST CASE

| | | | | | |
|---|---|---|---|---|---|
| PDS:QUE9:PAR | NoErr | 30000 | 1000 | 2000 | On Button |
| PDS:SYST? | 3:2:2 | 1000 | 1000 | 2000 | System On |

For performance reasons, this file is sorted on the sixth column. That is, the file is sorted on stimulus number and not on state, which hampers readability.

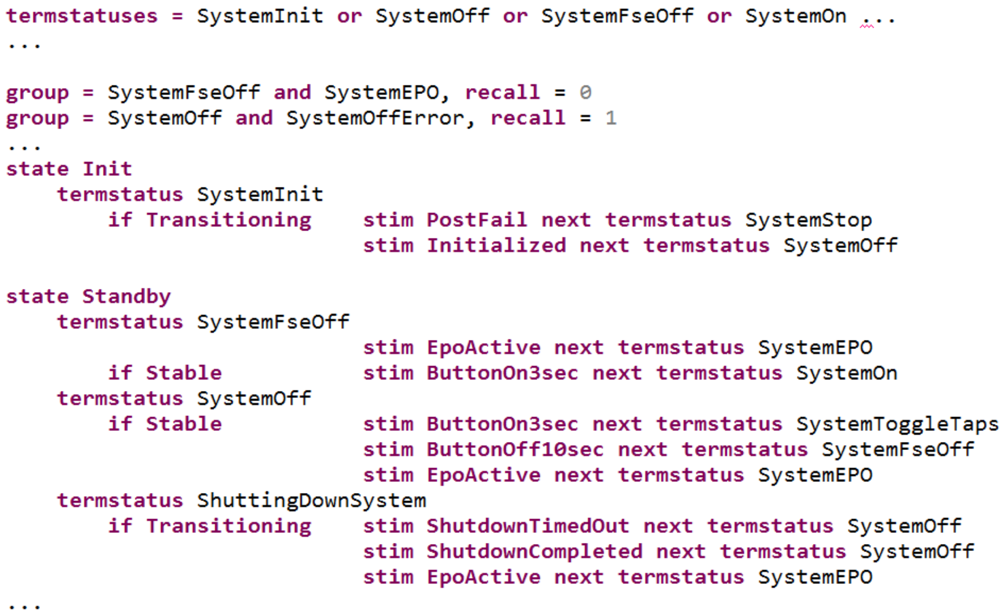
To test the state machine, there is an automated test tool running on a companion PC that connects to the controller of the power control unit via Ethernet. It can inject stimuli and ask the current state. A test case is a comma separated file. Table II shows two lines of a test case.

1) The first column is the network command that is send from the test tool to the power control unit (QUE injects a stimulus into the state machine and SYST asks the current state).
2) The second column is the expected response from the power control unit to the test tool (“NoErr” means that the command is successfully parsed and “3:2:2” is the current state, with main state 3 (Operational), substate 2 (Stable), and recall 2).
3) The third column is the time (in milliseconds) that the test tool waits before it sends the next command to the power control unit (in this example, the test tool waits 30 seconds between the QUE and the SYST command).
4) The fourth column is a time-out (in milliseconds) on the reply of the power control unit; within this amount of time the power control unit should send a message to indicate that it accepts the command.
5) The fifth column is a time-out (in milliseconds) on the response of the power control unit (as specified in the second column).
6) The sixth column contains comments.

A test case fails on a wrong response or on a time-out of the accept message or the response.

Every night a test suite is executed multiple times on two power control units with Jenkins [17] and the developers will find the results of the test execution in their mailbox. If test cases fail, a lot of time is spent investigating the cause and solving it in either the software of the power control unit or the test case. Test cases often fail because of timing issues in which the power control unit and the test tool are out of sync; this is almost always caused by the unreliable timing nature of LonWorks. The solution for such timing issues is an increase of the time bounds in the test cases. This results in long-lasting test cases with a lot of waiting time.

Concluding, the configuration file is hard to read, to change and to maintain, but has to be updated for every new system release. The same holds for the test cases which typically need to be updated manually for every new configuration file.



```
termstatuses = SystemInit or SystemOff or SystemFseOff or SystemOn ...
...

group = SystemFseOff and SystemEPO, recall = 0
group = SystemOff and SystemOffError, recall = 1
...
state Init
    termstatus SystemInit
        if Transitioning        stim PostFail next termstatus SystemStop
                                stim Initialized next termstatus SystemOff

state Standby
    termstatus SystemFseOff
                                stim EpoActive next termstatus SystemEPO
        if Stable               stim ButtonOn3sec next termstatus SystemOn
    termstatus SystemOff
        if Stable               stim ButtonOn3sec next termstatus SystemToggleTaps
                                stim ButtonOff10sec next termstatus SystemFseOff
                                stim EpoActive next termstatus SystemEPO
    termstatus ShuttingDownSystem
        if Transitioning        stim ShutdownTimedOut next termstatus SystemOff
                                stim ShutdownCompleted next termstatus SystemOff
                                stim EpoActive next termstatus SystemEPO
...
```

Fig. 4. Configuration DSL

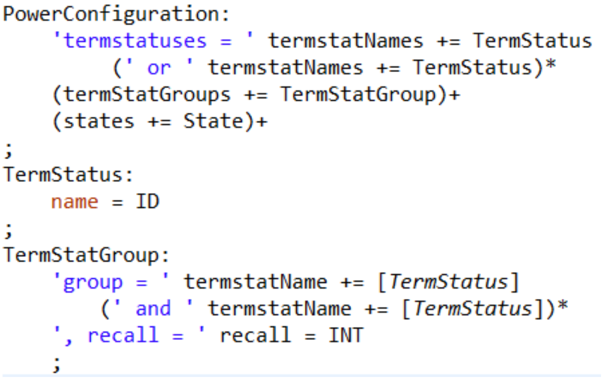
## III. Configuration DSL

To solve the problems mentioned above, we created a configuration DSL for the power control unit, using Eclipse, Xtext and Xtend [8]. This technology was chosen because the second author was familiar with it and the availability of a manual [18].

We explain the configuration DSL based on an instance fragment of the language, as shown in Fig. 4. Since the main states and their substates are always the same, there is no need to define them explicitly. The main purpose is to define the recalls and their transitions. To improve readability, the first part of the DSL instance defines meaningful names for the required status of the terminals, here called *termstatus*. Since several termstatuses might correspond to the same required settings of the terminals, the second part of the DSL groups the termstatuses and associates a recall number with each group. The third part defines the low-level state machine, where for a main state, a termstatus, and a stimulus we define the next termstatus. A transition might have a condition - indicated by the *if* keyword - on the current substate. Note that each termstatus belongs to exactly one main state, so the *next* relation implicitly defines the next main state.

The grammar for this language has been expressed in Xtext; a fragment is depicted in Fig. 5. Based on this grammar, the Xtext framework generates an editor for the language with, for instance, content assist. This makes it easy to define instances of the languages, such as the instance shown in Fig. 4.

The Xtext framework also provides suitable primitives for language validation and the generation of files from instances. In our application, the Xtend language has been used to generate a configuration file from language instances. This generator produces a line in the configuration file for every rule in the language instance. After the generator has generated all the rules, it sorts the lines on the value in the sixth column before writing it to a configuration file. A fragment of a generated configuration file is shown in Table III.



```
PowerConfiguration:
    'termstatuses = ' termstatNames += TermStatus
         (' or ' termstatNames += TermStatus)*
    (termStatGroups += TermStatGroup)+
    (states += State)+
;
TermStatus:
    name = ID
;
TermStatGroup:
    'group = ' termstatName += [TermStatus]
        (' and ' termstatName += [TermStatus])*
    ', recall = ' recall = INT
    ;
```

Fig. 5. Grammar of the Configuration DSL

The first eight columns, till the # sign, are the same as the manually created configuration file described in the previous section. In the comment part, the generator writes the first letter of the source main state, the source substate name, the stimulus, followed by the first letter of the target main state and the target substate name.

Additionally, we have generators that yield for every language instance a set of UML state diagrams, at several levels of abstraction, using PlantUML [19]. An impression of a generated diagram is given in given in Fig. 6 (not readable for reasons of confidentiality).

TABLE III
GENERATED CONFIGURATION FILE

```
6 0 0 00000000 00000000 109 19 2     # E.SystemEPO -> BUTTON_ON10SEC -> O.SystemOn
2 2 1 00000000 00000000 112 4 3      # S.SystemOff -> BUTTON_ON3SEC -> O.SystemToggleTaps
2 2 0 00000000 00000000 112 4 2      # S.SystemFseOff -> BUTTON_ON3SEC -> O.SystemOn
2 -1 1 00000000 00000000 115 6 0     # S.SystemOff -> BUTTON_OFF10SEC -> S.SystemFseOff
5 -1 2 00000000 00000000 117 21 5    # O.SystemOnError -> BUTTON_OFF -> S.ShuttingDownSystemError
3 2 2 00000000 00000000 117 5 5      # O.SystemOn -> BUTTON_OFF -> S.ShuttingDownSystem
3 2 3 00000000 00000000 134 7 2      # O.SystemToggleTaps -> TIMER_EXPIRED -> O.SystemOn
```

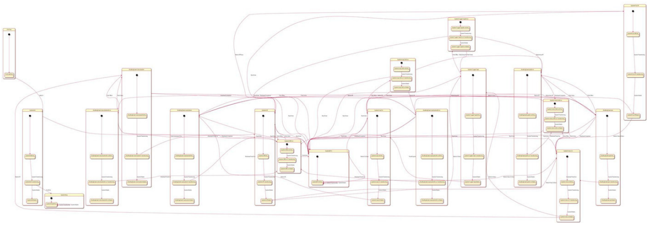
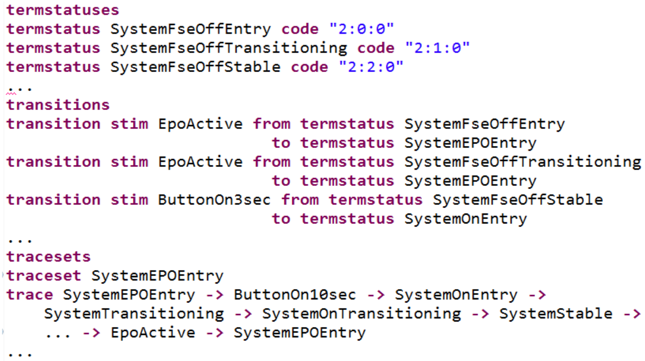


Fig. 6. State Diagram of the Configuration DSL



```
termstatuses
termstatus SystemFseOffEntry code "2:0:0"
termstatus SystemFseOffTransitioning code "2:1:0"
termstatus SystemFseOffStable code "2:2:0"
...
transitions
transition stim EpoActive from termstatus SystemFseOffEntry
                             to termstatus SystemEPOEntry
transition stim EpoActive from termstatus SystemFseOffTransitioning
                             to termstatus SystemEPOEntry
transition stim ButtonOn3sec from termstatus SystemFseOffStable
                             to termstatus SystemOnEntry
...
tracesets
traceset SystemEPOEntry
trace SystemEPOEntry -> ButtonOn10sec -> SystemOnEntry ->
    SystemTransitioning -> SystemOnTransitioning -> SystemStable ->
    ... -> EpoActive -> SystemEPOEntry
...
```

Fig. 8. Test DSL

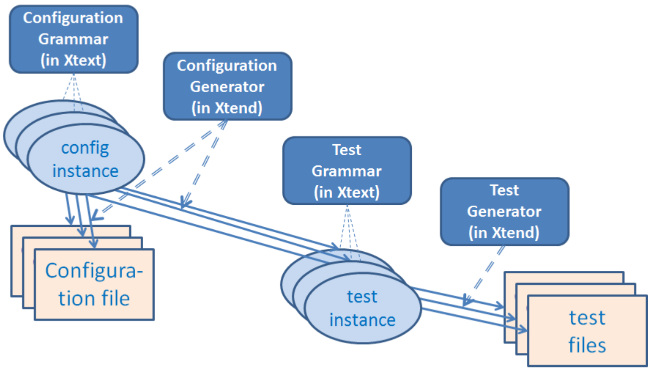

Configuration Grammar (in Xtext)
Configuration Generator (in Xtend)
config instance
Configura-tion file
Test Grammar (in Xtext)
Test Generator (in Xtend)
test instance
test files

Fig. 7. Overview of the two DSLs

## IV. Test DSL

Given a formalized representation of the configuration files, we investigated the possibilities to generate test cases from this DSL. A preliminary attempt to define a generator for test cases indicated that this was possible, but became rather complex since it included strategies to define test cases as well as translations to the current test format. To separate these issues, we aimed for a more abstract representation, related to the work described in [20] concerning a tool-independent representation of test design techniques. In our context, we defined a second DSL for the definition of tests to obtain a representation of tests which is independent of the current test techniques. Fig. 7 shows the relation between these DSLs. Note that with the Xtext approach there is no separate definition of the abstract grammar with meta models.

The test DSL is explained using the instance fragment depicted in Fig. 8. Note that this instance is generated from an instance of the configuration DSL. In the first part, for each termstatus of the configuration instance three extended termstatuses are generated corresponding to the three substates, by adding *Entry*, *Transitioning*, and *Stable* behind the name. The string after keyword *code* matches the first three columns of the VisualState configuration file; it is obtained using the code of the main state of the termstatus, the code of the substate, and the recall number defined in the configuration instance.

The second part lists all possible transitions. This is used to generate a report about coverage of termstatuses and transitions and to generate additional tests for stimuli that should not lead to a transition. In the third part, one or more trace sets are defined. Each trace set has one or more traces. A trace consists of an initial extended termstatus, and a number of pairs consisting of a stimulus and a next extended termstatus. Each trace starts and ends with the same extended termstatus, which makes it possible to run a number of traces in one go. Note that this requirement makes the generation of an instance of the test DSL a bit more complicated.

The generator of the trace DSL generates a test case, as shown in Table IV, for each trace in the language instance. Every Transitioning and Stable termstatus will result in a line in the test case with a SYST command that expects the string defined after the *code* key word in the *termstatuses* part of the instance. Since Entry substates are not observable (except for the first one), they are omitted. A stimulus will result in a line in the test case with a QUE command. Note that the third column is slightly different from table II; instead of a waiting time, now also an event can be specified. Such an event is used to synchronize with the power control unit and avoids long waiting times. It makes the testing process much faster.

Additionally, we used PlantUML to generate a visualization of a test trace as a sequence diagram. An examples is depicted in Fig. 9.

TABLE IV
PART OF A GENERATED TEST CASE

| | | | | | |
|---|---|---|---|---|---|
| PDS:SYST? | 6:00:00 | 2500 | 1000 | 2000 | SystemEPOEntry |
| PDS:QUE9:PAR | NoErr | 2500 | 1000 | 2000 | ButtonOn10sec |
| PDS:FWV? | 3.0.0.0 | T016_TRANS | 1500 | 2000 | SystemTransitioning |
| PDS:SYST? | 3:01:02 | 1000 | 1000 | 2000 | SystemOnTransitioning |
| PDS:FWV? | 3.0.0.0 | T017_STABLE | 1500 | 2000 | SystemStable |
| PDS:SYST? | 3:02:02 | 1000 | 1000 | 2000 | SystemOnStable |

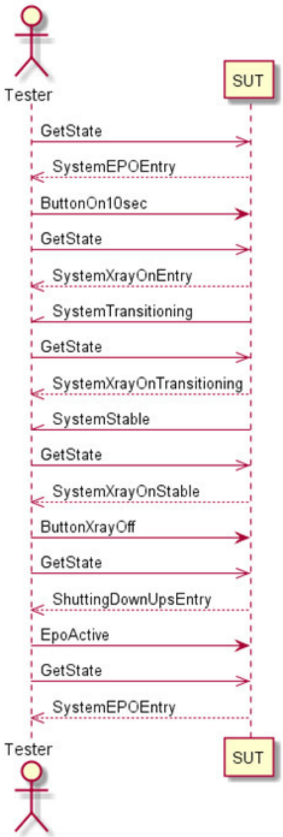


Fig. 9. Generated Sequence Diagram of a Test Case

The generator also generates a coverage file. Based on the selected termstatuses, transitions and traces, it creates a list of covered states, uncovered states, covered transitions and uncovered transitions. See Fig. 10.

Typically, we create two trace sets. The first trace set covers all transitions and is used for state and transitions coverage. A disadvantage of this set is that it also tests Emergency Power Off (EPO), which implies that only the controller of the power control remains powered; the distribution panels will lose power including the processor inside. Since this will rarely happen during normal usage, a second set is created with more realistic user scenarios where the distribution panels stay powered. Jenkins is configured to run these tests every weekend many times to test the reliability of the software running inside the power control. Outside the weekend, the full test set is run every night.

```
Covered states
 SystemFseOffEntry
 SystemFseOffTransitioning
 SystemFseOffStable
 SystemToggleTapsEntry
 SystemToggleTapsTransitioning
 SystemToggleTapsStable
 ...
Uncovered states
 PrePostEntry
 PrePostTransitioning
 PrePostStable
 ...
Covered transitions
 EpoActive
 SystemTransitioning
 SystemStable
 ButtonOn3sec
 ButtonOff10sec
 ...
Uncovered transitions
 MdsOn
 Initialized
 PdmCommissionTmo
 PostFail
 ...
```

Fig. 10. Coverage File

In the work described above, we used our own algorithm to generate test cases. As a next step we investigated the use of an existing tool to generate the test cases. We selected the SAL (Symbolic Analysis Laboratory) framework [21], [22], [23] which includes an automated test generator. The generator of the configuration DSL has been extended to generate two SAL files, corresponding to the two trace sets described above. When the test generator of SAL is supplied with a test goal - in our case the goal is to cover all transitions - it will yield traces that satisfies the goal. From this information, an instance of the test DSL is generated automatically using a small script.

## V. CONCLUSION

We summarize the results in Sect. V-A. The additional generation of analysis models is described in Sect. V-B. Sect. V-C addresses the questions of Sect. I. A brief description is future work can be found in Sect. V-D.

### A. Results

We started with the configuration DSL and generated a configuration file for the current system release. This generated file was successfully tested on the target hardware. Comparing the generated file with the existing one, we found a number of issues in the existing file. It contained a non-existing transition and a number of transitions were missing. As a next step, we made a DSL instance for the next system release and generated the configuration file. The size of this new configuration file is about twice the size of the current file, which indicates the increasing complexity of our system releases.

With the generated test cases, approximately twice as much transitions are covered as the manually written tests. The manually written test cases only made transitions from the Stable substates. The generated test cases also make transitions from the Entry and Transitioning substates, which leads to twice as much transition coverage.

The manually written cases were very time-dependent, with many long waiting times. They could still fail by a slow response of hardware, which required some analysis and typically a further increase of the waiting times. By having all concepts described in a clear and concise way using DSLs, we could make the test cases much more efficient. Instead of waiting all the time, the test tool now synchronizes with the power control unit and immediately resumes the test case once the power control unit has reached the desired state.

### B. Analysis models

In addition to the generated configuration and test files, we generated a number of analysis models. From the configuration DSL we generated models for the simulation tooling of POOSL [24] and the model checker mCRL2 [25]. The generators for mCRL2 and POOSL combine the behaviour of the high-level state machine and the low-level state behaviour described by the (generated) configuration file, into one state machine describing the complete system start-up and shutdown behaviour. Implicitly, these generators define the semantics of the DSLs. The use of multiple tools increases the confidence in the correctness of the generators.

The advantage of having a separate test DSL is that we can also generate tests or checks for the analysis models. For POOSL, we generated a tester process that communicates with the generated state machine. Every test trace results in a method that applies stimuli to the state machine process and checks whether the returned responses are as expected. The results are written to a file with a test report. For mCRL2 we generated properties to express that the expected traces occur in the generated state machine. This property can then be checked by the mCRL2 tools.

Using POOSL and mCRL2, we detected problems in the DSL instance, e.g., the simulation in POOSL revealed a missing condition. Moreover, we can detect whether a termstatus occurs in multiple main states. An alternative would be to use the validation possibilities of the Xtext/Xtend framework.

### C. Evaluation

We discuss the questions listed in the introduction (Sect. I):

- *How much time is needed to learn the tools and techniques?*
  Clearly, the learning curve for new techniques depends on previous education and knowledge. The work reported here was mainly done by the first author who independently learned the DSL techniques from the manual mentioned before [18]. With a Master's in Computer Science, including course about grammars and formal techniques, the basic part of the manual requires 4 hours to install the tools and to redo the examples of the manual. This was enough to get started with the case study.
- *How much effort is needed to migrate the current legacy component to a component which is defined by a high-level human-usable DSL?*
  It took about 35 hours to create the two DSLs presented here and to integrate them with the power control unit and the test tool. This step was sufficient to demonstrate the usefulness of the approach to management. In later increments, we added the generators for the analysis models and the use of SAL for test generation. Since the adaptation of grammars and generators is relatively easy and fast, the approach supports an incremental way-of-working. The Eclipse/Xtext framework is quite mature and provides many basic features such as syntax highlighting, auto completion, and content assist.
- *Does the DSL approach support the combination with analysis techniques such as simulation tools and formal model checkers?*
  Given earlier experience with POOSL, mCRL2, and SAL, it was straightforward to write generators for these languages. For each of the three languages mentioned, this took about 5 hours. The generators to visualize the state diagram and test traces using PlantUML requires only a few hours of work.
- *What are the benefits of introducing these new techniques compared to the current way of working?*
  The complexity of our configuration files is expected to increase quickly; as already mentioned above, the file size almost doubled for a new product release. With an investment of only 35 hours, we are ready to deal with this increasing complexity. We can now create the configuration and test cases for the power control unit in a readable, easy to change and maintainable format. The tests are now 3 times faster with a double coverage. It is expected that for a new product release we need only 8 hours instead of the estimated 60 hours.

Our experience is in accordance with a recent report about the state of practice in model-driven engineering [26]. It shows that most successful applications of model-driven development use small DSLs.

### D. Future Work

Future work includes an extension of the configuration DSL such that also the configuration file which contains the details of the recalls (i.e., the required setting of the high and low voltage terminals) can be generated. Similar to the low-level state machine, also this configuration file is hard to read, to change, and to maintain.

As a next step, we intend to remove the Visual State framework, and generate the full state machine directly in the C programming language. The generators for POOSL, mCRL2, and SAL are a good indication that this will be feasible.

## Acknowledgment

We thank the anonymous reviewers for a number of useful suggestions for improvement.

## References

[1] A. van Deursen, P. Klint, and J. Visser, "Domain-specific languages: An annotated bibliography," *SIGPLAN Notices*, vol. 35, no. 6, pp. 26–36, 2000. doi: http://dx.doi.org/10.1145/352029.352035

[2] M. Fowler, *Domain Specific Languages*. Addison-Wesley Professional, 2010.

[3] M. Voelter, S. Benz, C. Dietrich, B. Engelmann, M. Helander, L. C. L. Kats, E. Visser, and G. Wachsmuth, *DSL Engineering - Designing, Implementing and Using Domain-Specific Languages*. dslbook.org, 2013.

[4] J. Tolvanen, R. Pohjonen, and S. Kelly, "Advanced tooling for domain-specific modeling: MetaEdit+," in *The 7th OOPSLA Workshop on Domain-Specific Modeling*, 2007.

[5] P. Klint, T. van der Storm, and J. Vinju, "EASY meta-programming with Rascal," in *Generative and Transformational Techniques in Software Engineering III*, ser. Lecture Notes in Computer Science. Springer, 2011, vol. 6491, pp. 222–289.

[6] L. Kats and E. Visser, "The Spoofax language workbench. rules for declarative specification of languages and IDEs," in *The 25th Conference on Object-Oriented Programming, Systems, Languages, and Applications, OOPSLA 2010*, 2010. doi: http://dx.doi.org/10.1145/1869459.1869497 pp. 444–463.

[7] Software Technology Group, TU Dresden, "EMFText," http://www.emftext.org/, 2011, version 1.4.0.

[8] L. Bettini, *Implementing Domain-Specific Languages with Xtext and Xtend*. Packt Publishing Ltd, 2013.

[9] J. Kärnä, J.-P. Tolvanen, and S. Kelly, "Evaluating the use of domain-specific modeling in practice," in *The 9th OOPSLA workshop on Domain-Specific Modeling*, 2009.

[10] I. Nagy, L. Cleophas, M. van den Brand, L. Engelen, L. Raulea, and E. Mithun, "VPDS: A DSL for software in the loop simulations covering material flow," in *17th Int. Conf. on Engineering of Complex Computer Systems (ICECCS)*, 2012, pp. 318–327.

[11] K. Chandrasekaran, S. Santurkar, and A. Arora, "Stormgen - a domain specific language to create ad-hoc storm topologies," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 2. IEEE, 2014. doi: http://dx.doi.org/10.15439/2014F278 pp. 1621–1628.

[12] "Meta programming system (MPS)," http://jetbrains.com/mps, 2015.

[13] M. Voelter, D. Ratiu, B. Schaetz, and B. Kolb, "Mbeddr: An extensible C-based programming language and IDE for embedded systems," in *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity (SPLASH '12)*. ACM, 2012. doi: http://dx.doi.org/10.1145/2384716.2384767 pp. 121–140.

[14] M. Voelter, "Generic tools, specific languages," Ph.D. dissertation, Delft University of Technology, 2014.

[15] "LonWorks," http://www.echelon.com/technology/lonworks/, 2015.

[16] "VisualState," http://www.iar.com/Products/IAR-visualSTATE/, 2015.

[17] "Jenkins," http://jenkins-ci.org/, 2015.

[18] A. Mooij and J. Hooman, "Creating a domain specific language (dsl) with Xtext," http://www.cs.ru.nl/J.Hooman/DSL/, 2015.

[19] "PlantUML," http://plantuml.sourceforge.net/, 2015.

[20] M.-F. Wendland, "Abstractions on test design techniques," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 2. IEEE, 2014. doi: http://dx.doi.org/10.15439/2014F316 pp. 1575–1584.

[21] N. Shankar, "Symbolic analysis of transition systems," in *Abstract State Machines: Theory and Applications (ASM 2000)*, ser. Lecture Notes in Computer Science, no. 1912. Springer, 2000, pp. 287–302.

[22] ——, "Combining theorem proving and model checking through symbolic analysis," in *CONCUR'00: Concurrency Theory*, ser. Lecture Notes in Computer Science, no. 1877. Springer, 2000. doi: http://dx.doi.org/10.1007/3-540-44618-4_1 pp. 1–16.

[23] G. Hamon, L. de Moura, and J. Rushby, "Automated test generation with SAL," SRI International, CSL Technical Note, January 2005.

[24] B. D. Theelen, O. Florescu, M. Geilen, J. Huang, P. van der Putten, and J. Voeten, "Software/hardware engineering with the parallel object-oriented specification language," in *Proceedings of MEMOCODE'07*. IEEE, 2007. doi: http://dx.doi.org/10.1109/MEMCOD.2007.371231 pp. 139–148.

[25] S. Cranen, J. Groote, J. Keiren, F. Stappers, E. de Vink, W. Wesselink, and T. Willemse, "An overview of the mCRL2 toolset and its recent advances," in *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*. Springer, 2013. doi: http://dx.doi.org/10.1007/978-3-642-36742-7_15 pp. 199–213.

[26] J. Whittle, J. Hutchinson, and M. Rouncefiled, "The state of practice in model-driven engineering," in *IEEE Software*. IEEE, 2014. doi: http://dx.doi.org/10.1109/MS.2013.65 pp. 79–85.

DOI: 10.15439/2015F117
ACSIS, Vol. 5

# A Quality Attributes Approach to Defining Reactive Systems Solution Applied to Cloud of Sensors

Artur Skowroński
Schibsted Tech Polska Sp. z o.o.
ul. Armii Krajowej 28
Krakow, Poland
Email: artur.skowronski@schibsted.pl

Jan Werewka
AGH University of Science and Technology
Department of Apllied Computer Science
Krakow, Poland
Email: werewka@agh.edu.pl

***Abstract*—Reactive systems have been investigated and used for a long time. Due to new methods and new technology development, the reactive systems needs their redefinition. These systems are currently an interesting topic for IT (Information Technology) solution providers. In this paper the authors try to define a new view of the architecture of reactive systems, because reactive systems are evolving and there is no clear definition of them. The starting point of the investigation was the reactive manifesto which defines reactive systems by four main features (quality attributes): responsiveness, resilience, elasticity and message driven interoperability. The mentioned quality attributes are the basis for developing a system solution. For each of the quality attributes, a set of tactics are proposed to maintain attribute required behavior. The suitability of the proposed tactics was investigated for Reactive Sensor Middleware which is part of a CoS (Cloud of Sensors) in the PaaS (Platform as a Service) layer. A cloud of sensors for pollution monitoring in urban areas was used as an example. Verification of the tactics has confirmed that some of the proposed tactics are suitable for the selected CoS subsystem.**

***Index Terms*—Keywords Reactive systems, reactive manifesto, software architecture, quality attributes, tactics, cloud of sensors, pollution sensing**

## I. Introduction

IN THIS paper an approach is presented for the purpose of defining a software architecture model for a class of systems. The class considered here are reactive systems. In the classic book [1] on reactive systems design, systems are described by their characteristics: highly interactive, nonterminating process, interrupt driven, state-dependent response, environment-oriented response, parallel processes, usually stringent real-time requirements. The starting point of the analysis of reactive systems was the reactive manifesto [2], which defines reactive systems by its features. The manifesto stated that "a coherent approach to systems architecture is needed, and we believe that all necessary aspects are already recognized individually: we want systems that are responsive, resilient, elastic and message driven". The systems described by these features are called reactive systems. The reactive manifesto went through two revisions. The scalable and event driven traits from the previous version are replaced by elastic and message driven in the current version. We found this clarification interesting, especially when observing changes in dependency diagram which is part of both revisions. In the previous version the traits are all connected bidirectional lines, suggesting that all of them depends on each other, which was not very informative. That changed in Reactive Manifesto 2.0. The current iteration contains far more interesting inner dependencies. Message Driven pattern was defined as basis for all other traits. Responsiveness seems to be the main goal of reactive systems, because all other traits depends on it.

The goal of this paper is to develop an system solution based on selected quality attributes responsiveness, resilience, elasticity and message driven pattern. For the predefined set of quality attributes architectural tactics are proposed. The tactics will be used in determining architectural models. The approach is based on the classic software architecture development process proposed by Software Engineering Institute [3][4]. In the literature different solutions are proposed in the field. In [5] a general approach is proposed for embodying nonfunctional requirements (NFRs) into software architecture using architectural tactics. In [6] the influence of quality properties on decision making regarding software architecture was investigated.

## II. Quality Attributes of Reactive Systems

Quality attributes are referred to as Non Functional Requirements (NFR) and represent a desirable behavior of the system and are key success factors in developing software architecture. In next subsections three quality attributes are considered: elasticity, resilience and responsiveness.

### A. Elasticity

Elasticity is the ability of a system to scale resources up or down with minimal latency for different environment behavior during system runtime for different time periods. The ability can be reached manually or automatically. For the reactive systems the following sub attributes can be distinguished:

Consistent System Load. The load of IT systems can be not evenly distributed. The system should be able to scale both up and down - changing dynamically and automatically the amount of resources allocated. The goal is to handle user requests in a predictable, consistent manner. Attribute metrics should represent a standard level of usage of resources. It‘s lower boundary should not be set too low (the system is wasting resources, over-provisioned) while it‘s upper boundary

978-83-60810-66-8/$25.00 ©2015, IEEE

should not be set to high - in this case, we are in constant risk of throttling the system at any moment (under-provisioning).

Latency in allocating and deallocating resources. The system should be able allocate resources to achieve elasticity corresponding to the current load. Due to that time is a critical criteria - the amount of time needed to set up an additional application process should be as short as possible. The application should be small and granulated and it‘s startup time really quick beginning from the “cold” system and ending with the ability of handling user requests. The criteria chosen are time of allocating and de allocating new resources. The attribute metrics are numeric - startup time from the state before allocating new resources to the moment, when new resources are able to handle new users.

Scalability. An elastic system should be scalable in a predictable way. Its performance should improve proportionally to added resources. The overhead of adding new application instance should be as small as possible. It is desirable to achieve linear proportion between those two values (performance to capacity of added resources).

### B. Resilience

A resilient system is one that delivers a service that can be justifiably trusted when facing changes [9]. Resilience is related to a system’s ability of maintaining service provision without deviating from the fulfillment of system goals, despite changes that might affect the system or its environment. An example of resilience evaluation is presented in [7]. For description of resilience the following sub attributes are selected based on the “technologies” defined in [9].

Evolvability. It is ability to respond effectively to change. Within evolvability, an important topic is adaptivity, i.e., the capability of evolving while executing and retaining the notion of justified confidence.

Assessability. It is based on verification and evaluation. Classically, verification and evaluation are performed offline in a pre-deployment stage. In reactive systems assessment and evaluation has to be performed at run-time, during operation.

Usability. Computing systems have already pervaded all activities of our life, hence the importance of usability.

Diversity. Diversity should be advantageous in order to prevent vulnerabilities, e.g. have single points of failure.

From our perspective, there are additional sub attributes which resilient system should have:

Automaticity. Reactive systems, due to constant changes, should not be administrated by people only. A system should react automatically on changes in it (e.g. a situation when a given service is down or a new deployment is ready) and perform a predefined strategy, mitigating the occurrence of human errors which could cause the system to close down. In a resilient system, infrastructure should automatically resize itself to keep required quality.

Rationality. The system should be able to provide value even when it is partially inaccessible. In that case the system is built from the isolated micro services, it shouldn‘t happen that the system is not responding when one trivial part is not able to respond (e.q. we have parts of system unaccessible due to power blackout). This feature cannot be easily added to the system in the later stages and should be taken into account during the design process. A system should know which parts are important and cannot be missed (responding to a failure) and which parts are trivial (it’s just add value). Rational systems don‘t fail if there is no reason to do that.

### C. Responsiveness

Responsiveness refers to the ability of a system to fulfill assigned tasks within a given time as seen by the user.

Consistent response time. It is important to achieve consistent response time for a system request. Typically, mean response time is used as a metrics. Unfortunately, it doesn‘t say much about a time consistency of a request. Two systems with exactly the same mean can have a highly different pattern of behavior. One system can be predictable and consistent, the other may have quick as well as slow response times. For both systems their mean metrics will still be similar. Consistency increase user trust in the system reliability. The attribute metrics may be a standard deviation of response time for a full request-response loop. It describes how a system behaves in the long term in a given time and presents information about the best and worst cases.

Adapting data processing to the given usecase. Data from the system should be accessible as fast as they bring the value for the end client. Topic becomes more important when we are talking about communication between systems which are better adapted to fast data acquisition and usage. Internet of Things presents both new opportunities and challenges, that’s why Fast Data term becomes more and more important in synergy with Big Data systems.

## III. Tactics for Reactive Systems

A tactic is a design decision that aims to improve one specific design concern of a quality attribute [5]. Software architects utilize a rich set of proven architectural tactics and patterns to help satisfy specific quality concerns. Architectural patterns have an overarching impact on a software system, and are typically selected early in the design process. They determine the overall style of the design and include well-known solutions schemes [8]. Tactics and patterns are known architectural concepts; the work [9] provides more specific and in-depth understanding of how they interact. In the next subsections tactics are proposed for a previously selected set of quality attributes.

### A. Elasticity

*1) Ability to scale part of a system independently:* If we want to achieve usage of resources in a sufficient way, we need the ability to scale a different part of the system in response to its growing demand. Today, applications no longer rely on monolithic architecture. Parts of the system are developed in different technologies and they can have a very distributed necessity of system resources.

Sharding. Sharding is a specific type of database partitioning and its role is to split a database into smaller pieces, called shards, which are easier to manage, faster and less vulnerable to problems of global locking, meanwhile being easier to replicate due to its reduced size. Shards shouldn't share information between them, giving the ability to spread data between different physical instances and scale them independently, without the necessity of maintaining high-end, high-power systems. A common strategy is e.g. sharding data geographically, where we can take into account problem of latency too.

*2) Decrease overhead of single components:* Starting the executing of software components is the biggest factor in system latency. Below, are some tactics proposed, which are used to reduce software overheads

Containerization. Containerization and Software Containers are not new topics. However, they gain attraction in recent years, thanks to the support of IaaS Providers. In contrast to virtual machines it uses the kernel of the host machine and runs on isolated user space instances. Containers can be prepackaged and quickly distributed, with a minimal starting time of a single instance. It's worth to mention that there are particular Operating System Solutions created directly to work with Containerized software.

Lightweight technology stack. The bigger codebase, the longer load time can be felt by user. Splitting application into smaller pieces can drastically shorten time needed to run the new instance. Due to that fact, when want to achieve ability to run application instances dynamically on demand, we shouldn‘t relay on heavyweight enterprise technologies with huge amount of dependencies. Each application should have as little dependencies as possible and use external libraries/framework only when it‘s necessary.

*3) Allowing for Resource Balancing:* In the system which rely on the input from user, who can join or disconnect in any moment, amount of received data can vary dramatically over time. Having that in mind system should have ability for automatic resources balancing. Humans are error prone and too to slow to react while working with rapidly changing load. System should be able to allocate and deallocate resources to maximize ratio between throughtput and economic cost.

Implementing Backpressure. Backpressure in terms of responsive systems means the ability to acquire a feedback message from request passed through the system, e.g. by passing the message through the queue. Acquiring feedback is necessary to scale software in the correct way, e.g. the utilization of component resources. The implementation of backpressure is a nontrivial problem due to the asynchronous model of the communication. However, it is especialy important while dealing with the everlasting stream of data - there is possibility that the system under load will be overloaded by the incoming data.

Resource Managing based on Kernel Sharing Layer. An approach is used based on kernel features to provide the abstraction and isolation of system resources, instead of them on system-level defined rules, rather than monitoring the application and pooling its state through metrics values, such systems respond to application demands, by increasing its resources and spawning new instances (e.g. Apache Mesos [10] with Apache Aurora). It's solution fitted for Server Racks and Clusters to hide abstraction of multitenancy systems.

Immutable Designing. The less mutable state inside application, the easier application is to scale. In that case instances can‘t be easily replicated and a client can be served only by the instance with which started communicating. If the application is stateless, the load balancer can pass user input to whatever instance of application has free resources at the moment. The Domain Driven Design methodology is a good tool to evaluate which part of the application should be mutable and which not. Using good suited tools, such as functional programming languages or using message driven approach can be also benefitial.

### B. Resilience

The resilience of the system is an offshoot of both dependability and availability, defined to better suit the demands of an architecture based on micro services. Lack of availability of microservices sums up. The important thing is to design a system in such way that when one of the elements fails it will not bring down the whole application functionality.

*1) Replication:* The most obvious way to achieve high availability of all systems parts is to provide redundancy. This is especially preferable when a considered system is stateless and every request can be processed by any instance.

Bulkhead. Bulkheads are used in ships to create separate watertight compartments which prevents the ship from sinking. The idea can be effectively used in computer systems. In a similar manner, a computer system should have redundant components which are easily replicable whenever something happens to the system and one of its counterparts.

*2) Delegation:* In contrast to the classic synchronous method calls, a reactive system cannot use exception and exception handling due to its isolated nature. Thrown exceptions don‘t have a chance to reach component which is able to handle it. Information about failure should be delegated to another component able to resolve it.

Feedback supervisor. In the reactive system a supervisor should exist, which is a special component existing outside the standard flow of the systems and has information about the whole system. Whenever a failure occurs, corresponding information with the whole bounded context should be send to the supervisor. It‘s a great way to decouple the standard flow of the system from the failure support and error handling mechanism (e.g. Netflix‘s Hystrix).

*3) Isolation:* The main problem of distributed systems is the possibility of partial failure. We don‘t want errors propagate over the systems, dragging whole infrastructure down. Isolation is an important trait of the system created from the micro services, which assures that failure of one part does not spread over whole system.

State and behavior containment by Bounded Context. Every component should be as small as possible and enclose specific problem domains inside the bounded context. Boundaries are connected by messaging protocols. This ensure that the system architecture reflects the problem domain making it easy to evolve. It also promote component composability and modularization on the architectural level.

Containment of failures. It should be possible to contain a failure inside the block where it occurred. The failure shouldn't be promoted to the next block, inhibiting "disease spreading". No error should be able to cascade through the system. Whenever we do not provide the fallback, we should fail-fast to not saturate system resources and pass failure to the supervisor.

### C. Responsiveness

The responsiveness of the systems was investigated for a long time. In [5] six general principles for the synthesis of responsive software systems are presented: fixing, locality design, processing versus frequency tradeoff, shared resources, parallel processing, and centering. In the current systems developing a new approach for responsiveness tactics is essential.

*1) Deferred data validation:* The biggest difference between local software and software working throughout the network is the fact, that the distributed software has a far much longer feedback loop for each request. Whenever a user performs any action, it needs to wait to validate it on the server side, which can be a long operation striking user out of context. That's why it is important to sustain for the user an illusion of local work.

Normalization of data. System need to be able to cope with different type of inputs communicating on both a different protocols and data quality. That bring a necessity of bringing normalization layer which is able to retrieve common values from system, marking all data with an input specific metadata which can be used to additional analytics.

Data store synchronization. Thanks to better technology, it is possible to use data storage both on the client as on the server side. During work with web applications the user has feeling to work with native application by providing him with two data stores, a local and remote one. This is especially handy when both client and server are written in the same technology, sharing a common codebase. The user is working on the local copy bringing a short feedback. The local copy synchronization is performed in the background. The user is informed only whenever conflict happens.

Multiplayer game style data validation. To achieve smoothness of experience, programmers introduced a sophisticated system of the multiplayer game. Each player plays in his local environment and information about his actions is passed to the server, which confirms if its actions are possible to be done. The server has godlike power over each player and if he finds conflicts, resolves them and informs players about a verdict. Thanks to that each player has his own smooth experience.

*2) Sustaining consistent response time:* It is necessary to receive data as soon as they are able to be processed by system and end user. Synchronization should be done in the background. If our system responds to fast, responses it will be placed inside a buffer which should be maintained. If it will be too slow, we risk lowering the consistency of the overall system.

Streaming based data store To suit needs of the Fast Data system, our technology stack need to be adapted to processing not only huge amount of data, but also need to be able to process them in the fast way. Such a system needs to be able to combine storing huge amount of data from the many different concurrent inputs as a data warehouse, it also need to be able to deliver stream result in a quick way with a most current results for a waiting clients. An example of such a solutions can be Apache Storm and Apache Spark.

Non-blocking client-server communication. Communicating in a synchronous way is the biggest issue when trying to achieve consistent response times. A server communicating in a synchronous way is always waiting for a response and a one long request can delay a whole queue of operations. Each request should be responded to in the deferred way.

Worst case scenarios designing. To ensure that a response will be produced in reasonably, consistent time, it is important to use algorithms with low complexity. To provide a good level of stability in distributed systems, we need to bound a system with realistic timeouts. It is important not to break connections in case the data are processed correctly, but too long.

## IV. A Cloud of Sensor with Reactive Sensor Middleware

The proposed tactics for the reactive system will be verified for RSM (Reactive Sensor Middleware) which is part of a CoS (Cloud of Sensors) in the PaaS (Platform as a Service) layer.

The proposed RSM is very important because it is assumed that the number of sensors will increase rapidly. The physical sensors and a well-defined communication interface will be delivered by sensor providers. The sensor providers will deal with service installation, sensor infrastructure maintenance, and the sensor data offering to sensor service consumers over the Internet. The service consumers can use the sensor data with their own or other providers' applications, which are integrated with the physical sensors. The described solution is known as a CoS (Cloud of Sensors). This makes it possible for users to lease the sensing hardware and associated applications instead of buying the complete infrastructure.

There are similar solutions to CoS such as cyber physical cloud (CPC), the Internet of things (IoT) or Fog Computing [11], which is a paradigm that extends Cloud computing to the edge of the network and services can be hosted at end devices. All these solutions [12] have sensors and a cloud as an integral part of their architecture. CPS consists of computer (cyber) systems that interact with the physical world. CPC is simply a CPS with cloud integration. Thus, CPC is CPS with a cloud as its backbone for computation and communication. In contrast, the Internet of Things (IoT) enables pervasive and ubiquitous interconnection in near real-time on a massive scale with different remote devices (mostly sensors) that can be uniquely identified, located, and communicated with. Cost effective and scalable IoT solutions can be achieved using cloud-centric architecture.

Cloud of Sensors Solution Architectures Existing research into CoS and related architectures proposes various solutions.

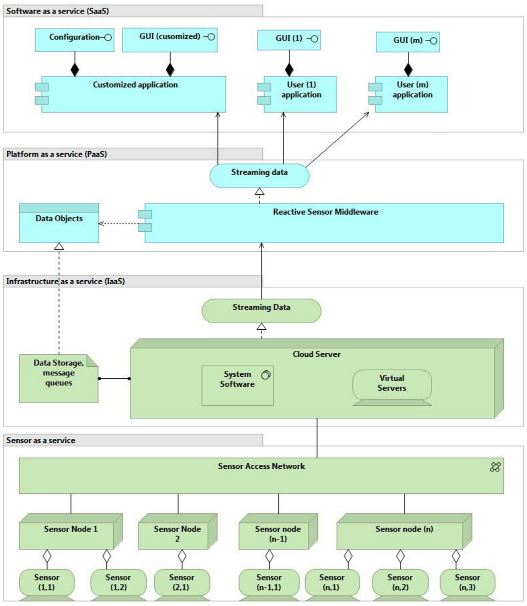


Fig. 1. CoS structure with Reactive Sensor Middleware located in PaaS layer

Some interesting examples are given below.

Paper [13] presents the design and evaluation of a Data Quality-Aware Sensor Cloud (DQS-Cloud) which is based on a cloud-based sensor data services infrastructure. The objective of the paper is to make a DQ as a multidimensional space in which all sensor devices produce multiple quality parameters or metadata such as accuracy, delay, frequency, latitude, longitude, sensor type, etc. Paper [14] proposes a new infrastructure called Sensor-Cloud infrastructure which virtualizes a physical sensor as a virtual sensor using cloud computing. An important issue is the developing of mathematical models (e. g. [15]) for the virtualization of sensor node resources. In [16] two alternative architectures for service management in IoT and sensor networks are discussed: based on Open Service Gateway (OSGi) framework and Remote Services for OSGi (R-OSGi) bundle.

For CPS systems other solution architectures are more suitable. In [17] a unified 5-level architecture is proposed as a guideline for implementation of CPS (Cyber Physical Systems). CPS solutions can be extended using SOA (Service-Oriented Architecture) solutions.

The CoS solution used for tactics verification of reactive systems, using cloud-centric architecture with Reactive Sensor Middleware in the PaaS layer is outlined on Fig. 1.

Pollution Monitoring in an Urban Area A cloud of sensors may have very broad applications. As an example an atmospheric pollution monitoring system was chosen in the paper. The goal of such a system is on-line pollution monitoring for decision making and the development of environmental policies, with the goal of reducing the impact of pollution on ecosystems and human health.

Some distinguished examples are given here. In [18] an interoperable system for air quality information management is proposed. The system is based on open-source standards-compliant tools and designed to develop a Spatial Data Infrastructure (SDI). In [19] a Pollution-Sense system for air pollution monitoring and control is presented. Participatory sensing combines the use of everyday mobile devices, such as cellular phones, GPS technology and location-based services, and sensors, to form interactive, bidirectional mobile sensing information systems. The system should provide large amounts of pollution data in time and space with different granularities. In [20] a method is described for the automatic detection of air pollution and fog using sensors mounted on vehicles. The described system consists of sensors which acquire their primary data from cameras and Light Detection and Recognition (LIDAR) instruments. In [21] a sensor cloud based on WSN (Wireless Sensor Networks) is proposed which virtualizes the wireless sensors and provides sensing as a service to users.

Different users of pollution monitoring systems can be distinguished. Some examples of users of such systems are: (1) Single persons or families (e.g. planning an excursion in an urban area); (2) People with allergies or breathing problems like asthma (e.g. planning to go outside); (3) Schools (e.g. planning different sport activities outside for children); (4) Early warning systems for municipal operation; (5) Urban planning used for air pollution reduction.

There are different types of pollution sensors which can be wearable, mounted on vehicles or UAVs, or installed in fixed positions (buildings, weather stations, etc.). The specialization of pollution sensors may differ significantly: (1) Personal environmental sensors, which are wearable sensors connected to or integrated with smart phones; (2) Pollution sensor stations installed on fixed points or on vehicles measuring air pollutants; (3) LIDAR (Light Detection and Recognition) used to detect small concentrations of air pollutants.

From the above discussion it is clear that providers of the sensor as a service layer can be individual persons or organizations with the motivation to receive payment for service proportional to sensor usage. The presented discussion aims only to be an overview of possible pollution system monitoring solutions, and is not intended to demonstrate the full picture.

## V. Verification of Tactics

The proposed set of tactics will be verified for an RSM which is part of a CoS placed in the PaaS layer. Of course not all of the proposed tactics for reactive systems will be suitable for RSM. Only the tactics most useful while building the aforementioned system will be selected.

**Elasticity** is a very important aspect of a system based on dynamically attached external data sources (in our case sensors). We do not know in advance the size of the streams of data that our application will be exposed to. For this reason correctly implemented resource balancing needs to be a core part of the infrastructure. Proper implementation of backpressure tactics is especially important as the system needs to be able to control the flow of data. A perfect solution would appear to involve quickly running and disabling new application instances when there is an urgent load increase or decrease, in order to setup and disable containers on the shared Resource, especially when the application is stateless and implemented using a light technology (a good example of such a stack is Node.js Amazon Lambda). In this case, for a cloud of sensors a sharding tactic should be proposed. Usually data received from sensors are closely related to geographical locations or data sensor types, therefore the data should be logically grouped to ensure processing efficiency. This makes it possible to easily separate data in order to mitigate latencies during querying of the data store.

**Resilience** is defined as the time taken by a system to return to an acceptable state after failure. The investigated cloud of sensors is highly dependent on the external data, therefore constant supervision is essential. A component which monitors the behavior of intermediate elements is needed, and this can be achieved by using a feedback supervisor tactic. A good level of isolation in the system should be provided as data can be acquired from different inputs and the instances collecting the data are independent. A containment of failure tactic can be used to prevent the failure of one instance impacting others. All instances should not be deployed in a single cluster, instead the system should use a bulkhead tactic which splits it into different physical locations, thereby mitigating the potential for a critical situation in which the system is unresponsive as a whole.

An important aspect of the design of the discussed sensor cloud from a **responsiveness** perspective is the use of a normalization of data tactic. This ensures that the data store always has data in a proper, common format and is able to execute all necessary analytic operations quickly and responsively, without intermediate steps. Also, providing non-blocking client-server communication is important if we want to achieve consistent response times for clients of our cloud solution.

## VI. Summary

The current reactive system's needs must be examined from different perspectives due to the emergence of new system solutions and new technologies. One such factor is the big data processing issue, caused by the need for real time data stream acquisition and visualization, which makes it important to perform architecture refactoring of reactive systems. The approach presented, which starts with the analysis of quality attributes and their tactics, seems proper at this stage of development. The investigation of the suitability of the proposed tactics for Reactive Sensor Middleware which is a part of CoS (Cloud of Sensors) placed in the PaaS (Platform as a Service) confirms the approach. The next stage will be experimentation on other real system examples, analyzing suitable architecture patterns and finally defining a reference architecture model for reactive systems.

## References

[1] R. J. Wieringa, "Design methods for reactive systems," 2003.

[2] "The reactive manifesto, published on september 16 2014. (v2.0), http://reactivemanifesto.org." [Online]. Available: http://reactivemanifesto.org

[3] L. Bass, P. Clements, and R. Kazman, *Software Architecture in Practice*. Boston, MA, USA: Addison-Wesley, Inc., 1998. ISBN 0-201-19930-0

[4] R. Wojcik and et al., "Attribute-driven design version 2.0, tr-023." SEI, Carnegie Mellon Univ, 2014.

[5] S. Kim, D.-K. Kim, L. Lu, and S. Park, "A tactic-based approach to embodying non-functional requirements into software architectures." in *EDOC*. IEEE Computer Society, 2008. ISBN 978-0-7695-3373-5 pp. 139–148. [Online]. Available: http://dblp.uni-trier.de/db/conf/edoc/edoc2008.html#KimKLP08

[6] H. R. E. Majidi, M. Alemi, "Software architecture: A survey and classification," *2010 Second International Conf. on Communication Software and Networks*, pp. 460–464, 2010.

[7] J. Cámara, P. Correia, R. de Lemos, and M. Vieira, "Empirical resilience evaluation of an architecture-based self-adaptive software system," ser. QoSA '14. New York, NY, USA: ACM, 2014. doi: 10.1145/2602576.2602577. ISBN 978-1-4503-2576-9 pp. 63–72. [Online]. Available: http://doi.acm.org/10.1145/2602576.2602577

[8] J.-C. Laprie, "From dependability to resilience," in *38th IEEE/IFIP Int. Conf. On Dependable Systems and Networks*, 2008.

[9] N. B. Harrison and P. Avgeriou, "How do architecture patterns and tactics interact? a model and annotation," *J. Syst. Softw.*, vol. 83, no. 10, pp. 1735–1758, Oct. 2010. doi: 10.1016/j.jss.2010.04.067. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2010.04.067

[10] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. H. Katz, S. Shenker, and I. Stoica, "Mesos: A platform for fine-grained resource sharing in the data center," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2010-87, May 2010. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-87.html

[11] S. Yi, C. Li, and Q. Li, "A survey of fog computing: Concepts, applications and issues," in *Proceedings of the 2015 Workshop on Mobile Big Data*, ser. Mobidata '15. New York, NY, USA: ACM, 2015. doi: 10.1145/2757384.2757397. ISBN 978-1-4503-3524-9 pp. 37–42. [Online]. Available: http://doi.acm.org/10.1145/2757384.2757397

[12] V. Sehgal, A. Patrick, and L. Rajpoot, "A comparative study of cyber physical cloud, cloud of sensors and internet of things: Their ideology, similarities and differences," in *Advance Computing Conference (IACC), 2014 IEEE International*, Feb 2014. doi: 10.1109/IAdCC.2014.6779411 pp. 708–716.

[13] A. Kothari, V. Boddula, L. Ramaswamy, and N. Abolhassani, "Dqs-cloud: A data quality-aware autonomic cloud for sensor services," in *Collaborative Computing: Networking, Applications and Worksharing, 2014 International Conference on*, Oct 2014, pp. 295–303.

[14] M. Yuriyama and T. Kushida, "Sensor-cloud infrastructure - physical sensor management with virtualized sensors on cloud computing," in *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, Sept 2010. doi: 10.1109/NBiS.2010.32. ISSN 2157-0418 pp. 1–8.

[15] S. Misra, S. Chatterjee, and M. Obaidat, "On theoretical modeling of sensor cloud: A paradigm shift from wireless sensor network," *Systems Journal, IEEE*, vol. PP, no. 99, pp. 1–10, 2014. doi: 10.1109/JSYST.2014.2362617

[16] D. Wilusz and J. Rykowski, "Comparison of architectures for service management in iot and sensor networks by means of osgi and rest services," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 2. IEEE, 2014. doi: 10.15439/2014F324 pp. pages 1207–1214. [Online]. Available: http://dx.doi.org/10.15439/2014F324

[17] J. Lee, B. Bagheri, and H.-A. Kao, "A cyber-physical systems architecture for industry 4.0-based manufacturing systems," *Manufacturing Letters*, vol. 3, no. 0, pp. 18 – 23, 2015. doi: http://dx.doi.org/10.1016/j.mfglet.2014.12.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S221384631400025X

[18] F. D'Amore, S. Cinnirella, and N. Pirrone, "Ict methodologies and spatial data infrastructure for air quality information management," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 5, no. 6, pp. 1761–1771, Dec 2012. doi: 10.1109/JSTARS.2012.2191393

[19] D. Mendez, A. Perez, M. Labrador, and J. Marron, "P-sense: A participatory sensing system for air pollution monitoring and control," in *Pervasive Computing and Communications Workshops, 2011 IEEE International Conference on*. doi: 10.1109/PERCOMW.2011.5766902 pp. 344–347.

[20] P. Sallis, C. Dannheim, C. Icking, and M. Maeder, "Air pollution and fog detection through vehicular sensors," in *Modelling Symposium (AMS), 2014 8th Asia*, Sept 2014. doi: 10.1109/AMS.2014.43 pp. 181–186.

[21] S. Madria, V. Kumar, and R. Dalvi, "Sensor cloud: A cloud of virtual sensors," *Software, IEEE*, vol. 31, no. 2, pp. 70–77, Mar 2014. doi: 10.1109/MS.2013.141

# 8th International Symposium on Multimedia Applications and Processing

## ORGANIZED BY

SOFTWARE Engineering Department, Faculty of Automation, Computers and Electronics, University of Craiova, Romania "Multimedia Applications Development" Research Centre

## Background and Goals

Multimedia information has become ubiquitous on the web, creating new challenges for indexing, access, search and retrieval. Recent advances in pervasive computers, networks, telecommunications, and information technology, along with the proliferation of multimedia mobile devices - such as laptops, iPods, personal digital assistants (PDA), and cellular telephones - have stimulated the development of intelligent pervasive multimedia applications. These key technologies are creating a multimedia revolution that will have significant impact across a wide spectrum of consumer, business, healthcare, educational and governmental domains. Yet many challenges remain, especially when it comes to efficiently indexing, mining, querying, searching, retrieving, displaying and interacting with multimedia data.

The Multimedia - Processing and Applications 2015 (MMAP 2015) Symposium addresses several themes related to theory and practice within multimedia domain. The enormous interest in multimedia from many activity areas (medicine, entertainment, education) led researchers and industry to make a continuous effort to create new, innovative multimedia algorithms and applications.

As a result the conference goal is to bring together researchers, engineers, developers and practitioners in order to communicate their newest and original contributions. The key objective of the MMAP conference is to gather results from academia and industry partners working in all subfields of multimedia: content design, development, authoring and evaluation, systems/tools oriented research and development. We are also interested in looking at service architectures, protocols, and standards for multimedia communications - including middleware - along with the related security issues, such as secure multimedia information sharing. Finally, we encourage submissions describing work on novel applications that exploit the unique set of advantages offered by multimedia computing techniques, including home-networked entertainment and games. However, innovative contributions that don't exactly fit into these areas will also be considered because they might be of benefit to conference attendees.

## Call for Papers

MMAP 2015 is a major forum for researchers and practitioners from academia, industry, and government to present, discuss, and exchange ideas that address real-world problems with real-world solutions.

The MMAP 2015 Symposium welcomes submissions of original papers concerning all aspects of multimedia domain ranging from concepts and theoretical developments to advanced technologies and innovative applications. MMAP 2015 invites original previously unpublished contributions that are not submitted concurrently to a journal or another conference.

Papers acceptance and publication will be judged based on their relevance to the symposium theme, clarity of presentation, originality and accuracy of results and proposed solutions.

## Topics

Topics of interest are related to Multimedia Processing and Applications including, (but are not limited) to the following areas:

- Audio, Image and Video Processing
- Animation, Virtual Reality, 3D and Stereo Imaging
- Big Data Science and Multimedia Systems
- Cloud Computing and Multimedia Applications
- Machine Learning, Data Mining, Information Retrieval in Multimedia Applications
- Multimedia File Systems and Databases: Indexing, Recognition and Retrieval
- Multimedia in Internet and Web Based Systems
- E-Learning, E-Commerce and E-Society Applications
- Human Computer Interaction and Interfaces in Multimedia Applications
- Multimedia in Medical Applications
- Entertainment and games
- Security in Multimedia Applications: Authentication and Watermarking
- Distributed Multimedia Systems
- Network and Operating System Support for Multimedia
- Mobile Network Architecture
- Intelligent Multimedia Network Applications
- Future Trends in Computing System Technologies and Applications

## Best Paper Award

A best paper award will be made for work of high quality presented at the MMAP Symposium. The technical committee in conjunction with the organizing/steering committee will decide on the qualifying papers. Award comprises a certificate for the authors and will be announced on time of conference.

## Steering Committee

**Amy Neustein,** Boston University, USA, Editor of Speech Technology

**Lakhmi C. Jain,** University of South Australia and University of Canberra, Australia

**Ioannis Pitas,** University of Thessaloniki, Greece

**Costin Badica,** University of Craiova, Romania

**Borko Furht,** Florida Atlantic University, USA

**Harald Kosch,** University of Passau, Germany

**Vladimir Uskov,** Bradley University, USA

**Thomas M. Deserno,** Aachen University, Germany

### Publicity Chair

**Amelia Badica,** University of Craiova, Romania

**Adriana Schiopoiu Burlea,** University of Craiova, Romania

### Organizing Committee

**Dumitru Dan Burdescu,** University of Craiova, Romania

**Costin Badica,** University of Craiova, Romania

**Marius Brezovan,** University of Craiova, Romania

**Liana Stanescu,** University of Craiova, Romania

**Cristian Marian Mihaescu,** University of Craiova, Romania

### Event Chairs

**Brezovan, Marius,** University of Craiova

**Burdescu, Dumitru Dan,** University of Craiova, Romania

### Program Committee

**Badica, Amelia,** University of Craiova, Romania

**Böszörmenyi, Laszlo,** Klagenfurt University, Austria

**Burlea Schiopoiu, Adriana,** University of Craiova

**Camacho, David,** Universidad Autonoma de Madrid, Spain

**Cano, Alberto,** University of Cordoba, Spain

**Cardoso, Jaime S.,** Universidade do Porto, Portugal

**Cretu, Vladimir,** Politehnica University of Timisoara, Romania

**Debono, Carl James,** University of Malta, Malta

**Fabijańska, Anna,** Lodz University of Technology, Poland - Institute of Applied Computer Science, Poland

**Fomichov, Vladimir,** National Research University Higher School of Economics, Moscow, Russia., Russia

**Giurca, Adrian,** Brandenburg University of Technology, Germany

**Grosu, Daniel,** Wayne State University, United States

**Groza, Voicu,** University of Ottawa, Canada

**Grundspenkis, Janis,** Riga Technical University, Latvia

**Kabranov, Ognian,** Cisco Systems, United States

**Kannan, Rajkumar,** Bishop Heber College Autonomous, India

**Korzhik, Valery,** State University of Telecommunications, Russia

**Kotenko, Igor,** St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science, Russia

**Kriksciuniene, Dalia,** Vilnius University, Lithuania

**Lamas, David,** Tallin University, Estonia

**Lau, Rynson,** City University of Hong Kong, Hong Kong S.A.R., China

**Lloret, Jaime,** Polytechnic University of Valencia, Spain

**Logofatu, Bogdan,** University of Bucharest, Romania

**Luna, Jose,** University of Cordoba, Spain

**Mangioni, Giuseppe,** DIEEI - University of Catania, Italy

**Mannens, Erik,** Ghent University

**Mihaescu, Cristian,** University of Craiova, Reunion

**Mocanu, Mihai,** University of Craiova, Romania

**Morales-Luna, Guillermo,** Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Mexico

**Ogiela, Marek,** AGH University of Science and Technology, Poland

**Ohzeki, Kazuo,** Shibaura Institute of Technology, Japan

**Õunapuu, Enn,** Tallinn University of Technology, Estonia

**Paltoglou, Georgios,** University of Wolverhampton, United Kingdom

**Popescu, Dan,** CSIRO, Sydney, Australia, Australia

**Querini, Marco,** Department of Civil Engineering and Computer Science Engineering

**Rutkauskiene, Danguole,** kaunas university of technology

**Salem, Abdel-Badeeh M.,** Ain Shams University, Egypt

**Sari, Riri Fitri,** University of Indonesia, Indonesia

**Smedberg, Asa,** Stockholm University, Sweden

**Stanescu, Liana,** University of Craiova

**Tejera, Mario Hernández,** University of Las Palmas de Gran Canaria, Spain

**Trausan-Matu, Stefan,** Politehnica University of Bucharest, Romania

**Trzcielinski, Stefan,** Poznan University of Technology, Poland

**Tsihrintzis, George,** University of Piraeus, Greece

**Vega-Rodríguez, Miguel A.,** University of Extremadura, Spain

**Velastin, Sergio,** Kingston University, United Kingdom

**Virvou, Maria,** University of Piraeus, Greece

**Watanabe, Toyohide,** University of Nagoya

**Wotawa, Franz,** Technische Universitaet Graz, Austria

**Zurada, Jacek,** University of Louisville, United States

DOI: 10.15439/2015F184
ACSIS, Vol. 5

# Shape and colour recognition of dishes for the purpose of customer service process automation in a self-service canteen

Tomasz Kryjak, *Member IEEE*
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Krakow, Poland
Email: tomasz.kryjak@agh.edu.pl

Damian Król
Email: damians.krol@gmail.com

***Abstract*—In the article a vision system for shape and colour recognition of dishes (plates, bowls, mugs), which can be used to automate the process of customer service in a self-service canteen is described. In consists of three basic components: object segmentation using so-called background model subtraction, shape recognition using geometric invariant moments and SVM classifier, as well as colour recognition using a Gaussian model. In addition, recognition in case of close or abut objects using a distance transform like approach is presented. The solution was evaluated on a dedicated test stand with controlled LED lightning. A 98% accuracy was obtained on over 100 test images, which indicates that the solution could be used in business practise.**

## I. Introduction

IN TODAY'S world a rapid automation of customer services in different areas can be observed. Among numerous examples worth to mention are: the growing number of vending and ticket machines, automatic cash registers in supermarkets or recently emerging touch-screen kiosks for ordering and paying in fast food restaurants.

The customer service process in a self-service canteen can also be subjected to automation. This topic is particularly relevant in the context of the currently ongoing economic and social changes. The model, which assumes eating lunch or dinner after work at home changes to dining in canteens, lunch-bars or restaurants. Therefore, an increasing number of various types of self-service restaurants can be observed. There, the customer takes a tray, collects the meals and proceeds to the cash desk, where the content of the tray is priced by the cashier and the payment is made.

The automation of the described process requires the use of two modules: the pricing of products on the tray and the payment collection. The second issue is already well know and applied on a large scale: payment terminals (with PIN based authentication and so-called contactless cards) and CDMs (cash deposit machines). However, the automatic pricing of meals or dishes is still a challenge.

The topic of vision based food classification is addressed in several research papers. It should be noted that the issue is very challenging due to the great variety of possible meals, which are often quite similar. Another big problem is the segmentation of a particular food type on a plate.

In the work [1] the GrabCut segmentation is used to extract individual food types, as well as SURF features and SVM classifier. The accuracy of the system is more than 81%. In paper [2] the concept of using local and global features is described. Additionally, the results from several individual classifiers are combined to improve the accuracy, which for the system is about 80%. Similar results were obtained in the work [3]. The authors of the work [4] used a combination of deformable part-based and a texture model. For particular food recognition a multi-view multi-kernel SVM was utilized. The accuracy of this system reaches 90%.

This short analysis reveals that the topic of food recognition is up to date and intensively researched. However, the high complexity limits the accuracy of the systems to about 90%. This is insufficient for an automated pricing system in a self-service canteen. It is also worth noting, that for similar systems several patents can be found – for example from the SRI company [5]. In addition, an interesting system to automate the pricing of bakery products developed by Brain Corporation is presented in a video available at [6].

In this paper a pricing system based on video analysis of the shape and colour of dishes is presented. To the best knowledge of the authors, this is the only described in the literature vision system operating on this principle. The input to the system is a photo (single video frame) of a tray with one or more dishes (plates, bowls, mugs). Then, the individual objects are extracted and their shape and colour is determined. To use this information in the pricing task it has to be assumed that on a given day, on a particular dish (defined by shape and colour) and single meal is served. For example: on a *large round, green* plate pork chop with french fries and on a *large round, red* plate fish with potatoes is served. It is worth noting that such a system quite strictly defines the operation rules of the canteen. The meals have to be placed by the staff on appropriate dishes. This can be done on a regular basis (applying to the customer's request) or prepared in advance (the problem of keeping the meal warm). In addition, it should

The work presented in this paper was supported by AGH University of Science and Technology project number 15.11.120.476.

978-83-60810-66-8/$25.00©2015, IEEE

be assured that the boarder of the dish is free form parts of the meal.

The remainder of this paper is organized as follows. Section II contains a general overview of the proposed vision system. Detailed information about particular modules i.e. object segmentation, shape recognition, colour recognition and close or abut object recognition are given in Sections IV, V and VI. Evaluation of the system is presented in Section VII. The paper ends with a summary and indication of future research directions.

## II. Overview of the proposed vision system

This section provides basic information about the discussed vision system. Is has been implemented in C++ programming language using the OpenCV image processing library [7] and the Qt GUI library [8].

The designed and implemented algorithms were evaluated on a specially constructed test stand, which consisted of:

- digital camera – a typical USB camera with resolution 640 × 480 pixel was used (Logitech Webcam Pro 9000),
- illuminator – in order to provide good lighting conditions for image acquisition, a LED based illuminator was developed. It was a source of strong, white and diffused light. This reduced the problems of shadows and disturbances caused by external lightning,
- housing – to partially isolate the workspace from external lighting. It was also the mounting point of the illuminator and the camera,
- worktop – it was used for proper tray positioning.

In the application image processing and recognition was performed in two modes:

- continuous (for every frame) – tray and hand presence detection, background model acquisition and update,
- on-demand (at user request) – object segmentation and recognition followed by item pricing.

### A. Continuous mode

In order to provide proper classification, it is necessary to correctly place the tray in the field of view of the camera (i.e. in the workspace). It was assumed that the positioning in the axis perpendicular to the typical tray movement will be forced by two guide-rails (movement to and from the user), while in the other axis will be controlled by automatic detection of square markers. The markers are placed in the workspace, in a distance corresponding to the tray width. Using information about their visibility, the presence and position of the tray can be determined.

The second module working in the continuous mode is the detection of the so-called empty workspace i.e. a situation when in the camera's field of view there are no objects: plates, tray or even dirt. It such case, the current image is saved and than used in the segmentation phase as a background model (details in Section III).

The third module is responsible for detecting the presence of objects that have a common part with the boarder of the workspace. This prevents from starting the image analysis, when the user still holds the tray or has his hand in the camera's field of view. A detailed discussion of the three modules is presented in the work [9].

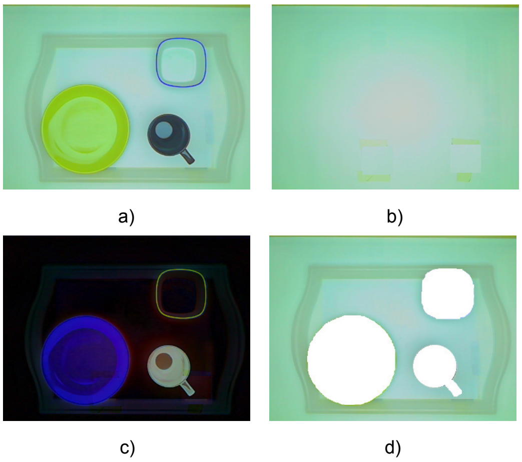


Fig. 1. Object segmentation using background subtraction. a) current workspace image, b) background model (image of an empty workspace), c) differential image, d) thresholding result overlaid on the input image (with same additional post-processing)

### B. On demand mode

When the tray is correctly positioned and the user has removed his hand from the workspace, the analysis can be started by pressing a button. It consists of the following steps:

- object segmentation,
- object shape recognition (also for close or abut objects),
- object colour recognition.

These steps will be discussed in detail later in the paper. The final result is the information about detected objects – their shapes and colours. On this basis, it is possible to price the tray.

## III. Object segmentation

The segmentation of objects (i.e. dishes/plates) is one of the most important elements of the described vision system. On its accuracy, to a large extent, depend the further processing steps: shape and colour recognition.

Taking into account the specificity of the task, the purpose of segmentation is to isolate:

- the correct shape of the objects – for classification,
- continuous edges of the objects, thick enough to retrieve colour information – the interior of the dishes can not be considered as a reliable source of information about colour, as it usually contains meals.

In addition, the method should reduce the possibility of connecting separate objects due to occurring shadows and have a tolerance to variable lighting conditions.

In the proposed solution object segmentation is based on the so-called background subtraction approach. From the current image with tray and plates, a background image is subtracted (view of an empty workspace). Thresholding the differential image allows to extract individual objects. The concept is illustrated in Figure 1. It is worth noting that in Figure 1b there are no markers visible on the worktop. This is due to a marker removal procedure, that involves replacing the marker's ROI with a workspace image from another location. This allows to obtain correct segmentation of the tray content, because markers are not detected.

The method has two main advantages. Its concept and the obtained results are easy to interpret i.e. all detected objects are regarded as dishes. Furthermore, it is very efficient. It should also be noted, that the empty workspace detection and background model update procedures allow to compensate lighting changes (e.g. naturally occurring during the day), which could affect the segmentation accuracy.

Limitations include: the need that objects or at least their boarder, have a colour different from the workspace and the requirement that the tray has the same colour as the workspace background – the tray should be transparent for the segmentation procedure.

In the current version of the algorithm, the segmentation is carried out in the RGB colourspace. First, the absolute value of the difference between the current image and the background model is computed – separately for each component. Then the maximal difference is selected and compared with a thershold. The obtained object mask is post-processed using morphological closing with a $5 \times 5$ square structuring element. Finally, a hole filling procedure is applied to eliminate the influence of the plate content on the segmentation result. A detailed description of the segmentation procedure is presented in the work [9].

## IV. Object shape recognition

The object shape recognition is a two-step process. First, a feature vector, which describes each shape, is generated. Then, this vector is assigned to a pre-defined class (i.e. classified). The input to this procedure is a binary object mask (containing a single object) and the output the class to which this object belongs.

### A. Feature vector

To describe the shape of an object (i.e. to generate a feature vector) a common approach involves the use of shape descriptors. This task is not easy, because on one hand a good differentiation of shapes is required (e.g. squares, circles, ellipses, etc.) is required. On the other hand the description must be insensitive to scaling, translation, rotation, affine transformations and some disruptions (e.g. "ragged" shape edge). In the literature a lot of different shape descriptors are described. They can be roughly divided into contour based (only the edge is analysed) and area based (the whole object is analysed). In the paper [10] four of them were evaluated: Fourier descriptors, curvature scale space descriptor, angular radial transform and image moments. The experiment showed that the geometric moments are a good solution for shape description. Furthermore, their are implemented in the popular image processing library OpenCV [7].

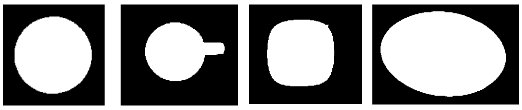


Fig. 2. The used shapes: circle, mug, rsquare, elipse

In the described system geometric invariant moments, also often referred to as Hu moments, are used. They have been proposed in the work [11] and are the basis of many shape recognition approaches. For example, in [12] they are utilized for human action recognition.

### B. Classifier

As a classifier the Support Vector Machine (SVM) algorithm was used. It was originality proposed by Vladimir Vapnik [13] and is one of the most popular machine learning algorithms. This is due to: intuitiveness of the method, good accuracy, high computational efficiency and quite simple and quick learning procedure (possible to automate because of small number of parameters).

In the basic version, the SVM is a binary linear classifier. However, a modification was proposed (so-called kernel trick), that allows the classification of non-linear problems. It involves the transformation of the feature vectors to a new space with higher dimensionality. Often as the kernel the Gaussian radial basis function (RBF) is used (the default solution in OpenCV library).

To enable the classification of more than one object type the multi-class problem is transformed into multiple binary classifications. For example, for three shapes $S_1, S_2, S_3$ three classifiers are required: $C_1$ to distinguish $S_1$ from $S_2$ and $S_3$, $C_2$ to distinguish $S_2$ from $S_1$ and $S_3$ and $C_3$ to distinguish $S_3$ from $S_1$ and $S_2$.

### C. Training dataset preparation

The use of a machine learning approach requires the preparation of three datasets: training, validation and test. The first is used to train the classifier i.e. to obtain the required parameters. The second to test various options (e.q. SVM parameters or number of used features). The last to asses the final solution. Typically, the input dataset is divided at a ratio of 60%, 20%, 20%.

In the current version of the system the following shapes are used:

- circle (big plate, small plate, bowl),
- mug (circle with a handle),
- rounded square (rsquare) (plate, small bowl),
- elipse (platter).

The used shapes are presented in Figure 2. Using these templates three datasets: training, validate and test were generated. For this purpose the templates were: scaled, rotated, subjected to affine transform, disturbed.

A template and exemplary samples (rotated and disturbed) are presented in Figure 3.

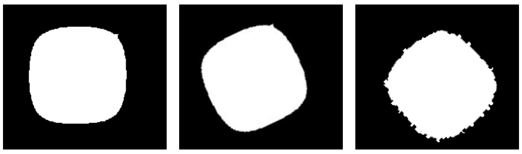


Fig. 3. Shape template (*rsquare*) and two samples: rotated and disturbed

Finally, the training dataset consisted of 2696 samples (673 for each shape), the validation dataset of 748 samples (187 for each shape) and test dataset of 876 samples (219 for each shape). For each sample the first three Hu moments were computed. They formed a feature vector used in training and evaluation of the classifier. The features were subjected to normalization given by: $s_n = (s - f_{min})/(f_{max} - f_{min})$, where $f_{max}$ and $f_{min}$ are respectively the largest and smallest value of the feature (particular Hu moment).

### D. Classifier training and evaluation

Training and evaluation of the classifier was performed on the prepared training, validation and test datasets. The SVM with RBF kernel available in OpenCV was used. The classifier was trained with the `train_auto` function, which performs a multiple cross-validation procedure to select the best SVM parameters.

During the experiments, the impact of number of used Hu moments on the classification performance was evaluated. The accuracy $ACC = TD/(TD + FD) * 100\%$ measure was utilized, where $TD$ – number of correct classifications, $FD$ – number of incorrect classifications.

The obtained results are summarized in Table I. The analysis indicates that using only the first two Hu moments should allow to obtain very good classification accuracy. For this case an experiment on the test dataset was carried out and the following results were achieved: *circle* – 100%, *elipse* – 100%, *mug* – 100%, *rsquare* – 100%.

### E. Results and comments

The proposed shape recognition method achieved almost 100% accuracy. It turned out, that Hu moments are well suited for distinguishing simple geometric shapes. For the considered application the use of the first two invariants resulted in satisfactory performance. The SVM classifier is very easy to use (lots of libraries, available in OpenCV and Matlab) and fast during classification. The solution was designed to allow a simple extension of the feature vector – for example adding other Hu moments or shape descriptors. It is also worth to notice, that the automatic generation of training samples significantly facilitated the evaluation of the approach.

## V. Object colour recognition

Colour, next to shape, is the second feature which is used to identify a particular dish in the system. In this section the method of obtaining colour samples, evaluation of various colour models, as well as the used solution is presented. In the current version of the system colours: blue, orange, green, brown[1] are recognized.

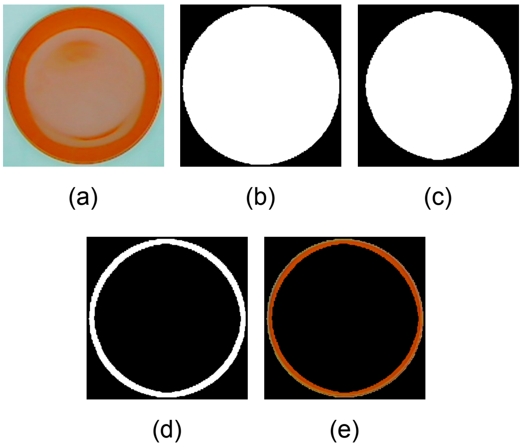


Fig. 4. Edge extraction demo. (a) current image, (b) object mask, (c) mask after erosion, (d) subtraction of (b) and (c), (e) the extracted edge

### A. Obtaining colour samples

The process of obtaining samples used to build the colour model is an element of the system calibration procedure. During its course, the dishes with particulars colour are placed in the workspace. Then object segmentation and edge extraction are performed. Reliable colour information can be only obtained from a narrow edge of the dish, mainly due to presence of meals in its central part.

The edge is extracted using a morphology-based approach. First, the input mask (obtained in the segmentation stage) is subjected several times to erosion with a square structural element of size $3 \times 3$ . Then the input mask and erosion result are subtracted. Finally, the edge with colour samples is obtained. Examples of this procedure is presented in Figure 4.

### B. The analysed colour models

In this subsection the analysed colour spaces and colour models (Gaussian and histogram based) are described.

*a) Colour spaces:* In the experiments several common colour spaces were used:

- RGB (*Red, Green, Blue*) – the basic colour space used in input (cameras) and output (displays) devices,
- YCbCr – a colourspace with a luminance (Y) and two chrominance (Cb, Cr) components. It is used in image and video stream compression (JPEG/MJPEG),
- CIE Lab – colour space similar to YCbCr, but defined to be perceptually uniform i.e. the distance (Euclidean) between two colour samples corresponds with the difference noted by a human.

The HSV (Hue, Saturation, Value) colour space was not analysed, because of the angular coordinates of the H component, which causes some difficulties in models using mean and standard deviation.

[1] This colours were available in the used dish set

TABLE I
CLASSIFICATION RESULTS FOR DIFFERENT FEATURE VECTORS

| | circle | | elipse | | mug | | rsquare | |
|---|---|---|---|---|---|---|---|---|
| Hu moments | Train | Validate | Train | Validate | Train | Validate | Train | Validate |
| 1 | 99.9629 % | 99.8656% | 89.9703 % | 90.5914 % | 89.8217% | 89.7849 % | 99.9629 % | 100 % |
| 1,2 | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| 1,2,3 | 99.9629 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % |

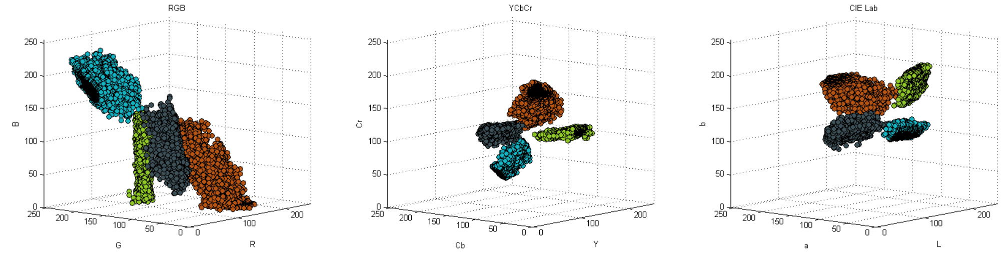


Fig. 5. Colour model samples displayed on a 3D scatter plot. From left: RGB, YCbCr, CIE Lab

TABLE II
STANDARD DEVIATION ANALYSIS FOR USED COLOUR SPACES

| Colour | RGB | YCbCr | CIE Lab |
|---|---|---|---|
| orange | 14352,61 | 2282,61 | 2351,64 |
| blue | 4030,69 | 932,80 | 150,01 |
| green | 3966,19 | 569,31 | 230,09 |
| brown | 11199,39 | 424,27 | 679,77 |

In the first step, the used colour samples were displayed on a 3D plot – Figure 5. On this basis it can be concluded that for the YCbCr and CIE Lab colour spaces the samples are more separated from each other than for RGB. This observation is also confirmed by calculating the standard deviations of the individual components of the samples. In Table II the product of standard deviations for each component is presented. It illustrates the dispersion degree of samples around the mean value. The analysis confirms the earlier observation that YCbCr and CIE Lab have better properties than RGB. It should also be noted that YCbCr and CIE Lab are quite comparable, with a slight indication of the latter. However, due to much more efficient conversion between RGB and YCbCr comapring to RGB and CIE Lab in the further analysis the YCbCr colour space is used.

The obtained colour samples are used to create a representation (model) for classification. In this work models based on Gaussian distribution, 1D histograms and 3D histograms were evaluated. More advanced approaches like Gaussian Mixture Models [14], artificial neural networks [15] or typical machine learning [16] approach were not considered. However, if in future versions of the system, more, especially quite similar, colours should be recognized, than this methods could provide better performance and reliability.

*b) Gaussian distribution model:* For each vector of training samples assigned to a particular colour, mean and standard deviation can be computed. On that basis a Gaussian model can be build. However, there are two options available. In general, it is assumed that the individual colour components are interdependent. Then the probability that a given pixel belongs to a particular model is given by:

$$p(x|\theta) = \frac{1}{(2\pi)^{n/2}\sqrt{(det\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \quad (1)$$

where: $x$ – given pixel, $\theta$ – given colour, $n$ – number of colour components $(3)$, $\Sigma$ – covariance matrix, $\mu$ – mean value vector.

It can be also assumed that the components are independent, which simplifies the equation:

$$p_c(x|\theta) = \frac{1}{\sqrt{(2\pi S)}} e^{-\frac{(x-\mu)^2}{2S}} \quad (2)$$

where: $S$ – standard deviation.

In this case the probability is obtained separately for each component. Therefore, to define it for a pixel, the particular results should be summed up: $p(x|\theta) = p_{c1}(x|\theta) + p_{c2}(x|\theta) + p_{c3}(x|\theta)$.

Thus, for a given pixel the probability of belonging to a particular colour model is obtained. The computations are performed for the whole edge of the dish. At this point, there are two possibilities. First, individual decisions about colour assignment for each pixel can be made. Than the obtained "votes" can be summed up to perform a final classification. Another way involves the summing up of the probabilities and making the decision for the entire image. In the experiments, the second approach was used.

*c) 1D histogram model:* Another common used colour model is a histogram. It can be one-dimensional or three-dimensional (described in the next paragraph). In the first case the given colour is described by three separate histograms, one for each colour component. An important parameter of the method is the number of histogram bins. If the value is less than 256 (assuming that the colour component values are in range 0-255), then nearby values are aggregated.

Colour classification can be performed in two ways. In the first, using the pixel values as an "address" – the corresponding histogram values are obtained and summed. The resulting number is a measure of probability that the pixel has a given colour (histogram normalization is assumed). The numbers, like in the Gaussian model, should be summed up for the whole image. Following this procedure for each colour model (i.e. $4 \times 3$ histograms) and then selecting the maximal value allows to perform the recognition.

In the second approach, a histogram is computed for edge pixels. Then it is compared with the model using a distance measure between two histograms. The most common are correlation, intersection and Bhattacharyya distance.

*d) 3D histogram model:* A three-dimensional histogram can also be used as a colour model. A certain disadvantage of this approach is its high memory complexity – in the default case the histogram has $256^3$ bins. Therefore, aggregation (bin number reduction) is frequently used. The classification can be done in two ways – analogous to those described above for the 1D histogram.

### C. Evaluation and results

When evaluating different colour models, two aspects should be addressed. Firstly, high classification accuracy is required. This means that for each test image the correct result should be obtained. Unfortunately, during research it turned out that this is not a sufficient criterion. The test image database does not include all possible cases that may occur during operation of the system. This applies particularly to potential changes in lightning. Therefore, it is advisable to propose and use an additional measure that will determine which model is better, even in the case when more than one has 100% accuracy on the test dataset.

In this research a classification certainty factor was used. For methods with a probability output the following coefficient can be computed: $cS = \frac{P(m)}{\sum P(c)}$, where: $P(c)$ denotes the probability than an object (i.e. a dish edge) belongs to the colour class $c$ and $P(m)$ is the maximum of this factor. In the ideal case this value should be one, which indicates that all pixels were correctly classified.

For methods based on histogram comparison this coefficient should be modified. In case of correlation, a positive value is obtained when the classification is correct (positive correlation). For other colours the value is usually negative. Therefore, it seems to be eligible to use the absolute value and repeat the above described scheme. When using the histogram intersection, the $cS$ coefficient can be used directly. However, for the Bhattacharyya distance it is necessary to subtract the final values from 1 ($cS = 1 - cS$).

The finally implemented model was selected after a series of experiments. A dataset of 64 images containing dishes in different workspace locations and orientations was used. Additionally, possible lightning conditions were simulated: brightening, darkening and adding gradient (using appropriate gamma correction).

In the first stage, the models with 100% classification accuracy on the test dataset were selected. This were the Gaussian model with dependencies between components and 3D histogram. However, it is worth noting that the remaining models obtained results ranging from 97% to 99%.

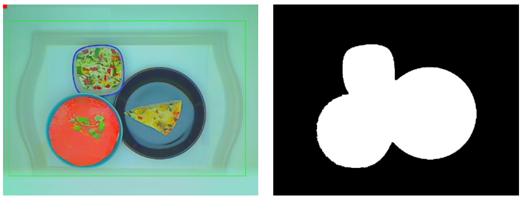


Fig. 6. Example of abut objects. On the left the input image, on the right the corresponding object mask

In the second stage the classification certainty coefficient was evaluated. For the Gaussian model the value $cS = 0.99$ was obtained. Similarly for the probability version of 3D histogram, as well as correlation, intersection and Bhattacharyya histogram distance (using 256, 128 and 64 bins). Therefore, all this approaches should be regarded as very robust.

It should be noted that there was no significant impact of the number of histogram bins on the classification performance. This could be due to the considered colours, which were quite different from each other (separated in the colour component space). Finally, the Gaussian model was used because of its lower memory complexity.

## VI. Close or abut object recognition

In this section a procedure of shape and colour recognition for close or abut objects is described. An example of such a situation is presented in Figure 6.

It should be noted that in the initial vision system specification it was assumed that objects should be placed apart from each other i.e. on the object mask each of them should appear separately. Therefore, when close or abut object are detected, a message for the user is generated: "Please correct the positioning of the dishes". However, to speed-up the service and increase system capabilities an automatic procedure for this case was developed. The issue is quite complex and at least two main problems have to be addressed and solved:

- determining that two or more objects are segmented as one (are close or abut) – distinguishing this situation from typical arrangement of dishes on the tray,
- detecting abut objects with the same colour. In a general case the information about colour can not be used in this procedure, also due to possible influence of food or meal colour on the segmentation results.

### A. Abut objects detection

A situation in which two or more objects are close or abut can be detected by analysing:

- object's shape – there is no recognition by the SVM classifier (i.e. the abut objects form a new shape, which is not similar to any other recognized by the system),
- object's size – it is larger than other recognized by the system.

Thus, if during the shape analysis an object with unrecognised shape and large area appears, it is regarded as a collection of abut or close objects. Here it is assumed, that several small objects can not "form" a large, recognized object (e.g. a big plate formed by several small mugs). Correct recognition in this case seems to be very difficult and would require more sophisticated segmentation algorithms.

### *B. Different concepts of separating objects*

During preliminary research different methods of separating abut objects were analysed. The most straightforward solution is the use of information about edges, because this allows a human to correctly recognize connected objects. Unfortunately, the approach of: detecting edges using Sobel or Canny, thickening them and subtracting from the input mask does not allow to properly separate objects. Mainly due to discontinuities of detected edges and disturbances caused by meals or non-uniform lightning. Furthermore, the obtained object masks are usually smaller then the input ones, which causes difficulties during colour sample extraction.

Another promising solution is the Hough transform, but it has a quite high computational complexity, especially for shapes, whose analytical description requires multiple parameters (circle, ellipse) and it can not be used for all kind of shapes (non-analytical curves).

However, the use of some kind of "knowledge" about the recognized shapes seems to be the key to proper connected object segmentation. This is the result of the observation that knowledge of how the objects look like, in combination with edges and colour information make the task of separating objects quite easy for a human. Therefore, basing on the results published in [17] a distance transform (DT) based solution was used in the proposed system.

### *C. The proposed shape recognition method*

The starting point for the shape recognition algorithm development was the analysis of the distance transform method described in [17]. It uses two of the mentioned observations: object edge analysis and pattern detection (i.e. knowledge).

In the first step of the method, for the input image edges are determined. The resulting binary edge mask is used to compute the so-called distance transform (DT). It is an image in which a pixel value represents the distance (Euclidean, chessboard, Manhattan, quasi-Euclidean) to the nearest edge. An example is presented in Figure 7c.

In the developed algorithm, extracting individual objects requires the use of shape templates in the form of an edge mask and a binary mask. The template is moved across the DT image using a sliding window approach. For each location, the following factor is computed:

$$D_{coef} = \frac{1}{|T|} \sum_{t \in T} DT(t); \tag{3}$$

where: $|T|$ – number of pixels for a given template, $DT(t)$ – distance transform value for pixel $t$ from template $T$. The location with the best match for a given template is characterized by the smallest $D_{coef}$ value.

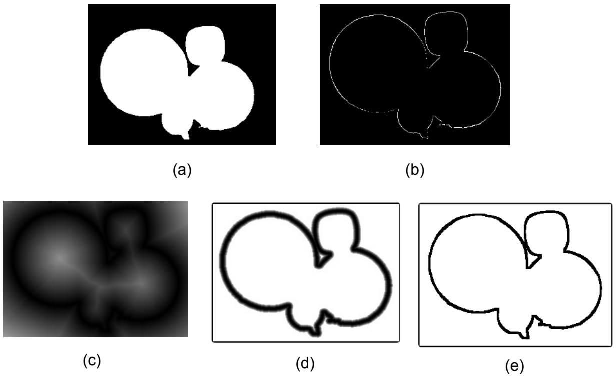


Fig. 7. Distance transform example: (a) input image – object mask, (b) binary edge mask, (c) DT result, (d) modified DT result, (e) binary DT result

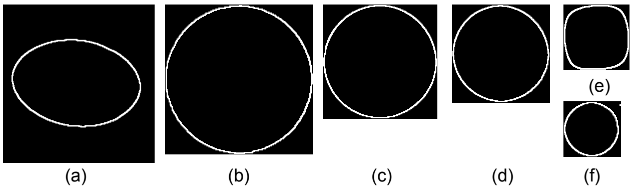


Fig. 8. Used shape templates: (a) – elipse, (b) – big plate, (c) – small plate, (d) – bowl, (e) – small bowl (rounded square), mug (without handle)

In the general case, to ensure proper operation for this type of algorithm, template modifications like translation, scaling and rotation are required. Translation is "built-in" in the sliding window approach. Scaling is not required, because the distance between the camera and object is fixed. However, rotating the template is necessary for selected, asymmetrical shapes like an ellipse.

The used shape templates are presented in Figure 8. It should be noted, that for the *mug* template the shape was simplified to a circle (the handle was omitted). This eliminates the rotation necessity, but does not affect the recognition performance, as there are no objects of similar size and shape in the system.

During preliminary research three variants of the DT based approach were considered: basic DT, modified DT and DT reduced to a binary mask. In the modified version the initial DT image was transformed according to the following formula: $DT(i,j) = DT(i,j)^{\alpha}$ (during tests $\alpha = 2$). Such an approach promotes locations with good template match. An example image is presented in Figure 7d.

A further simplification is the assumption that the DT image has a binary form – in a small neighbourhood of the edges the DT values are set to 0 and in the remaining part to a maximal value. This further enhances the promotion of good-match locations. An example is presented in Figure 7e. It is worth noting, that in this case computing the "DT" image involves only performing a morphological dilation of the edge mask,

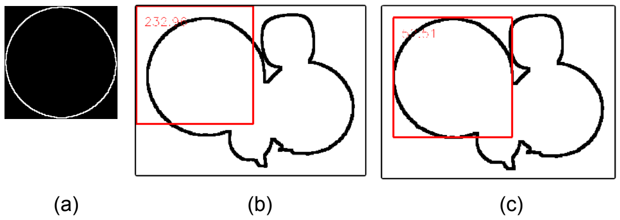


Fig. 9. Template search with sliding window approach: (a) template, (b) initial situation ($D_{coef} = 232.96$), (c) best match ($D_{coef} = 52.51$)

which greatly improves the computational efficiency.

Initial experiments demonstrated a similar performance of the three described DT variants. Therefore, in the final version of the system, the binary DT approach was used, since it is most computationally efficient. However, it is worth considering why such a simple solution obtains comparable results with the "full" DT transform. First, in the described system, the algorithm operates only on the object's outer edge mask, whereas the original solution used all edges of the input image. This significantly limits the number of edges that could potentially affect the shape recognition. Second, the used templates are quite simple, rigid, with well defined size and only slight variation caused by perspective issues. Finally, due to strictly controlled lighting conditions the object's boarder segmentation is very reliable.

### D. Shape recognition procedure for connected objects

During designing the solution, the primary goal was to ensure high recognition reliability. Therefore, an iterative search procedure was proposed. In a single iteration one most probable shape is found.

First, the edge mask and the DT binary image are computed. Next, for each shape template the sliding window procedure is applied and the best location (lowest $D_{coef}$) stored. At the configuration stage, it can be determined if a particular template should be rotated. In the current version of the system, the rotation is only performed for the *ellipse* with 30° step, which results in 5 different templates. For the non-symmetrical template *rsquare* it turned out that no rotation is necessary. Example of the search procedure is presented in Figure 9.

After obtaining the $D_{coef}$ (compare Equation (3)) values for all used shape templates, the minimum is selected. In this way, the most probable shape is detected.

In the next step this shape is removed from the object mask. The new mask is subjected to filtration. This involves the removal of connected components with area smaller than a preset threshold (so-called area open) and morphological processing. An example in shown in Figure 10. The described iteration is repeated until all object are removed.

*a) Template localization improvement:* During experiments it was noted that the obtained, in the above described procedure, shape locations are not very accurate. This has a great impact on the subsequent colour extraction method, particularly for dishes with narrow edge like a bowl. In order to compensate this error a simple template position improvement method was proposed. First, using the initial location estimate, a ROI from the input object mask is extracted. It size is slightly bigger than those of the considered template. Next, in this ROI a sliding window approach is used to find the location, where subtracting the template mask from the object mask results in removing maximal number of pixels. This is then regarded as the improved object location.

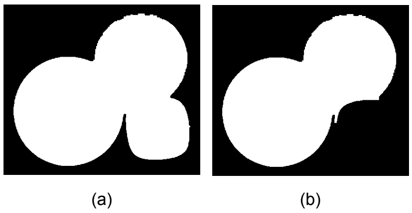


Fig. 10. Shape removal example: (a) – input mask, (b) – mask after shape removal (*rsquare*).

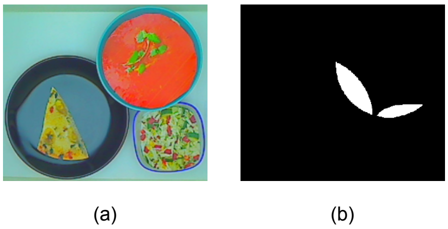


Fig. 11. Example of obscuration area detection: (a) – input image, (b) – obscuration mask.

*b) Analysis of the detected objects:* The result of the above described procedure is the information about all detected objects i.e. their shape and location. For the purpose of colour recognition edge extraction should be performed. However, in case of close or abut objects often some obscuration occurs. For example a bowl could cover a part of a plate. Extracting colour samples in this areas could lead to errors.

Therefore, such regions should be detected and excluded from further analysis. This is done using a so-called obscuration mask – example presented in Figure 11.

In the next step, for the detected objects, edge masks are obtained. This is done in three steps:

- determining the logical AND (intersection) of the objects mask and template mask,
- extraction the edge of this object (using the described prior morphological approach),
- determining the logical AND of the edge mask with the negation of the obscuration mask.

The obtained masks are used in the colour recognition procedure. An example of the described approach is presented in Figure 12. It is worth to notice, that despite some minor errors in determining the localization, the final detection is correct.

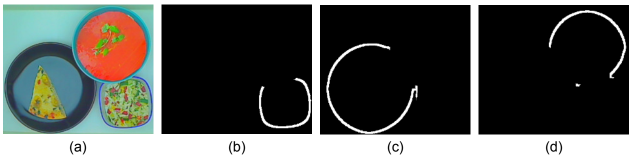


Fig. 12. Example of abut objects splitting. (a) – input image, (b), (c), (d) – edge masks for the detected objects

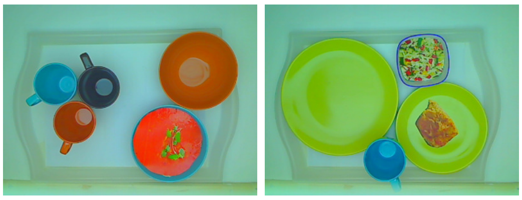


Fig. 13. Examples of test images used in evaluation of the close or abut object recognition procedure

### E. Evaluation and parameter selection

The impact of different parameters on the connected object separation procedure was evaluated. For this purpose 30 test images with different close or abut objects were used. Two examples are presented in Figure 13.

The accuracy of the procedure was evaluated by comparing the returned shape and colour recognition with reference ground truth (more details in Section VII). The following parameters of the method were analysed: template edge thickness, image resolution, sliding window step.

For the first parameter it was observed, that above a certain edge thickness value, the obtained results were incorrect. Finally, for edge extraction an erosion with 7 × 7 square structuring element was used. This results in 3-4 pixels thick edges. In case of reducing the input image resolution, the size of structuring element should also be reduced.

The resolution of the analysed image (object mask) has a great impact on the performance of the described procedure. This is directly related to the sliding window approach and $D_{coef}$ calculation. Therefore, reducing the image size brings significant acceleration to the entire vision system. In the experiments it was determined that changing the size from 640 × 480 to 192 × 144 (30% of the input value) does not result in loss of accuracy – all objects present on test images were correctly recognized.

The sliding window step value is a compromise between template location accuracy and calculation speed. During experiments it was determined, that the value 4 pixel (vertical and horizontal) is suitable for the used resolution of $192 \times 144$.

Additionally, in an experiment is was proved that switching off the template localization improvement procedure leads to errors in colour classification for the bowl shape.

## VII. Vision system evaluation

The developed vision system was subjected to a series of test on the designed stand (compare Section II). Their goal was to evaluate basic functionalities like:

- proper tray position detection,
- empty workspace detection,
- presence of hand in the workspace detection,
- object segmentation,
- shape and colour recognition,
- recognition of close or abut objects.

In addition, elements of the algorithms that require further improvement and modification were identified.

Evaluation of the first three functionalities were conducted on-line i.e. the system behaviour was observed during real-time analysis of the video stream acquired by the camera. The tray position was detected correctly. The only potential problem that occurred was the partial visibility of a marker (the marker was not fully covered by the tray). However, even if it was detected as an object, due to small size it was excluded from further analysis. In addition, it is assumed that the user will co-operate i.e. place the tray as accurate as possible (supported by the feedback provided by the system).

The empty workspace detection module worked correctly in case of proper camera placement and marker position calibration. However, it should be noted that presence of some kind of dirt in the workspace may cause errors – be recognized as an object. Consequently, the background model update mechanism will fail. It is worth mentioning that a frequent background model update is necessary for proper segmentation. During experiments, despite the used LED illuminator, the external lighting changes had impact on the segmentation.

The presence of hand in the workspace detection works correctly. The only exceptional case is when the user wears a sweater or sweatshirt with long sleeves in colour similar to the workspace. In the future, skin-colour regions segmentation could be considered. Whereby, the result interpretation could be quite difficult – some food could have colour similar to skin.

Other functionalities were tested both on-line and off-line on selected test images.

### A. Test images annotation

In order to automate the vision system evaluation and parameter selection process a tool for object annotation was designed. Each object (dish) present on a test image can be described by: location (rectangle inscribed in the object), shape, colour.

For the purpose of frame annotation, a simple GUI application was created. It allows to browse a directory with images and set information about location, shape and colour of different dishes.

### B. Object segmentation evaluation

The used segmentation method worked correctly when the background model was up to date. The only observed problems occurred when cast shadows, present due to specific dish location, caused incorrect operation of the abut object recognition procedure.

TABLE III
SHAPE AND COLOUR OF DISHES USED IN THE EVALUATION.

| Shape name | C 1 | C 2 | C 3 | C 4 |
|---|---|---|---|---|
| `circle_huge` [big plate] | brown | blue | green | |
| `circle_big` [small plate] | brown | blue | green | orange |
| `circle_small` [bowl] | brown | blue | green | orange |
| `mug` [mug with handle] | brown | blue | | orange |
| `rsquare` [small bowl, plate] | —- | | | |
| `elipse` [platter] | —- | | | |

### C. Object shape and colour recognition evaluation

Evaluation of the object shape and colour recognition process was performed by comparing the results returned by the application with prepared manual annotation (ground truth). A series of test scenarios was prepared: from simple one with one dish, to complex involving close or abut objects, as well as presence of additional objects (like napkins or cutlery). The used dishes – they shapes and colours – are summarized in Table III.

For the test procedure 81 images with single dish, 9 with multiple dishes (not close) and 33 with close or abut dishes were prepared. The current version of the system returned an incorrect result in 2 out of 123 cases (accuracy 98%). Both errors were related to a quite significant shape change due to location far from the centre of the camera's optical axis. This could be eliminated by proper camera calibration.

## VIII. Conclusions

In this paper a vision system able to recognize the shape and colour of objects was described. It can be used to automate to process of customer service in a self-service canteen. It is built of four main modules: object segmentation, shape classification, colour classification and recognition of close or abut objects. When designing each module, different possibilities were considered and those with high efficiency and low computational complexity were selected. The final system meets the design constrains i.e. accuracy above 95% and computing time < 1s on an typical PC.

The solution can be further developed in several directions. One of them is the broadly understood relaxation of restrictions. A good example is the use of tray with colour different than the workspace (i.e. not "transparent" for the segmentation). Another one would be the determination of the minimum edge thickness required for proper colour recognition. This would allow to use dishes with only a thin colour boarder instead of "full" colour and therefore use more different colours.

In addition, the use of a lens distortion correction module or another camera should be considered. In a long term perspective, a food recognition module could also be added. An interesting and promising direction is the development of a smart camera vision system to perform the described image processing, analysis and recognition.

## Acknowledgements

The work presented in this paper was supported by the Malopolska Regional Development Agency under the program "Knowledge, Practice, Experience – the key to success in business" in cooperation with Qumak S.A., Kobierzynska 2, 30-363 Krakow

The authors wish to thank Qumak S.A. Krakow company, and particularly Mr. Konrad Pogódź for supporting this research, as well as Dr. Zbigniew Mikrut for help and many valuable comments.

## References

[1] Y. Kawano and K. Yanai, "Real-Time Mobile Food Recognition System," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013. doi: 10.1109/CVPRW.2013.5 pp. 1–7.

[2] F. Zhu, M. Bosch, N. Khanna, C. Boushey, and E. Delp, "Multiple Hypotheses Image Segmentation and Classification With Application to Dietary Assessment," *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 1, pp. 377–388, Jan 2015. doi: 10.1109/JBHI.2014.2304925

[3] V. Bettadapura, E. Thomaz, A. Parnami, G. Abowd, and I. Essa, "Leveraging Context to Support Automated Food Recognition in Restaurants," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, Jan 2015. doi: 10.1109/WACV.2015.83 pp. 580–587.

[4] H. He, F. Kong, and J. Tan, "DietCam: Multi-View Food Recognition Using a Multi-Kernel SVM," *Biomedical and Health Informatics, IEEE Journal of*, vol. PP, no. 99, pp. 1–1, 2015. doi: 10.1109/JBHI.2015.2419251

[5] SRI, "`http://www.sri.com/engage/products-solutions/food-recognition-technology` (last access 03.05.2015)."

[6] BrainCorporation, " `http://www.diginfo.tv/v/12-0145-r-en.php` (last access 03.05.2015)."

[7] OpenCV, "`opencv.org` (last access 03.05.2015)."

[8] QT, "`www.qt.io` (last access 03.05.2015)."

[9] T. Kryjak, "Segmentation of dishes for the purposes of customer service process automation in a self-service canteen (under review)," 2015.

[10] A. Amanatiadis, V. Kaburlasos, A. Gasteratos, and S. Papadakis, "Evaluation of shape descriptors for shape-based image retrieval," *Image Processing, IET*, vol. 5, no. 5, pp. 493–499, August 2011. doi: 10.1049/iet-ipr.2009.0246

[11] M.-K. Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 179–187, February 1962. doi: 10.1109/TIT.1962.1057692

[12] V. Megavannan, B. Agarwal, and R. Venkatesh Babu, "Human action recognition using depth maps," in *Signal Processing and Communications (SPCOM), 2012 International Conference on*, July 2012. doi: 10.1109/SPCOM.2012.6290032 pp. 1–5.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995. doi: 10.1023/A:1022627411411. [Online]. Available: http://dx.doi.org/10.1023/A:1022627411411

[14] Z. Fu and L. Wang, "Color Image Segmentation Using Gaussian Mixture Model and EM Algorithm," in *Multimedia and Signal Processing*, ser. Communications in Computer and Information Science, F. Wang, J. Lei, R. Lau, and J. Zhang, Eds. Springer Berlin Heidelberg, 2012, vol. 346, pp. 61–66. ISBN 978-3-642-35285-0. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-35286-7_9

[15] S. Mikrut, Z. Mikrut, A. Moskal, and E. Pastucha, "Detection and recognition of selected class railway signs," *Image Processing & Communications.*, vol. 19, no. 2-3, p. 83–96, 2015. doi: 10.1515/ipc-2015-0013

[16] M. Querini and G. Italiano, "Color classifiers for 2d color barcodes," in *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, Sept 2013, pp. 611–618.

[17] D. Gavrila, "A bayesian, exemplar-based approach to hierarchical shape matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 8, pp. 1408–1421, Aug 2007. doi: 10.1109/TPAMI.2007.1062

DOI: 10.15439/2015F154
ACSIS, Vol. 5

# Segmentation of Cerebrospinal Fluid from 3D CT Brain Scans Using Modified Fuzzy C-Means Based on Super-Voxels

Abdelkhalek Bakkari
Lodz University of Technology
Insitute of Applied Computer Science
18/22 Stefanowskiego Str.,
90-924 Lodz, Poland
Email: bakkari.abdelkhalek@hotmail.fr

Anna Fabijańska
Lodz University of Technology
Insitute of Applied Computer Science
18/22 Stefanowskiego Str.,
90-924 Lodz, Poland
Email: anna.fabijanska@p.lodz.pl

***Abstract*—In this paper, the problem of segmentation of 3D Computed Tomography (CT) brain datasets is addressed using the fuzzy logic rules. In particular, a new method which combines Fuzzy C-Means clustering and the idea of super-voxels is introduced. Firstly, the method applies the extended Simple Linear Iterative Clustering (SLIC) method to divide image into super-voxels, which are next clustered by Modified Fuzzy C-Means algorithm. The method deals with 3D images and performs fully three dimensional image segmentation. Ten samples are supplied proving that our Modified Fuzzy C-Means (MFCM) together with super-voxels are apt to take into account a large diversity of special domains that appear and which are inappropriate solved adopting classical Fuzzy C-Means approach. The results of applying the introduced method to segmentation of the Cerebro-Spinal Fluid (CSF) from the brain ventricles are presented and discussed.**

## I. Introduction

DIVIDING an image into coherent regions, that are somehow homogeneous and uniform leads to image segmentation.

One of the most popular clustering algorithms used for image segmentation is the Fuzzy C-Means (FCM) approach [1]. Since the method has a lot of advantages (e.g. it provides the best results for overlapped data sets of pixels) it is especially popular when the segmentation of medical images is required [2]. In particular, there was a significant number of attempts to apply FCM clustering for brain segmentation [3], [4], [5]. These works however consider mainly MRI datasets. To the best of our knowledge, there are only few works regarding brain segmentation from CT datasets.

Despite its popularity, the FCM algorithm has also some disadvantages, which limit its application to segmentation of 3D CT medical datasets. The main limitation of the algorithm is in particular its high computational complexity, intensive memory workload and unacceptably long time of computations. These result from the necessity of processing billions of voxels contained within a scan.

Therefore, the most of the existing FCM-based algorithms dedicated to 3D image segmentation are in fact 2.5D approaches. This means that they perform FCM segmentation slice-by-slice and then compose 3D result by combining 2D results obtained from single slices [6], [7].

To overcome the above mentioned limitations of FCM algorithm and make the the method available also in the case of 3D images this paper proposes a solution which incorporates the idea of super-voxels into the Fuzzy C-Means clustering approach. In particular the proposed approach extends the idea of super-pixels into supre-voxels. The Super-voxels are next clustered using FCM algorithm according to statistical features extracted using the co-occurrence matrix.

The proposed approach is next applied to extract the CSF from 3D CT datasets of brain.

The following part of this paper is divided into five sections. Firstly, in Section II, the technical background and a brief review on super-voxels and Fuzzy C-Means techniques is presented. Next, in the Section III datasets used in this paper are characterised. This is followed in Section IV by the description of the introduced approach. The results of the method are presented and discussed in Section V. Finally, Section VI concludes the paper.

## II. Theoretical Background and Related Works

### A. Fuzzy C-Means Clustering

FCM is an algorithm proposed by Bezdek [8] as an alternative for K-means clustering [9]. According to FCM algorithm, each datum point is a part of a cluster whose degree is governed by its membership grade.

What is distinct about FCM is that it divides a collection of $N$ vectors into $c$ fuzzy groups with a cluster centre for each group. It is worth noting that a datum point may be a part of many groups and it gets a membership grade ranging between 0 and 1.

The role of FCM revolves around having $c$ as the number of clusters, $c_i$ as the cluster centre of fuzzy group $i$ and the parameter $m$ as the weighting indicator for every fuzzy integrating group. Through optimizing the function of FCM, the fuzzy subdivision can be conducted.

978-83-60810-66-8/$25.00©2015, IEEE

The membership $J_{FCM}(U,V)$ and the cluster centres are determined according to the following equation [9]:

$$J_{FCM}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d^2(x_k, v_j) \tag{1}$$

where: $u_{ik}$ is a matrix of size $(c \times d)$, $d = \| x_j, v_i \|$ is the Euclidean distance between the centroid $v_i$ and each pixel $x_j$, $U = u_{ik}$ represents the matrix of fuzzy partition, $V = v_1, v_2, ..., v_n$ is are class centres, $m$ is the fuzzy factor with $m > 1$ and $c$ is the class number.

Because of its advantages, FCM approach has been widely applied in segmentation of medical images. Application of the method to brain segmentation is especially popular. Additionally, numerous improvements to the FCM method have been proposed.

One of the major proposals in the medical image segmentation is concerning the adoption of the spatial distance into the clustering based segmentation as initiated by Tolias and Panas [10], [11], [12]. Furthermore, Liew [13] proposed an automatic segmentation of 3D dimensional Magnetic Resonance Imaging (MRI) brain images. They used a local spatial distance into the FCM algorithm adopting a new dissimilarity index instead of the standard Euclidean distance. In addition, they created a cluster prototype with variation of 3D multiplicative bias field [14].

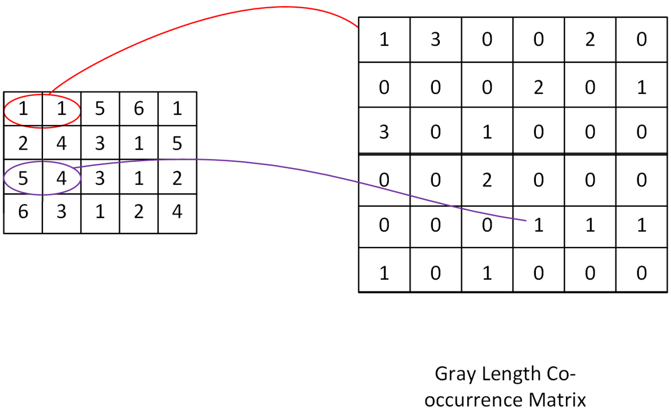
In the same context, there is an approach which provides the extraction of some features. It can incorporate the intensity information for the voxels neighbours [15]. Moreover, the fuzzy logic supply to segment 3D image under the consideration of the following three information: position, boundary and intensity knowledge [16]. This method aims to extract the three portions of the brain such as the left cerebral hemisphere, right cerebral hemisphere, cerebellum and brain stem. One popular technique involves adopting Fuzzy C-Means stand on the local spatial continuity [14]. It takes into account the voxel neighbour information and the intensity variation.

Regarding to other methods, [17] adopted a new method for segmentation of 3D datasets, based on Fuzzy C-Means. This approach is applied only to the three views; sagital, coronal and axial. Furthermore, the extraction of CSF is reported in [6]. This approach is based on the fuzzy inference rules. It is focused on the information obtained by the fuzzy information granulation.

The use of the Fuzzy C-Means may present some constraints, especially, when applied to brain segmentation. Brain image segmentation from CT scans faces the numerous numbers of challenges due to the characteristics of the images: poor image contrast, high-level speckle noise, weakly defined boundaries and boundary gaps. The traditional Fuzzy C-Means method is often unable to perform adequately on these images complex extension. Therefore, to overcome the above drawbacks, this paper proposes a new method based on the second statistic feature by the use of the co-occurrence matrix.

## B. 3D Co-occurrence Matrix

The co-occurrence matrix stores information about the occurrence of couples of pixels in the image. It takes into consideration the neighbouring pixels and the spatial relationship of pixels. In the case of intensity images, the co-occurrence matrix is also called grey-level spatial dependence matrix or grey level co-occyrence matrix (GLCM). This matrix is determined based on pixel intensity values. The idea of the GLCM determination is explained in figure 1. In particular, the (5×4) matrix shown on the left represents image with pixel intensities represented by numbers, while the (6×6) matrix on the right represents the corresponding GLCM.



| | | | | |
|---|---|---|---|---|
| 1 | 1 | 5 | 6 | 1 |
| 2 | 4 | 3 | 1 | 5 |
| 5 | 4 | 3 | 1 | 2 |
| 6 | 3 | 1 | 2 | 4 |

| | | | | | |
|---|---|---|---|---|---|
| 1 | 3 | 0 | 0 | 2 | 0 |
| 0 | 0 | 0 | 2 | 0 | 1 |
| 3 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 2 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 0 | 0 | 0 |

Fig. 1: The idea of the gray level co-occurrence matrix determination: the input image (on the right) and the corresponding co-occurrence matrice (on the left).

The GLCM was introduced to describe two dimensional images. However, in the literature, there are approaches which concentrate on using the 3D co-occurrence matrix which involves a good description of the image information. For the first time GLCM was adopted to extract important features using the texture, called the second order statistics [18]. These include the homogeneity, the angular second moment, the entropy and the contrast. After that, the 2D Haralick texture feature was applied to medical images and extended to 3D domain [19]. Furthermore, it is adopted for the hyperspectral imagery as an image cube [18].

In the same context, the self organizing map (SOM) is a kind of artificial neural network founded on competing as well as unsupervised learning. The combination of SOM and FCM with the GLCM is assumed to extract the first and the second statistical features preceded by a segmentation of the input image [19]. The main inconvenient of 3D image segmentation performed in this way is that it involves only the 2D images, performing image segmentation slice-by-slice [6] [7].

## C. Super-voxels

A super-pixel can be defined as a set of connected pixels that posses similar attributes. Most commonly, pixel intensity or colour is regarded [20]. Figure 2 shows the super-pixel principle. The top image represents an input 2D image, while the bottom image is an output image after image division into super-pixels.

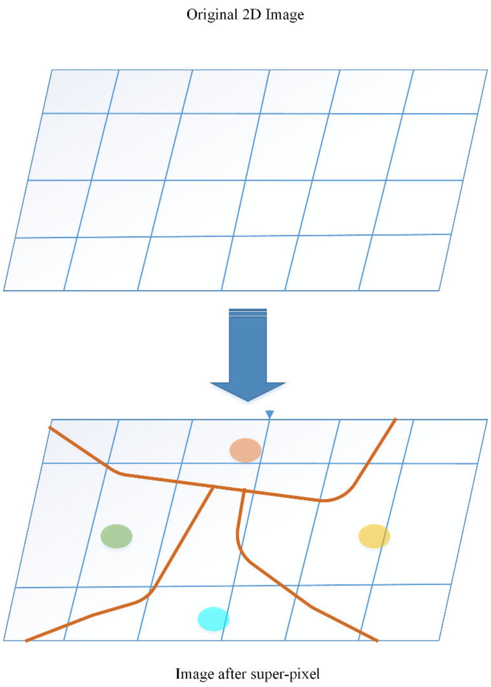


Fig. 2: A model of image segmentation using super-pixels algorithm.

The state-of-the-art is replete with approaches for image division into super-pixels, which have been proposed for 2D images. These include SLIC approach [21], [20] which uses linear iterative clustering, Turbo-pixels [22] which use curve evolution, NC super-pixels [23] based on the normalized graph cuts, FH super-pixels which use greedy segmentation proposed in [24] or methods based on energy minimisation framework as proposed in [25].

There are also some approaches which extend the idea of super-pixels into 3D images. For example, the approach proposed in [26] aims to divide a three dimensional image into blocks. Other method uses the super-voxel technique for processing a Voxel Of Interest (VOI) [27] instead of the whole voxels of the image. In [28], the convexity is considered as a metric for the super-voxel extraction. Moreover, a method of super-voxel combined with a clustering approach has been reported in [29] in order to extract statistical features. The same problem is treated in [30]. It aims to determine the region of interest adopting shift followed by the super-pixel algorithm. Its advantage is that it is applicable not only for 3 dimensions, but also from 1 to $N$ dimensions.

The method proposed in this paper extends Zero version of Simple Linear Iterative Clustering (SLICO) approach into three dimensions. SLICO method is widely used in the literature. It aims to divide the input image into a super-pixels, that commonly have a uniform and compact shape with better boundary stickiness. In this paper, we adopted the SLICO technique because it is fast to compute, memory efficient, and simple to use. The memory efficiency and low computational cost is especially important when segmentation of 3D images is considered.

Figure 3 shows the application of SLICO approach to a sample CT brain slice. In particular, the figure (3-a) represents the input 2D image, while the figure (3-b) shows the result of SLICO method.

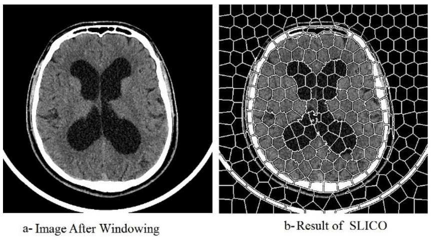


Fig. 3: SLICO approach for 2D image super-pixel segmentation: a- Image after windowing, b- Result of SLICO algorithm.

To the best of our knowledge the combination between the FCM and super-pixel is concerned only in [31] and [32]. The first work takes into account the super-pixel technique as a clustering objects in spite of the classical super-pixel. The second approach inspects a different strategy. An additional feature of segmentation is added (eg. extraction of CSF, white matter and gray matter). Both methods however are dedicated only to segmentation of 2D brain images.

## III. Input Data

Ten CT brain scans in Digital Imaging and Communications in Medicine (DICOM ) format were used in this paper. All of them present brain with the ventricular system enlarged due to the hydrocephalus. The images were adopted in order to extract the CSF contained within the brain ventricles, to test the proposed approach and to evaluate its performance in comparison to other methods. The average number of slice in the dataset was 215. Each slice had the spatial resolution of $512 \times 512$, the bit resolution of 12 and the slice thickness equal to 1.5000. In addition, the spacing between slices has $0.7500mm$.

The figure 4 shows 2D the selected slices composing a sample 3D CT brain scan. The slices are after intensity windowing.

## IV. Proposed Method

The Modified Fuzzy C-Means algorithm based on super-voxels is our proposed approach. The main idea behind this approach is to perform image division into super-voxels using the extended SLICO approach and then cluster the resulting regions using FCM algorithm.

The proposed method contains three main steps, namely: image pre-processing followed by an application of the super-voxels technique and finished by using the modified Fuzzy C-Means algorithm.

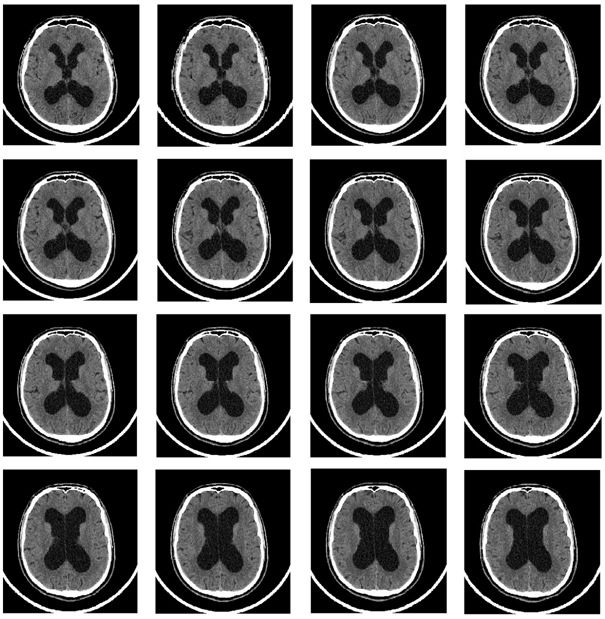


Fig. 4: Sample CT brain slices after pre-processing step.

The block diagram of the introduced approach is shown in figure 5. The details regarding each step of the introduced approach are given in the following subsections.

### A. Image pre-processing

The pre-processing is the fundamental task for the introduced approach. It is mainly operated by window and contrast adjustment. The modification of the window as well as the contrast values depends on the input image and the region of interest. The information about the desired window is usually given in the DICOM header.

In the preprocessing step, firstly image intensities are linearly transformed according to the rescale intercept and slope as described in equation ( 2).

$$New_{HU} = (RPV \times R_S) + R_I \qquad (2)$$

where: RPV is the raw (original) pixel value, $R_S$ is the rescale slope and $R_I$ represents the rescale intercept.

After applying the rescale/intercept transformation, image windowing is performed. Generally, crucial brain regions such as the cerebrospinal fluid, the white matter and the grey matter drop within the interval from 0 to 150 under Hounsfields Units (HU). Accordingly, the windowing procedure has to be achieved to highlight intensities within the region of interest. In particular, the original pixel values declined over the range are threshold to black or white. To obtain this, the window centre $W_C$ is set to 40, while, the window level $W_L$ is set to 80. Finally, the images converted to 8-bit grey scale format, where intensities range from 0 (black) to 255 (white). This procedure is described by equation ( 3).

$$GrayImage = 255\frac{W_{Max} - W_{Min}}{New_{HU} - W_{Min}} \qquad (3)$$

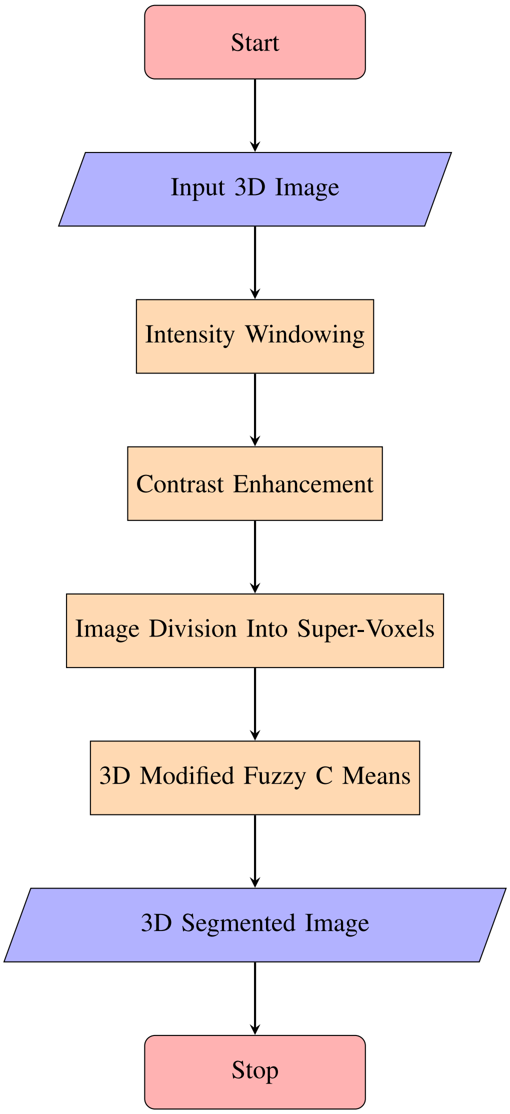


Fig. 5: The block diagram of the proposed method.

where: $W_{Max} = W_C + W_L/2$ and $W_{Min} = W_C - W_L/2$.

Figure 6 illustrates the original image and the image after the windowing. In particular, sub-figures 6 a and 6 c show sample images before windowing, while sub-figures 6 b and 6 d correspond to images after this procedure.

After intensity transformation, intensities corresponding with brain region and CSF region are highlighted.

### B. Super-voxels algorithm

In this paper, the SLICO super-pixels algorithm is extended to be adequate with three dimensional images and greyscale super-voxels. In order to do this, the initialization of the cluster is required. Thus, we called SLICO technique a super-pixel clustering. The second step is to calculate the spatial distance between the cluster centre and each voxel in the window of size $(7 \times 7 \times 7)$. Eventually, the new cluster centres have to be updated relatively to the spatial distance.

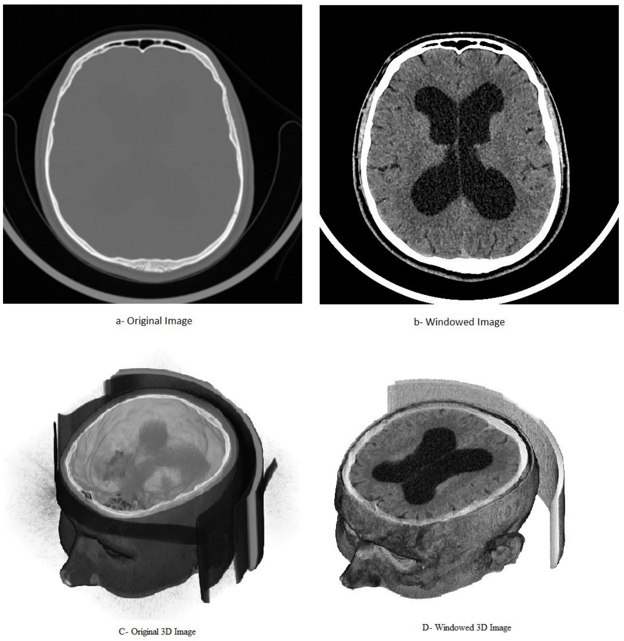


Fig. 6: Image windowing: a- 2D original image, b- 2D windowed image, c- 3D original image, d- 3D windowed image.

The proposed method presents several advantages compared to existing ones. Simple linear iterative clustering (SLICO) is an adaptation of k-means for super-voxels generation, with two important distinctions:

- The number of distance calculations in the optimization is greatly decreased. This reduce is due to the limiting the search space to a region proportional to the super-voxels size. This minimizes the complexity to be linear in the number of pixels $N$ and also independent of the number of super-voxels $k$.
- While simultaneously a providing control over the size and compactness of the super-voxels, a weighted distance measure combines colour and spatial proximity. SLICO is similar to the approach described in [17]. This latter is used as a pre-processing step for depth estimation, which was not fully explored in the context of super-voxel generation.

The algorithm 1 can be described step by step as follows [33]:

1) Use the vector $[lx, ax, bx, cx, xk, yk, zk]$ to represent each voxel, $[lx, ax, bx, cx]$ for the voxel colour vector (in the case of greyscale image, this vector = [0,0,1] ), $[xk, yk, zk]$ is the voxel position, then the voxel of our colour similarity and distance to produce super-voxels. The grid interval is $(S = \sqrt[N/K]{2})$; initialize the $K$ clusters centres.
2) In the area of where $(n*n*n)$, find the minimum gradient position $I(xk, yk, zk)$ is $(xk, yk, zk)$ position of the voxel $[lx, ax, bx]$ vector, $||.||$ is the norm.
3) Perform the following steps to know the cycle $E$ (residual error) $< a$ threshold:
   Each cluster centre $C_k$ is designed for $(7S \times 7S \times 7S)$ voxels area. The most appropriate voxel allocated to this cluster is for the greater value of $m$ .
4) After the clustering is complete, recalculate the cluster centres and $E$;
5) Connect similar regions.

Algorithm 1: Super-Voxels Algorithm (SLICO)

```
procedure
    /*Initialisation*/
    Initialize cluster coordinates $C_i = [lx, ax, bx, cx, xk, yk, zk]^T$
    Sampling voxels at regular grid steps S
    Move cluster centres to the lowest gradient position in a (7*7*7) neighbourhood
    for each voxel i do
     label l(i)=-1
    Set distance $d(i) = \inf$
    end for
    repeat
     /*Assignment*/
    for each cluster center $C_k$ do
    for each voxel in a $(7S*7S*7S)$ region around $C_k$ do
    Compute the distance D between $C_k$ and i

    if $D < d(i)$ then
        set $d(i) = D$
        set $l(i) = k$

    end if
    end for
    end for
    /*Update*/
    Compute new cluster centers
    Compute residual error E
    D1 distance between previous centers
    recomputed centers
    Until $E \leftarrow$ threshold
    Enforce Connectivity
end procedure
```

The adopted technique is described by Algorithm 1.

The improved SLICO technique adopts the typical compactness parameters (chosen as an initialization) applied to all super-voxels in the 3D image. In the case of a high smoothness in some regions with a high texture for the others, the SLICO provides a smoothness repeatedly super-voxels in the weak regions and extremely intermittent super-voxels in the textured regions.

SLICO is an improvement of SLIC proposed in [33] to solve that problem effectively. The compactness parameter must not be initialized by the user. SLICO precisely selects the compactness parameter adequate for each super-voxel separately. This achieves ordinary shaped super-voxels for both textured and non textured parts in the image.

The figure 7 represents the application of SLICO algorithm after its extension to three dimensions as proposed in this paper. The figures 7-a and 7-c show the original 2D and 3D images after windowing, while, the figures 7-b and 7-d show the 3D image after division into SLICO super-voxels.

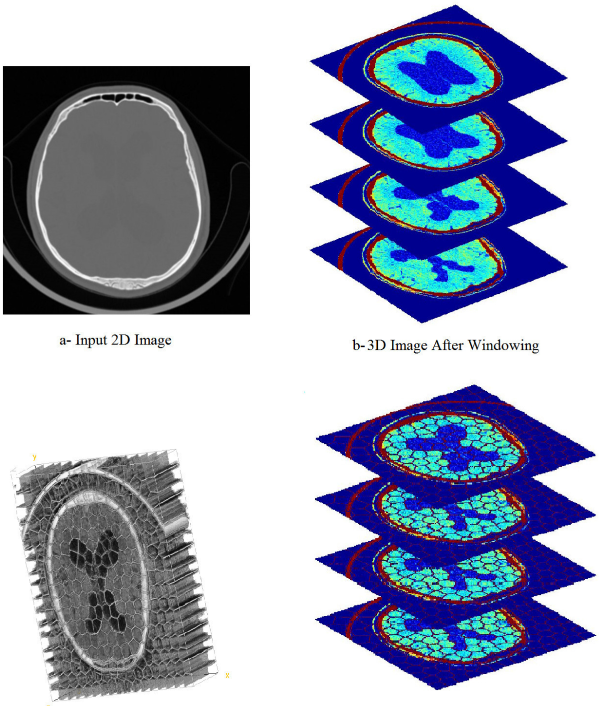

a- Input 2D Image
b-3D Image After Windowing
c-After SLICO Super-voxels
d- 3D Image After SLICO

Fig. 7: Results of the extended SLICO algorithm applied to 3D Image: a- input 2D image, b-after SLICO super-voxel, c-3D image after windowing, d-3D image after SLICO.

### C. Modified Fuzzy C-Means

The modified Fuzzy C-Means is a combination between a creation of the co-occurrence matrix and the standard Fuzzy C-Means algorithm.

Conventionally, many approaches of 2D image segmentation take into account the two specified parts: the segmentation technique and the representation system. Throughout this structure, the proposed approach is defined as an amendment of the Fuzzy C-Means algorithm, established on a co-occurrence matrix [34]. The Fuzzy C-Means algorithm can be adopted to compute the membership degree for each super-voxel. However, the FCM algorithm involved only the grey level and does not include the super-voxels spatial information with consideration of each other. For this reason, we determined the statistical attributes of the image after applying the super-voxels technique. This combination may help to impress this inconvenient.

The steps of 3D Modified Fuzzy C-Means method are specified as follows [34]:

1) Choose our input image after super-voxel technique.
2) Set the size of the sliding window.
3) Calculate of the co-occurrence matrix for the sake of extracting a peculiar image.
4) Perform the standard Fuzzy C-Means algorithm which is applied to the attribute image to attain the final segmented one.
5) Adopt the standard Fuzzy C-Means technique in order to extract the region of interest (CSF).

*1) Spatial feature correlation method:* In this paper, we used the co-occurrence matrix [35] as it is related to the presence of a voxel pair from the given image $I$. The co-occurrence matrix is made of important data that restore the class bias of $I$. Consequently, the proposed co-occurrence matrix performs a major role in image dividing.

The co-occurrence matrix, known also as spatial feature correlation method, describes the occurrence of voxel pairs in the distance denoted $d$ in a certain direction in accordance to the following equation:

$$Cooc(i,j,k,R) = card\left\{ \begin{array}{r} ((x,y,z),(x',y',z') \in D, checking\ R(d,\theta) \\ I(x,y,z) = i; I(x',y',z') = j \end{array} \right. \tag{4}$$

where $card(A)$ denotes the cardinality of the subset $A$, *checking* $R(d,\theta)$ expresses the relation between two voxels, $d$ is the Euclidean distance between two voxels. $\theta$ is the angle that describes the orientation of the two voxels among the horizontal direction. This angle can be equal to $0°$ or $45°$ or $90°$ or $135°$. Every element of the co-occurrence matrix $Cooc(i,j,k,R)$ conforms to the number of voxel pairs $(i,j,k)$. It expresses the number of occurrence of a voxel which has a gray-level value $j$. Therefore, this occurrence have to be related to a horizontal adjacency. Subsequently, the evaluation of the regions within the image is made through the use of the co-occurrence matrix. Therefore, the removal of the second statistical features will be simple. These features are the mean $Me$ (Eqn. 5), the variance $V$ (Eqn. 6), the Skinewski $Sk$ (Eqn. 7) and the Kurtosis $Ku$ (Eqn. 8).

$$Me = \frac{1}{M \times N \times S} \sum_{k=1}^{s} \sum_{i=-\frac{w-1}{2}}^{\frac{w-1}{2}} \sum_{j=-\frac{w-1}{2}}^{\frac{w-1}{2}} I(n+i, r+j, t+k) \tag{5}$$

$$V = \frac{1}{M \times N \times S} \sum_{k=1}^{s} \sum_{i=-\frac{w-1}{2}}^{\frac{w-1}{2}} \sum_{j=-\frac{w-1}{2}}^{\frac{w-1}{2}} (I(n+i, r+j, t+k) - Me)^2 \tag{6}$$

$$Sk = \frac{1}{M \times N \times S} \sum_{k=1}^{s} \sum_{i=-\frac{w-1}{2}}^{\frac{w-1}{2}} \sum_{j=-\frac{w-1}{2}}^{\frac{w-1}{2}} (I(n+i, r+j, t+k) - Me)^3 \tag{7}$$

$$Ku = \frac{1}{M \times N \times S} \sum_{k=1}^{s} \sum_{i=-\frac{w-1}{2}}^{\frac{w-1}{2}} \sum_{j=-\frac{w-1}{2}}^{\frac{w-1}{2}} (I(n+i, r+j, t+k) - Me)^4 \tag{8}$$

where $I$ denotes the image and $(M \times N \times S)$ represents the size of $I$, $(w \times w \times w)$ is the sliding windows as shown in the figure 9. The four features are obtained from the windows size $(7 \times 7 \times 7)$, described by the figure 8.

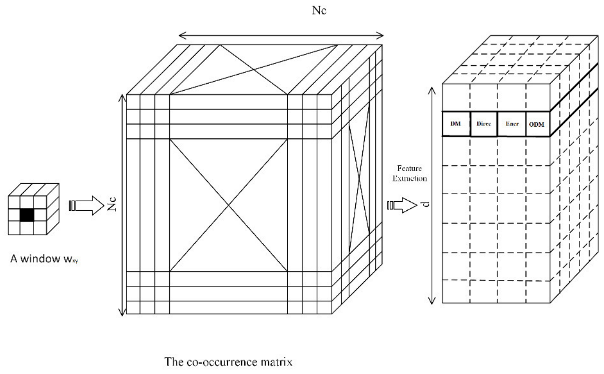


Fig. 8: A sliding window for statistical features extraction.

The window is centered at the voxels $(n, r, t)$ in order to extract a centred window around every voxels. Hence, in the figure 8, the vector that contains the statistical features (DM, Direc, Ener, ODM) is classified adopting the C-Means algorithm into $c$ classes.

The segmentation based on the C-Means algorithm divides the image in $c$ regions (classes). The dimensional scanning plan of an image is implemented voxel by voxel.

*2) Standard Fuzzy C-Means Algorithm:* After applying the SLICO super-voxels algorithm for dividing the image into super-voxels, we extracted the attribute image (Means, Variance, Skinewski and Kurtosis). The third step of the proposed method is to adopt the Fuzzy C-Means algorithm to each super-voxels of the obtained image.

The Fuzzy C-Means aims to minimize the weighted within class sum of squared error objective function [3] :

$$J_{FCM}(U,V) = \sum_{l=1}^{s} \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{lik})^m \|x_k - v_i\|^2 \tag{9}$$

where $x = [x_1, x_2, x_3, \ldots, x_n]^T$ is the data set, $U$ is composed by memberships $u_{ilk}$ of $k^{th}$ bit in the $i^{th}$ class and $m$ is the fuzzy factor with $m > 1$.

The proposed solution of the objective function can be attained using an iterative process, that is performed as follows:

1) Input of the original image which has a size $(M \times N \times D)$,
2) Initialize the parameters: the fuzzifier and the centres of classes,
3) Initialize the partition matrix $U^{(0)}$ based on random variables between 0 and 1,
4) Calculate of the Euclidean distance referring to the following equation :

$$d(x,y,z) = \sqrt{(z_2 - z_1)^2 + (y_2 - y_1)^2 + (x_2 - x_1)^2}, \tag{10}$$

where: $(x_1, y_1, z_1)$ are the coordinates of the first voxel, while $(x_2, y_2, z_2)$ are the coordinates of the second voxel.

5) Update of the prototype using the equation as follows:

$$b_i = \frac{\sum_{k=1}^{n} U_{ik}^m \times x_k}{\sum_{k=1}^{n} U_{ik}^m} \tag{11}$$

6) Calculate the partition matrix $U^{(k)}$ according to equation:

$$U_{lij} = \left[ \sum_{l=1}^{s} \sum_{k=1}^{c} \left( \frac{d^2(x_j, b_i, z_l)}{d^2(x_j, b_k, z_l)} \right)^{\frac{2}{(m-1)}} \right]^{-1} \tag{12}$$

7) Convergence test: repetition of the 4, 5 and 6 steps described by the following equation:

$$||U^{(k+1)} - U^{(k)}|| < \varepsilon \tag{13}$$

where $\varepsilon$ is the tolerance. It converges to zero.

## V. Experimental results

This section presents the results of applying the introduced approach to 10 sample CT images of brain. In particular, a region of CSF is extracted by the proposed method.The sub-figure (10-a) is the windowed 3D image, the sub-figure (10-b) is the result after applying the SLICO algorithm. The sub-figure (10-c) shows the image after applying the SLICO algorithm combined with the mathematical morphology (10-d) is the image after Modified Fuzzy C-Means algorithm. The Figure 11 presents an the results shown in 3D. While, the figure 12 represents the sample slice overlayed.

The proposed approaches were compared in terms of the accuracy and the execution time with the following approaches: Modified Fuzzy-C Means, the combination between SLIC and MFCM and the combination between SLICO and MFCM.

The results of accuracy comparison (in percentage) are given in the Table I. It was measured as follows :

$$Accuracy = \frac{\text{Number of correctley classified pixels}}{\text{Total number of pixels}} \times 100\% \tag{14}$$

The first column shows the case ID. This is followed by the accuracy of the MFCM. The third column represents the accuracy of SLIC combined with the MFCM and the last column shows the accuracy percentage of our proposed method (SLICO+MFCM). While Table II presents comparison of execution time between the Modified Fuzzy C Means, the combination between the Modified Fuzzy C-Means, the Modified Fuzzy C Means combined with the SLIC super-voxels algorithm and the the combination between SLICO and MFCM.

The execution time is given in the table II. Tests were

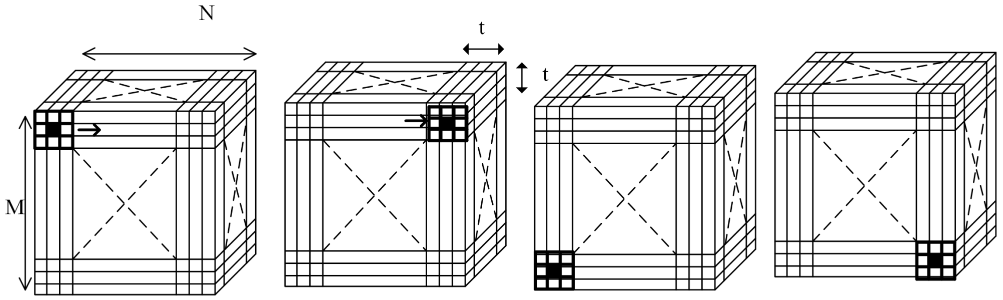


Fig. 9: The adaptive sliding windows from the left to the right and from the top to the bottom on an (M*N*D) Image.

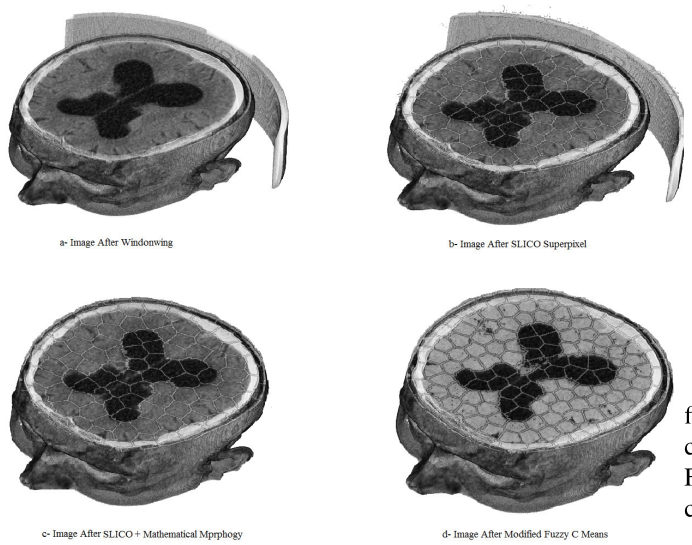

a- Image After Windonwing
b- Image After SLICO Superpixel
c- Image After SLICO + Mathematical Mprphogy
d- Image After Modified Fuzzy C Means

Fig. 10: Image segmentation result: a- Image after windowing, b- Image after SLICO supervoxels, c- After SLICO + mathematical morphology, d- After Modified Fuzzy C-Means.

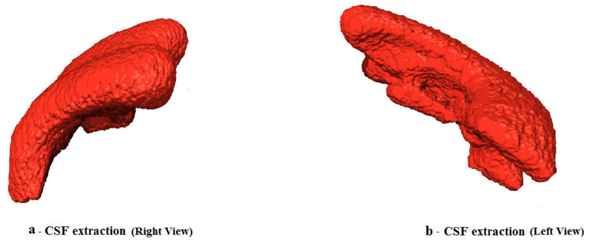

a - CSF extraction (Right View)
b - CSF extraction (Left View)

Fig. 11: SLICO combined with the MFCM Results in 3D.

performed on a PC computer with an Intel Core (TM) i5-3450 CPU 3.10 GHz, a 32 GB of RAM and a CUDA for Graphic Processing Unit using Graphic Parallel Unit Toolbox under Matlab 2013a version.

We can interpret the figure 12 and 13 that the two classes are correctly extracted for 2D and 3D images. The first class is the CSF region and the second one if for the rest of the image.

In our paper, we are interested in the CSF region. So, the figure 13 takes into account the region of Interest (CSF). It is clear that, our SLICO technique combined with the Modified Fuzzy C Means is more efficient than the SLIC technique combined with the Modified Fuzzy C-Means.

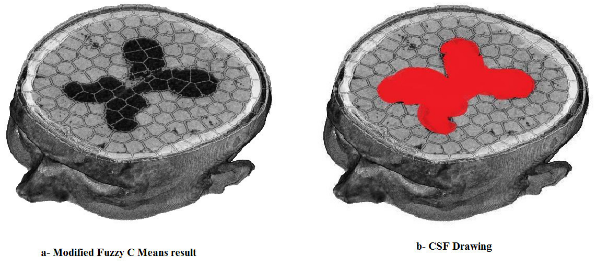


Fig. 12: Sample slice overlayed.

From the Table I, we can say that the 3D Modified Fuzzy C Means takes much time than the ameliorated version based on the GPU. Otherwise, the MFCM combined with SLICO technique is faster than the SLIC technique combined with the MFCM algorithm. The average time of the combination between SLICO technique and Modified Fuzzy C-Means is about 20.94 s. Althought, for The average time of the combination between SLIC technique and Modified Fuzzy C-Means is about 29.10.

Furthermore, the Table II demonstrates that the ameliorated MFCM combined with the SLICO is more accurate than the combination between the MFCM algorithm and SLIC supervoxels technique.

The extracted CSF from three dimensional image is showed in the figure 13. As can be seen in this figure, the visualization of the (VOI)Volume Of Interest using our prposed method (MFCM+SLICO) is more consistent than the Modified Fuzzy C-Means combined with the SLIC technique.

## VI. Conclusion

The segmentation method proposed in this article, is a novel region segmentation method based on the super-voxel technique and the modified Fuzzy C-Means algorithm while the Cerebro-Spinal Fluid (CSF) part has a good consistency. This method consists of three steps. In the first step, the intensity windowing and contrast enhancement are applied

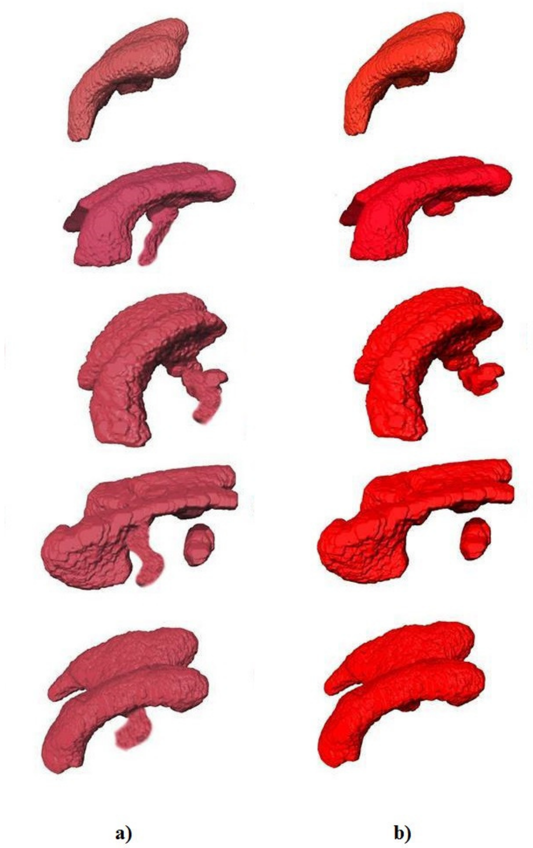


Fig. 13: CSF Visualization: a) SLIC combined with MFCM, b) SLICO combined with MFCM results.

TABLE I: Comparison the accuracy between original MFCM, SLIC algorithm combined with MFCM and SLICO algorithm combined with MFCM.

| Case ID | Accuracy MFCM (%) | Accuracy SLIC+MFCM (%) | Accuracy SLICO+MFCM (%) |
|---|---|---|---|
| 01 | 93,12 | 96,13 | 97,50 |
| 02 | 92,60 | 93,15 | 98,27 |
| 03 | 88,54 | 90,04 | 96,87 |
| 04 | 90,45 | 95,68 | 97,31 |
| 05 | 75,21 | 80,05 | 82,86 |
| 06 | 80,05 | 82,78 | 89,17 |
| 07 | 91,80 | 93,43 | 95,01 |
| 08 | 82,34 | 86,15 | 91,33 |
| 09 | 76,23 | 89,90 | 90,17 |
| 10 | 78,14 | 81,94 | 85,35 |

TABLE II: Comparison the execution time between original MFCM, SLIC algorithm combined with MFCM and SLICO algorithm combined with MFCM.

| Case ID | Time MFCM (s) | Time SLICO+MFCM (s) | Time SLIC+MFCM (s) |
|---|---|---|---|
| 01 | 1120,56 | 11,13 | 11,50 |
| 02 | 1240,60 | 14,30 | 14,55 |
| 03 | 1224,34 | 12,84 | 13,40 |
| 04 | 1149,57 | 11,68 | 12,16 |
| 05 | 1180,68 | 10,05 | 10,76 |
| 06 | 1202,85 | 12,78 | 13,23 |
| 07 | 1136,90 | 13,43 | 15,02 |
| 08 | 1210,42 | 13,15 | 13,44 |
| 09 | 1119,35 | 11,90 | 12,27 |
| 10 | 1127,14 | 11,94 | 12,68 |

to the input 3D CT image. In the second step, we adopted an image division into super-voxels. Then, a segmentation modified Fuzzy C-Means approach is applied in order to segment the image into two classes. Considerable evaluation results have demonstrated great potential on our new approach.

Regarding to the main objective of this research paper, there is no exist method suggested the combination of fuzzy logic rules with a super-voxel technique. Furthermore, the proposed method considers the neighbouring membership degree among the voxels of the images during the determination of a final classification which can be unable with traditional segmentation methods.

## Acknowledgment

This work is financially funded by the European Union under the Erasmus Mundus project. This research was funded also by the Ministry of Science and Higher Education of Poland from founds for science in years 2013-2015 in a framework of Iuventus Plus programme (project no. IP 2012 011272).

## References

[1] S. M. S. Szilágyi, László and Z. Benyó, "A modified fuzzy c-means algorithm for mr brain image segmentation," in *Image Analysis and Recognition*, vol. 4633, August 2007, pp. 866–877.

[2] S. C. L. Lee, Lay Khoon and W. J. Thong, "A review of image segmentation methodologies in medical image." *Advanced Computer and Communication Engineering Technology*.

[3] B. Irving, A. Cifor, B. W. Papież, J. Franklin, E. M. Anderson, M. Brady, and J. A. Schnabel, "Automated colorectal tumour segmentation in dce-mri using supervoxel neighbourhood contrast characteristics," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2014*. Springer International Publishing, 2014, pp. 609–616.

[4] P. B. Kanade and P. P. Gumaste., "Brain tumor detection using mri images." *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, vol. 03, pp. 146–150, 02 2015.

[5] S. K. Adhikari, J. K. Sing, D. K. Basu, and M. Nasipuri, "A spatial fuzzy c-means algorithm with application to mri image segmentation," in *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*. IEEE, 2015.

[6] Y. Hata, S. Kobashi, S. Hirano, H. Kitagaki, and E. Mori, "Automated segmentation of human brain mr images aided by fuzzy information granulation and fuzzy inference." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 30, no. 02, pp. 381–395, 02 2000.

[7] S. Kobashi, Y. Fujiki, M. Matsui, N. Inoue, K. Kondo, Y. Hata, and T. Sawada, "Interactive segmentation of the cerebral lobes with fuzzy inference in 3t mr images." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 01, pp. 74–86, 02 2006.

[8] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," *New York: Plenum Press*, 1981.

[9] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, pp. 32–57, September 1974.

[10] S. M. P. Yannis A Tolias, "On applying spatial constraints in fuzzy image clustering using a fuzzy rule-based system," *Signal Processing Letters, IEEE*, vol. 5, pp. 245–247, October 1998.

[11] A. Abdullah, A. Hirayama, S. Yatsushiro, M. Matsumae, and K. Kuroda, "Cerebrospinal fluid pulsatile segmentation-a review," in *Biomedical Engineering International Conference (BMEiCON), 2012*. IEEE, 2012, pp. 1–7.

[12] H. A. Y. S. M. M. Abdullah, A. and K. Kuroda, "Cerebrospinal fluid image segmentation using spatial fuzzy clustering method with improved evolutionary expectation maximization," in *Engineering in Medicine and Biology Society (EMBC) 2013 35th Annual International Conference of the IEEE*. IEEE, 2013, pp. 3359–3362.

[13] S. H. L. Liew, AW-C. and W. H. Lau, "Segmentation of color lip images by spatial fuzzy clustering." *Fuzzy Systems, IEEE Transactions on*, vol. 11, pp. 542–549, 2003.

[14] A. W.-C. Liew, S. H. Leung, and W. H. Lau, "Fuzzy image clustering incorporating spatial continuity," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 147, no. 2, pp. 185–192, 2000.

[15] P. Shen and C. Li, "Local feature extraction and information bottleneck-based segmentation of brain magnetic resonance (mr) images," *Entropy*, vol. 15, no. 8, pp. 3205–3218, 2013.

[16] G. Aubert and P. Kornprobst, "Mathematical problems in image processing: partial differential equations and the calculus of variations," *Springer-Verlag New York Inc.*, 2006.

[17] H. Khotanlou, O. Colliot, J. Atif, and I. Bloch, "3d brain tumor segmentation in mri using fuzzy classification, symmetry analysis and spatially constrained deformable models," *Fuzzy Sets and Systems*, vol. 160, no. 10, pp. 1457–1473, 2009.

[18] F. Tsai, C.-K. Chang, J.-Y. Rau, T.-H. Lin, and G.-R. Liu, "3d computation of gray level co-occurrence in hyperspectral image cubes," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2007, pp. 429–440.

[19] L. Tesař, A. Shimizu, D. Smutek, H. Kobatake, and S. Nawano, "Medical image analysis of 3d ct images based on extension of haralick texture features," *Computerized Medical Imaging and Graphics*, vol. 32, no. 6, pp. 513–520, 2008.

[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," 2010.

[21] C. Y. Ren and I. Reid, "Slic: a real-time implementation of slic superpixel segmentation," *University of Oxford, Department of Engineering, Technical Report*, 2011.

[22] A. Levinshtein, A. Stere, K. Kutulakos, D. Fleet, S. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows." *IEEE Transactions on Pattern Analysis. and Machine Intelligence*, vol. 31(12), pp. 2290–2297, 2009.

[23] X. Ren and J. Malik, "Learning a classification model for segmentation." in *Proceedings of the 9th IEEE International Conference on Computer Vision*, 2003, pp. 10–17.

[24] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation." *International Journal of Computer Vision*, vol. 59(2), pp. 167–181, 2004.

[25] O. Veksler, Y. Boykov, Mehrani., and P., "Supervoxels in an energy optimization framework." in *Proceedings of the 11th European Conference on Computer Vision*, 2010, pp. 211–224.

[26] A. Fabijanska and J. Goclawski, "The segmentation of 3d images using the random walking technique on a randomly created image adjacency graph," 2014.

[27] D. Mahapatra, P. J. Schuffler, J. A. Tielbeek, J. C. Makanyanga, J. Stoker, S. A. Taylor, F. M. Vos, and J. M. Buhmann, "Automatic detection and segmentation of crohn's disease tissues from abdominal mri," *Medical Imaging, IEEE Transactions on*, vol. 32, no. 12, pp. 2332–2347, 2013.

[28] H. E. Tasli, C. Cigla, and A. A. Alatan, "Convexity constrained efficient superpixel and supervoxel extraction," *Signal Processing: Image Communication*, vol. 33, pp. 71–85, 2015.

[29] B. Andres, U. Koethe, T. Kroeger, M. Helmstaedter, K. L. Briggman, W. Denk, and F. A. Hamprecht, "3d segmentation of sbfsem images of neuropil by a graphical model over supervoxel boundaries," *Medical image analysis*, vol. 16, no. 4, pp. 796–805, 2012.

[30] A. Foncubierta-Rodríguez, H. Müller, and A. Depeursinge, "Region-based volumetric medical image retrieval," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2013, pp. 867 406–867 406.

[31] S. Jia and C. Zhang, "Fast and robust image segmentation using an superpixel based fcm algorithm," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 947–951.

[32] S. Ji, B. Wei, Z. Yu, G. Yang, and Y. Yin, "A new multistage medical segmentation method based on superpixel and fuzzy clustering," *Computational and mathematical methods in medicine*, vol. 2014, 2014.

[33] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[34] A. Bakkari, E. Ben Braiek, I. Njeh, and A. Ben Hamida, "Automatic brain mr perfusion image segmentation using adaptive diffusion flow active contours based on modified fuzzy c means," in *Advanced Technologies for Signal and Image Processing (ATSIP), 2014 1st International Conference on*. IEEE, 2014, pp. 214–218.

[35] M. M. Morales DI and P. F, "Urban and non urban area classification by texture characterishics and data fusion," *International Geoscience and Remote Sensing Symposium*, vol. 6, pp. 3504–3506, 2003.

DOI: 10.15439/2015F333
ACSIS, Vol. 5

# An SDN-assisted System Design for Improving Performance of SVC-DASH

Cihat Cetinkaya, Yalcin Ozveren, Muge Sayit

Ege University
International Computer Institute
Izmir, Turkey
Email: {cihat.cetinkaya, muge.fesci} @ege.edu.tr;
yalcinozveren@gmail.com

*Abstract*— **Today, the most of the video streaming system provides quality adaptation and prefers to send their packets over HTTP. MPEG group has standardized Dynamic Adaptive HTTP Streaming (DASH) regarding this tendency on the adaptive HTTP streaming. Besides providing quality adaptation with a non-scalable codec, DASH standard also allows Scalable Video Coding (SVC) to adapt quality. Software Defined Networks (SDN) is a recently emerged networking paradigm. SDN enables to separate control and data plane of computer networks and hence provides flexibility to network operators to implement their own routing approaches. In this paper, we propose a system for increasing Quality of Experience (QoE) of SVC-DASH clients by utilizing SDN. Our experimental results show that the proposed system provides an increase in received video quality and decrease in outage duration and startup delay when compared to the performance of the client running over todays Internet implementing shortest path routing.**

## INTRODUCTION

CISCO estimates that IP video traffic will be up to 80% of global Internet traffic in 2018 [1]. The huge demand on IP video lead researches to develop video streaming applications providing high Quality of Experience (QoE) to the clients [2]. Video streaming applications need to adapt their bitrate according to the underlying network conditions to provide good performance in terms of QoE. Popular video streaming applications such as Youtube [3], Microsoft Silverlight [4], Apple HLS [5] send video packets over HTTP and perform quality adaptation by sending HTTP requests for receiving the video partitions with different qualities. HTTP protocol is preferred since it reduces the server load by HTTP caches while providing reliable data transfer.

In HTTP adaptive streaming, video files encoded at various rates are stored in the server. Each video file, i.e. representations, is separated into the small partitions called segments. The clients observe several parameters giving indication about underlying network capacity, such as buffer level or download duration to transfer the recent segments. As a result of observed parameters, the clients send their request indicating the selected representation for each segment. Hence, the selection of video quality changes over time regarding the client requests. Recently, MPEG group has standardized Dynamic Adaptive HTTP Streaming (DASH) [6]. DASH standard defines the formats of the segments and document containing the segment information.

With a non-scalable codec, video file should be separately encoded in order to obtain different representations having different quality. If scalable video coding (SVC) is used, video file is encoded once and video sequences with different quality can be extracted from this encoded file [7]. Hence, SVC provides the storage advantage. If SVC is used with HTTP streaming, storage advantage provides cache efficiency and optimizes the bandwidth usage since only one encoded video sequence is send to the HTTP cache for a video file [8]. Using SVC with DASH allows producing video sequences with large range of representations while providing cache efficiency [8].

Software defined networking (SDN) technology decouples the control and data plane of computer networks [9]. This separation allows network operators to implement different routing algorithms designed by considering specific application needs. In an SDN domain, a centralized device, called controller, has the network topology view. The controller determines the routing paths and sends related information to the forwarding elements. SDN paradigm offers a flexible platform to implement various routing algorithms designed for different network applications.

In this study, we propose a system for improving the performance of SVC-DASH clients running over SDN. We use quality of experience parameters such as received video quality, buffer underruns -i.e. outage durations-, and startup delay to measure the performance. SDN architecture enables to select different streaming path to transfer different video layer flows. We define a method for controller to learn which layers are streamed to which clients without communicating with the clients and the server, and provide path assignments using this information. We also give the design guideline of a SVC-DASH client running over SDN.

This work is funded by the Scientific and Technological Research Council of Turkey (TUBITAK) Electric, Electronic and Informatics Research Group (EEEAG) under grant 114E409.

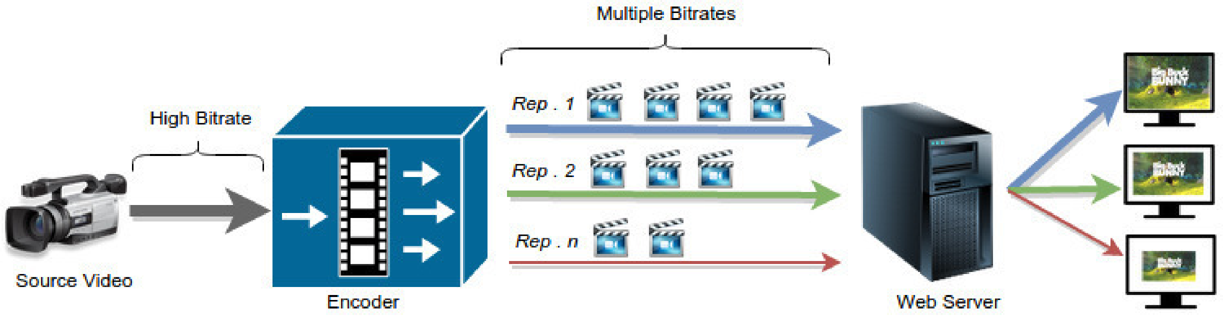


Fig. 1 DASH diagram showing different representations of the same video content.

The rest of the paper is organized as follows. In section two, the background of SDN and SVC-DASH are given with the related work proposed in the literature. In section three, the design of the proposed system is detailed. The performance of the proposed system is reported in section four. Finally, the conclusion is given followed by the references.

## BACKGROUND

### A. Dynamic Adaptive Video Streaming over HTTP

In DASH standard, more than one video file encoded at different bitrates for the same content is stored in a web server. In Fig. 1, a DASH diagram is given where a video content is encoded at three different bitrates with different resolutions.

In the server, each video sequence is separated into the segments. The segments can be independently decoded and has the data of fixed time interval. Information about the segments such as representation ID, URL address and bitrate are kept in Media Presentation Description (MPD) file. DASH standard defines MPD document and segment formats [6]. When a client connects to the DASH server, it gets the MPD file after establishing TCP connection between itself and the server. The quality adaptation is performed by client requests via selecting the segments with different qualities. Hence the complexity is given to the clients. DASH client decides which representation to be selected in each request. The clients adapt the video rate by changing the representations over time and send request to the server to get current segment of selected representation.

Rate adaptation algorithm is the algorithm that determines the policy of representation selection. The rate adaptation algorithms of DASH clients differ according to the parameter which is taken into account when selecting the next segment. This parameter can be measured bandwidth [10, 11] or the level of buffer fullness [12]. Some client algorithms use both of them to decide representation [13].

SVC provides to produce video sequences with various qualities in one file. The video content is encoded so as to consist of one base layer and one or more enhancement layers with SVC. While base layer can be decoded independently, it is required to base and *n*-1 enhancement layers to decode $n^{th}$ enhancement layer. Base layer has the minimum quality and each enhancement layers increase the quality. In SVC-DASH or Advanced DASH, each layer corresponds to one representation. The layer dependencies are defined in MPD. Hence, when a client requests for a segment, it may also request segments of lower layers according to the dependencies given in MPD document.

### B. Software Defined Networking

SDN paradigm enables to separate control and data planes of computer networks. This approach moves the intelligence from forwarding elements to an external device, called controller. Network operators can implement novel routing strategies taking requirements of network applications into account by programming the controller. The controller sends commands containing forwarding rules to the forwarding elements, i.e. switching devices. Another advantage that SDN provides is that the controller can have the network topology information such as traffic amount on the links, packet loss ratio or drop packets by communicating with the switches. The SDN architecture is given in Fig. 2.

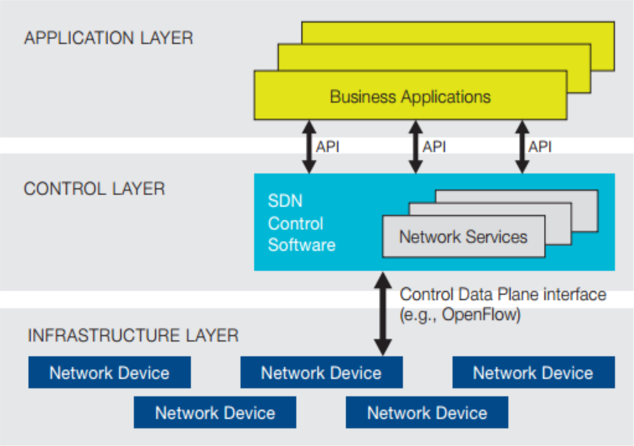


Fig. 2 SDN architecture showing decoupled structure of control plane and the data plane [9].

OpenFlow is the first protocol designed for providing communication between the controller and the switches. OpenFlow opens a secure communication channel. Switches update their forwarding tables according to the messages of the controller, sending via using OpenFlow protocol. When a switch gets a packet, it checks the header information of the packet, and performs an action regarding the defined rule for that information. In other words, the switches implements a <match, action> processing for each packet. Forwarding a packet, discarding a packet, or querying controller for asking the rules of a packet can be given as action examples.

### C. Related Work

Since both DASH standard and SDN paradigm have gained great attention from the academia and market, researchers has proposed several studies related on these topics recently.

In [14], dataset for SVC-DASH is presented. HTTP cache efficiency of SVC-DASH is discussed in [8]. In this study, it is shown that if SVC-DASH is used, then network traffic reduces when compared to the traffic amount of MPEG-DASH.

The performance in terms of QoE of DASH clients can be improved by utilizing SDN capabilities. In [15], the authors aim the provide fairness among DASH clients. For this purpose, the controller communicates with the clients and sends commands indicating which representation should be selected. This selection is determined according to the network topology information and client type. Since SDN provides controller to decide forwarding rules, different selection strategies for the flow routes between server and the clients is also proposed in several studies to increase QoE of DASH clients. In [16], a traffic shaping policy is determined for OpenFlow enabled switches, which gives priority to the HTTP flows. This approach provides DASH clients to achieve higher performance when compared to the case that HTTP flows has no priority. QoE centric path assignment over SDN networks for DASH clients is proposed in [17]. The controller communicates with DASH clients and if the performance decreases in terms of QoE parameters, the bottleneck point is detected by the information obtained from the switches. If the problem is in SDN network, then a different path is selected, otherwise the server is changed [17]. Path assignment by considering the bitrate of the segments and path capacities for increasing DASH clients' performance running over SDN networks is proposed in [18]. The server signals the controller to give information about the bitrate of the current segment and the paths are changed if the capacity of the current path is not enough to convey that segment packets.

Path assignment capability of SDN networks brings a natural approach to send packets of scalable coded video by sending base layer video over a different path. In [19], base layer video packets is send over a lossless path while the remaining traffic is send over the shortest path over SDN network. Besides sending base layer video packets over lossless path, forwarding enhancement layer packets over another path is proposed in [20]. The flow routes of base and enhancement layers packets are determined according to the output of an optimization model by considering packet loss and delay variation [21]. Quality adaptation techniques and flow route decisions considering the quality adaptation are not briefly discussed in these studies.

To the best of our knowledge, there is one approach proposed in the literature to route HTTP flows for SVC-DASH clients running over SDN. In [22], the path that has the smallest load is assigned to route base layer packets. Similar to base layer path assignments, the paths determined for enhancement layers are done based on layer priority; the packets of enhancement layer having higher priority are sent over a path having smaller load. The SVC-DASH client always requests all enhancement layers for each segment, i.e. it has no rate adaptation algorithm. Since rate adaptation is not implemented in that client software, there is no strategy defined for selecting the representation to be requested. The client sends requests in every 2 seconds [22].

In this study, we propose a system for SVC-DASH streaming over SDN. In the system, we think each video layer packets as a separate flow and assign streaming paths considering the bitrate of the layers. The paths of each layer can be the same if the available bandwidth is enough, but can be different, otherwise. Our approach differs from the studies in the literature in the following aspects. First, the controller does not communicate with the server or the clients to get information about the client requests. Second, we utilize different OpenFlow message types to detect received quality. And finally, we give an implementation guideline for SVC-DASH clients running over SDN.

## Flow Route Assignment for SVC-DASH over SDN

In this section we give the details of the proposed system for improving the performance of SVC-DASH clients running over SDN and give the details of SVC-DASH client implementation in the proposed system.

### D. SDN for SVC-DASH Clients

In the proposed system, flow routes i.e. streaming paths between the server and the clients are determined by the controller. The controller communicates with the switches in both proactive and reactive manner. It queries the switches periodically in order to obtain traffic information as proactive behavior. The controller calculates the available bandwidth of each path in the network by using the traffic information. Whenever the controller receives a message from a switch querying for a flow, it sends the forwarding rule to that switch. This is the reactive part.

Since the controller determines the paths for video flows between server and the clients, the switches need to communicate with the controller to learn the forwarding rules. The controller determines the paths for video layer packets; and then sends the forwarding rules related to the assigned path. Note that, the controller has to send the message containing forwarding rules to all switches along the chosen path.

As a path assignment strategy, the controller can assign the path having maximum available bandwidth for the newly joined client. Although the path having maximum available bandwidth is selected as streaming path for that client, sending each video layer packets over the same streaming path may decrease the capacity utilization. For example, suppose we have two paths where the capacity of first and second path equal to bitrate of base and first enhancement layer, respectively. In this case base layer packets can be sent over the first path and enhancement layer packets can be sent over the second path. But capacity of each path is not enough to transfer both layers packets.

Because of capacity utilization problem, we advocate to assign different streaming paths for the video layer packets requested by the same client for maximizing the network capacity utilization. For capacity utilization maximization, the best solution is to assign paths regarding the bitrate of the layers being sent. When a client joins the system, the controller makes system-wide path assignment for each client in the system by taking capacity of the paths and bitrates of the video layers into account.

In order to assign paths for the video layers packets of the clients, one approach is for controller to communicate with the clients or the server, to obtain requested layer information and to assign path according to the retrieved layer information. However, this reactive approach brings extra message complexity to the system. Instead of communicating with the server or the clients, we propose a method for controller to have the requested layer information without introducing extra messaging burden. We provide that the server sends each layer from a different TCP port and put this port information of video layers into the MPD file. The clients send video layer requests according to the given information in the MPD file. For example, if a client requests base and one enhancement layer, it sends these requests to defined ports of base and first enhancement layers, not the same port on the server. Hence, the switches look for a match of the tuples (client IP address, server IP address, port number) when receiving request messages of the clients. If this match does not present in its flow table, the switches sends a query containing the tuple.

SVC-DASH client establishes a TCP connection between the server and itself after connecting to the system. The first message of the TCP connection establishment is TCP SYN message being sent by the client. The OpenFlow enabled switch that newly joined client is connected to sends a PACKET_IN message to the controller after receiving TCP SYN message. The controller assigns paths upon getting a PACKET_IN message. Since the controller has not the information of what representation, i.e. which layers will be requested by the client; it determines the paths for base layer packets and for enhancement layer packets in case of there is enough network capacity.

```
Procedure Path_Assignment

1:  $c_i$: $i^{th}$ online client
2:  N: number of online clients
3:  $cp_p$: the capacity of path p

4:  $i = 0$;
5:  while $i$<N
6:      $p$: the path having max available bandwidth
7:      assign $p$ for the base layer packets of client $i$
8:      $cp_p = cp_p$ – bitrate of base layer;
9:      $i$++;
10: end while

11: $i = 0$;
12: current enhancement layer = first enhancement layer;

13: while current enhancement layer ≠ NULL && there
    exist any $p$ with $cp_p \geq$ bitrate of current enhancement layer

14:     while $i$<N
15:         $p$: the path having max available bandwidth
                                            (i.e. capacity)
16:         if ($cp_p \geq$ bitrate of current enhancement layer)
17:             assign $p$ for the current enhancement layer
                                packets of client $i$
18:             $cp_p = cp_p$ – bitrate of current enhancement layer;
19:         end if
20:         $i$++;
21:     end while

22:     if (current enhancement layer =
                                highest enhancement layer)
23:         current enhancement layer = NULL
24:     else
25:         current enhancement layer = next enhancement
                                                layer;
26:     end if
27: end while
```

Fig. 3 Path assignment algorithm running by the controller. The output of the algorithm determines the paths for all clients

The path assignment algorithm is given in Fig. 3. As given in the algorithm, the controller first assigns the paths having maximum available bandwidth for transferring base layer packets of all clients between the 5th and 10th lines of the algorithm. Since base layer packets are crucial, the controller has to assign paths for them even if the capacity of the network is not enough to send these packets properly. Upon completing the path assignment for the base layer packets, the controller assigns the paths for enhancement layer flows starting from the first enhancement layer for all clients in the while loop starting at line 14. But this time, if the network capacity is not enough to transfer the packets of an enhancement layer, then this assignment procedure is stopped. In other words, the paths for enhancement layer packets are assigned if and only if there is enough capacity.

After determining the paths for transferring the packets of video layers, the controller sends related flow information to the switches. The switches overwrite new forwarding rules to the existing entries in their flow tables. Note that at this point, although the switches have the information of rules for forwarding base and one or more enhancement layer packets, the clients may not request to receive any enhancement layer, in turn, these rules may not be used. Here, the controller determines the paths in proactive manner.

As defined in HTTP adaptive streaming, the clients request for a representation after it joins the system and gets MPD file from the server. Besides that, online clients continue to request representations according to their rate adaptation algorithms. At this point, there are three possible cases for a client:

(*i*) The controller may have assigned the paths for exactly same number of enhancement layers with the number of enhancement layers that a client requests to.

(*ii*) The controller may have assigned paths for *i* enhancement layers but the client requests more than *i* enhancement layers packets.

(*iii*) The controller may have assigned paths for *i* enhancement layers but the client requests less than *i* enhancement layers packets.

Clearly, for the first case, the forwarding rules for the requested representations, i.e. video layers flows are already determined on the flow tables of the switches.

In the second case, the switch gets a packet (or more than one packet) which flow information does not present in its flow table. It sends a PACKET_IN message to learn what to do with the received packet. This case occurred when the client measures available bandwidth and decides that it can receive more enhancement layers, which shows that the available bandwidth of at least one of the paths assigned for that client is higher than expected. Suppose the client receives video packets over n paths. The controller assigns the path having maximum available bandwidth among these n paths for the requested flow.

In the third case, the switches have more forwarding rule entry than it is required to. In this case, there are unused flow entries in the flow tables of the switches. The same situation can happen when a client decreases the quality by requesting less number of enhancement layers than that of it requested in previous period. In the OpenFlow protocol, each flow entry has an idle timeout which is determined by the controller. The switch removes a flow which has no matching packets in idle timeout and informs controller by sending FLOW_REMOVED message. In this work, the flow of base layers has no idle timeout where the enhancement layers is set to a certain threshold called flow timeout. When controller receives a FLOW_REMOVED message, it determines the path having minimum throughput among the paths which are assigned to transfer video layer packets for that client. The controller changes this path to a new path having maximum available bandwidth and sends related forwarding rules to the corresponding switches. The controller performs the second step in order to find better path which can be provide to increase the quality received by the client.

There is one thing left subtle but important. As stated in previous section, DASH clients request for the segments consecutively. If a segment is downloaded fast, the client waits for a while before it sends the request for the next segment. In other words, it enters the successive downloading and waiting periods. This phenomenon is known as on/off behavior [23]. On/ off period cause a traffic measurement problem if the controller measures of a current flow traffic when the client is in off period. If the client is in off period, then the controller measures the current traffic on that path as zero. In order to prevent that misconclusion, we have conducted a set of experiments and determine maximum length of off periods. Then we set idle timeout of the flows to a value higher than maximum off period length. Hence, we provide that the controller to sense traffic if there is at least one client receiving packets over corresponding path.

### *E. An Implementation of SVC-DASH Client Running over SDN*

Typically, two parameters are determined regarding the output of the rate adaptation algorithm: which representation (if SVC is used, which enhancement layers) will be requested and when it will be requested. In this section, we give the details of the implemented SVC-DASH client software.

When a client joins the system, it requests MPD file from the server as stated in DASH standard. Since we added the port information in MPD file, the client extracts the port information for each video layer from the MPD document besides the information of the bitrate of each video layer. The client also estimates the available bandwidth by calculating the download rate of the MPD document. Based on this calculation, the client determines the number of video layers to be requested for the first segment.

In SVC-DASH client implementation given in [22], the client requests base and enhancement layers segments consecutively. For example, if client decides to request video segment of second enhancement layer after executing rate adaptation algorithm; it sends request for base layer packets first. Upon receiving the base layer packets, it requests for the first enhancement layer packets, and after receiving first enhancement layer packets, it requests second enhancement layer packets.

Since the controller can assign different paths to the video layer flows in SDN network, the clients does not have to request video layers sequentially. In our SVC-DASH client implementation, the client requests video layers simultaneously after deciding which layers will be requested. Suppose the client decides to request one base and two enhancement layers. The clients send an HTTP GET message to the corresponding ports of the server at the same time. Hence, the client throughput is maximized since it does not have to wait downloading packets of a video layer to request the next layer packets for a segment.

In order to calculate throughput, we need a new method since each video layer packet transmitted over a different path. Suppose the client request base and $n$ enhancement layers. Let $t_{dt}$ represents the total time to download base and these $n$ enhancement layers. Then $t_{dt}$ and the throughput for the requested segment are calculated according to the formula given in (1) and (2), respectively. In the formulas, $e_i$ represents the $i^{th}$ enhancement layer, $b$ represents base layer. $t_x^{last}$ is the receiving time of the last packet of $x^{th}$ layer, $t_{request}$ is the sending time of the request and $br_x$ is the bitrate of the $x^{th}$ layer.

$$t_{dt} = \max(t_b^{last}, t_{e_1}^{last}, \ldots, t_{e_n}^{last}) - t_{request} \qquad (1)$$

$$\frac{br_b + \sum_{i=1}^{n} br_{e_i}}{t_{dt}} \qquad (2)$$

After calculating the throughput, the client selects the representation which has the maximum bitrate value lower than the throughput. The client sends next HTTP GET message upon downloading the requested layers packets of current segment.

## PERFORMANCE EVALUATION

### F. Experimental Setup

The experiments have been performed over a test bed shown in Fig. 4. SDN environment including OpenFlow-enabled switches and hosts that run SVC-DASH clients and SVC-DASH server are emulated using Mininet [24] software. SDN controller is implemented using Floodlight [25] software.

In the experiments, Big Buck Bunny video [14] which has one base layer and three enhancement layers are served by the SVC-DASH server. The bitrates of the video layers are given in Table 1. Note that the bitrates given in the table are cumulative because of inter-layer dependency. The video is divided into 300 segments each with a length of 2 seconds. Therefore the experiments last for 600 seconds. The clients join the system with interval of 30 seconds and never exit during the experiments. Flow idle timeout is set to 20 seconds.

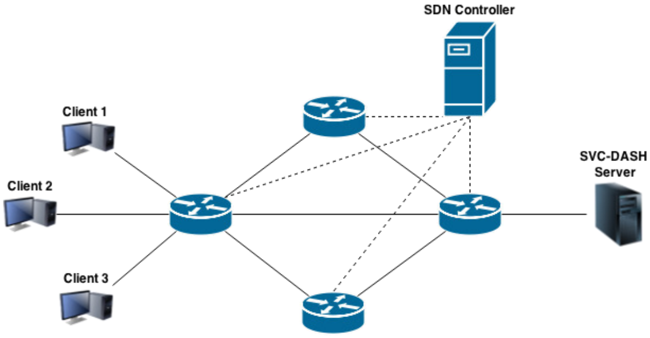


Fig 4. Experimental Testbed

TABLE I.
THE BITRATE DISTRIBUTION OF VIDEO LAYERS

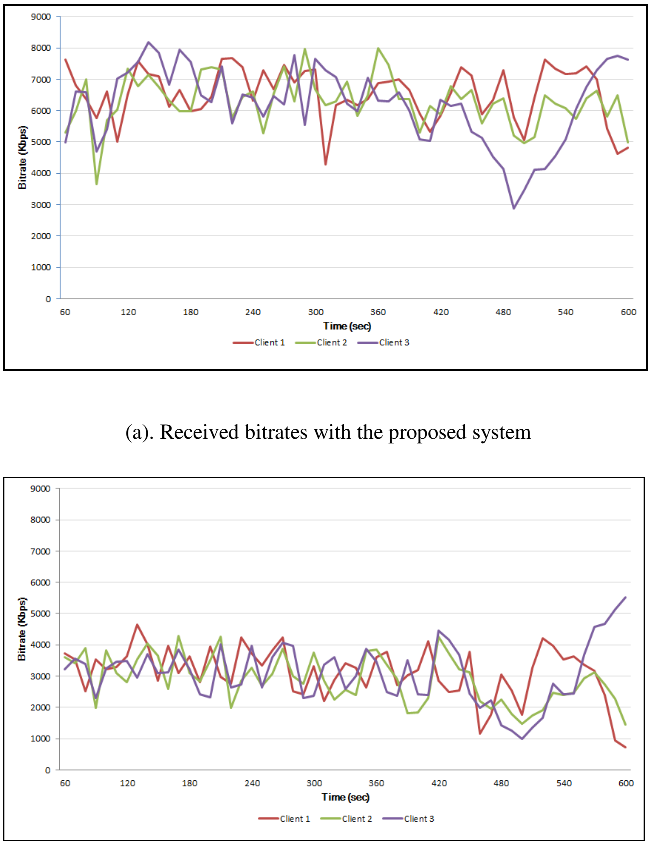
| Video Layer | *Bitrate* |
|---|---|
| Base | 1555 Kbps |
| Base + Enhancement 1 | 2700 Kbps |
| Base + Enhancement 1 + Enhancement 2 | 4547 Kbps |
| Base + Enhancement 1 + Enhancement 2 + Enhancement 3 | 6857 Kbps |

As seen from Fig. 4, there are three different paths between server and the clients in the network topology. The capacity of each path in the network equals to 6000 Kbps. This means that although the network capacity is enough to send base layer to each client, there is not enough capacity to send the highest enhancement layer to all clients in the system.

### G. Experiment Results

In order to show performance of the proposed system, we observe several QoE parameters: received bitrate, received video quality and, number and length of durations. We also perform experiments with same set of parameters by using traditional Internet shortest path routing and give the results comparatively. The topology includes only one shortest path with the length of one hop. Hence, the video packets are streamed over the same path with shortest path routing. Each experiment is repeated 10 times; averaged results are obtained and shown in the graphs.

In Fig. 5a and Fig. 5b, the average bitrates received by each client in the proposed system and in the system with shortest path routing are given as a function of time, respectively. While the averaged received bitrate values for the clients are in the range of 5000 Kbps and 8000 Kbps, these values are in the range between 2000 Kbps and 4000 Kbps with shortest path routing. In both approaches, the clients can achieve bitrate higher than the capacity of the paths. The main reason of that is on/off period pattern of the clients. A client may even use the full capacity of a path if other clients using that path are in their off periods.



(a). Received bitrates with the proposed system

(b). Received bitrates with the shortest path routing

Fig 5. Averaged received bitrate as a function of time

In order to show the quality of the received video, we measured the total length of the video segments that received from each layer and give the results in Fig. 6. In the figure, base and enhancement 3 denotes the minimum and maximum quality, respectively. During the streaming session, the clients in the proposed system experience the video quality provided by the enhancement 2 most of time, where the clients receive the base layer quality in the shortest-path approach. Furthermore, the clients in the shortest-path approach never experience the quality provided by the enhancement 3. Note that, the values given in the graph are cumulative. In other words, if receiving an enhancement layer is shown in the graph, it means base and all enhancement layers below the received enhancement layer are also received.

When the bandwidth is not enough to send the selected representation, the clients may drain their buffers and then start to experience outages. In Fig. 7, average durations in seconds observed in the proposed system and the shortest path approach are given. As it is seen, the proposed system has a lower number of outages. The reason of this is that the proposed system dynamically changes the paths and provides to increase network capacity utilization. Thus the clients get the maximum available bandwidth. But in the shortest path approach, the clients always use the same path. So their share of bandwidth decreases which causes outages to increase. In Fig. 8, CDF graph of total outage durations is given in order to show the distribution of outage durations.

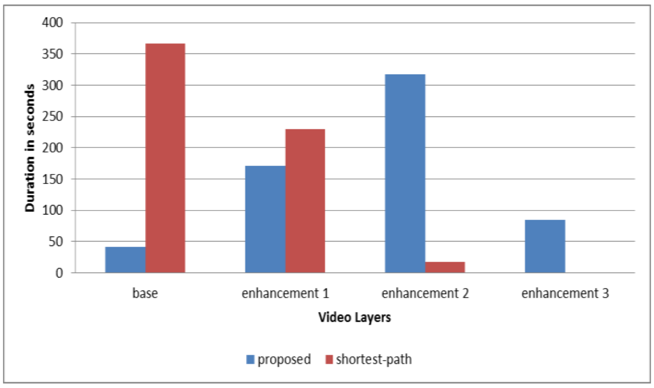


Fig 6. Received video quality in terms of layers

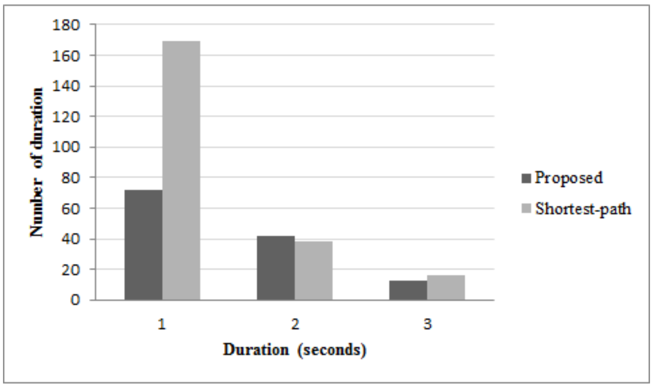


Fig 7. Total duration values of outages measured in the experiments

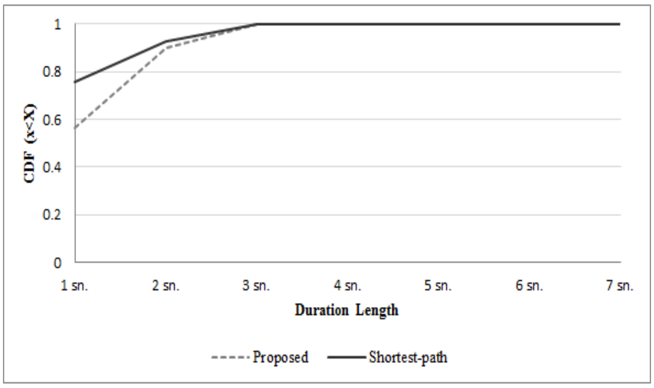


Fig 8. CDF graph of observed duration of outages

In Table 2, averaged startup delays experienced by each client are given. Startup delay equals to the elapsed time from requesting the first segment to playing the video. As seen from the table, limited capacity causes an increase in startup delays. The clients experience longer startup delay with shortest path routing.

TABLE II. STARTUP DELAY VALUES IN SECONDS

| | ***Proposed system*** | ***Shortest-path routing*** |
|---|---|---|
| Client 1 | 10 sec | 10 sec |
| Client 2 | 11 sec | 12 sec |
| Client 3 | 11 sec | 17 sec |

## CONCLUSION

In this paper, we propose a system for increasing QoE SVC-DASH client running over SDN. For this purpose, we think each layer of the video as a separate flow and assign paths by taking path capacity and the bitrate of the layers into account.

Our system differs from the similar studies in the literature as the controller does not communicate with the server and the clients to detect which video layers are sent to the clients. In order to detect requested number of enhancement layers, we utilize PACKET_IN and FLOW_REMOVED messages sent by the switches to the controller. The controller decides system wide re-routing of flows or change the path of a flow according to the received message type and capacity of the paths.

The performance of the proposed system is measured by considering the QoE parameters such as received video quality, outage duration and startup delay. When we compare the achieved performance of the proposed system with the traditional shortest-path routing of today's Internet, our system provides to yield better performance for each QoE parameter.

In the future, we are planning to measure the performance of the proposed system with larger number of clients, different types of video resolutions and with different types of network topologies.

## REFERENCES

[1] Cisco VNI, "Cisco Visual Networking Index: Forecast and Methodology, 2013–2018", White Paper, 2014.

[2] M. Dąbrowski, "Emerging technologies for interactive TV", in proc. of the Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 787–793, 2013.

[3] http://www.youtube.com

[4] Microsoft, IIS Smooth Streaming Transport Protocol, Sept. 2009; https://msdn.microsoft.com/en-us/library/ff469518.aspx/[MS-STTR].pdf.

[5] https://developer.apple.com/streaming/

[6] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the Internet", IEEE Multimedia, pp. 62–67, 2011.

[7] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 9, pp. 1103–1120, 2007.

[8] Y. Sanchez, C. Hellge, W.V. Leekwijck, Y.L. Louédec, and T. Schierl, "Scalable Video Coding based DASH for efficient usage of network resources", Position Paper for the Third W3C Web and TV workshop, Los Angeles, CA, USA, September 2011.

[9] Open Networking Foundation (ONF), "Software defined networking: the new norm for networks", White paper, 2012.

[10] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE", in proc. of the 8th international conference on Emerging networking experiments and technologies (CoNEXT), 2012.

[11] B. Rainer, S. Lederer, C. Müller, and C. Timmerer, "A seamless Web integration of adaptive HTTP streaming", in proc. of EUSIPCO, 2012.

[12] T.Y. Huang, R. Johari, and N. McKeown., "Downton abbey without the hiccups: Buffer-based rate adaptation for HTTP video streaming", in proc. of the ACM SIGCOMM workshop on Future human-centric multimedia networking (FhMN), NY, USA, 2013.

[13] L. D. Cicco, V. Caldaralo, V. Palmisano, and S. Mascolo, "ELASTIC: A client-side controller for dynamic adaptive streaming over HTTP (DASH)", in proc. of the IEEE 20th international Packet Video Workshop (PV), 2012.

[14] C. Kreuzberger, D. Posch and H. Hellwagner, "A scalable video coding dataset and toolchain for dynamic adaptive streaming over HTTP", in proc. of the ACM MMSys, Portland, Oregon, 2015.

[15] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, "Towards network-wide QoE fairness using OpenFlow-assisted adaptive video streaming", in proc. of the ACM SIGCOMM workshop on Future human-centric multimedia networking, 2013.

[16] M. S. Seddiki, M. Shahbaz, S. Donovan, S. Grover, M. Park, N. Feamster, and Y. Song, "FlowQoS: QoS for the rest of us", in proc. of the hotSDN workshop, 2014.

[17] H. Nam, K. Kimy, J. Y. Kimy, and H. Schulzrinne, "Towards QoE-aware Video Streaming using SDN", in proc. of the GLOBECOM, 2014.

[18] C. Cetinkaya, E. Karayer, M. Sayit, and C. Hellge, "SDN for Segment based Flow Routing of DASH", in proc. of the IEEE 4th International ICCE conference, 2014.

[19] S. Civanlar, M. Parlakisik, A.M. Tekalp, B. Gorkemli, B. Kaytaz, and E. Onem, "A QoS-enabled OpenFlow environment for Scalable Video streaming", in proc. of the IEEE GLOBECOM Workshops, pp. 351-356, 2010.

[20] H.E. Egilmez, B. Gorkemli, A.M. Tekalp, and S. Civanlar, "Scalable video streaming over OpenFlow networks: An optimization framework for QoS routing", in proc. of the 18th IEEE International Conference on Image Processing (ICIP), pp. 2241-2244, September 2011.

[21] H.E. Egilmez, S. Civanlar, and A.M. Tekalp, "An optimization framework for QoS-enabled adaptive video streaming over OpenFlow networks", IEEE Transactions on Multimedia, vol. 15, no. 3, pp. 710-715, 2013.

[22] S. Laga, T. Van Cleemput, F. Van Raemdonck, F. Vanhoutte, N. Bouten, M. Claeys, and F. Amd De Turck, "Optimizing scalable video delivery through OpenFlow layer-based routing", in proc. of the IEEE Network Operations and Management Symposium (NOMS), 2014.

[23] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. Begen, and D. Oran, "Probe and adapt: Rate adaptation for HTTP video streaming at scale", IEEE J. on Selected Areas in Comm., vol. 32, no. 4, pp. 719-733, 2014.

[24] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: Rapid prototyping for software-defined networks", in proc. of the ACM SIGCOMM Workshop on Hot Topics in Networks, pp. 1-6, USA, 2010.

[25] Floodlight. http://www.projectfloodlight.org/floodlight/.

DOI: 10.15439/2015F272
ACSIS, Vol. 5

# The Scalable Distributed Two-layer Content Based Image Retrieval Data Store

Stanisław Deniziak
Kielce University of Technology
al. Tysiaclecia Panstwa Polskiego 7,
25-314 Kielce, Poland
Email: s.deniziak@tu.kielce.pl

Tomasz Michno, Adam Krechowicz
Kielce University of Technology
al. Tysiaclecia Panstwa Polskiego 7,
25-314 Kielce, Poland
Email: {t.michno, a.krechowicz}@tu.kielce.pl

***Abstract*—The multimedia databases are becoming more and more popular nowadays. One of their main problem is a huge data amount storage. Another problem with multimedia databases is querying. Traditional approaches, based on textual keywords are not sufficient. More advanced techniques, incorporating image content features, should be used. In this paper we propose new multimedia database structure with ability of Content Based Image Retrieval which is based on our previous work: Query by Shape method (QS). Query by Shape is a method which is based on decomposing an object into features. Each feature may consists of shape primitive, a color or a texture. In this paper we only use shape primitives. In order to achieve high scalability and workload control, we propose a modified Scalable Distributed Two-layer Data Structure, as a storage. The modification incorporates adding tree structure, comparing algorithm and returning a set of results to the client.**

## I. Introduction

THE MULTIMEDIA databases are becoming more and more popular nowadays. There are many applications where they are needed, like social media portals (e.g. Facebook, Instagram, Flickr and Google+) or monitoring systems. Because modern cameras produce high resolution images, the amount of data which has to be stored is very huge. Another problem with multimedia databases is their querying. Traditional approaches, based on textual keywords are not sufficient. More advanced techniques incorporating image content features should be used.

In this paper we propose the idea of a multimedia database structure with ability of Content Based Image Retrieval which is based on our previous work described in [1], Query by Shape (QS). Query by Shape is a method which is based on decomposing an object into features [1]. Each feature may consists of shape primitive, a color or a texture. In this paper we only use shape primitives. As a data structure for storage we use a modified Scalable Distributed Two-layer Data Structure which is highly scalable, distributed data store [2]. The modification incorporates adding tree structure, comparing algorithm and returning a set of results to the client.

This paper is organized as follows. The Section II presents the survey of image retrieval algorithms. The Section III contains a short review of NoSQL data stores. The idea of Scalable Distributed Two-layer Data Structures is described in the Section IV. The Section V shows the motivation of our research. The idea of our database structure is presented in the Section VI. The conclusion of the research is given in the Section VII.

## II. Image Retrieval algorithms

In the area of multimedia databases, three types of retrieval algorithms can be distinguished: Keyword-Based Image Retrieval (KBIR), Content-Based Image Retrieval (CBIR) and Semantic-Based Image Retreival(SBIR).

The first group, KBIR, is based on the relational database approach, where images are stored in the database and they are described using keywords. During the query, the proper keywords should be given. The database structure is very simple but strongly relies on the textual annotations given by a human. This approach is prone to mistakes because of the subjective kind of descriptions [1]. Moreover it is very hard to cover the whole information, present in the image, using textual description [3], [4].

The CBIR algorithms are based on different approach than KBIR. Image features are used to index images and perform queries [1]. All algorithms in this group could be divided into two categories: low-level and high-level algorithms. The low-level algorithms process images globally, extracting features from the whole frame, using e.g. a normalized color histogram [5], a spatial domain [6], a difference moment and entropy [7] or an MPEG-7 image descriptors like shape and texture [8]. The low-level features used by CBIR algorithms are easy to compute but they are insufficient if the query is oriented on searching for similar objects rather than whole images, which incorporates separating the object from the background.

The high-level CBIR algorithms provide more reliable and precise results in this situation. The major part of the algorithms from this group are based on the regions extraction and graphs matching. A region is a group of similar pixels, most often grouped by colors. There are also methods which uses during region extraction: a set of primitives [9] or fuzzy pattern [10] detection, moment-based local operators [11] or parallelograms, ellipses, corners and arcs detection [12]. After region extraction, a graph is being constructed in order to store the relations between them [1]. During the multimedia database query, the graph-subgraph matching is performed

978-83-60810-66-8/$25.00©2015, IEEE

e.g. using the classic Ullman algorithm [13] or more advanced ones. There are algorithms that are automatic or semi-automatic with ability to present preliminary results to the user who may choose important regions and repeat the query [14]. Another group of CBIR algorithms allows queries without full knowledge about searched images or objects. One of the first and most successive approaches in this area uses a human drawn sketches which are compared with the corresponding sketches in the database, globally for the whole image [15]. Another example is Query by Shape method [1], which is our previous research and which is used as a base of this paper. The idea of the algorithm is to decompose objects into features like shape primitives or color features. The features are not used for region extraction but for constructing an object's sketch or skeleton which is strictly a graph. Also the matching algorithm is proposed which uses the $similarity$ coefficient. The $similarity$ informs how similar two graphs are. If they are the same, the $similarity$ is equal to 1, if they are completely different, it takes 0 value. All intermediate values indicates that graphs are partially similar.

The SBIR algorithms are algorithms that try to overcome the "semantic gap", which is the difference between what is present on the image and what a human could interpret [4], [16]. Most SBIR algorithms are based on textual description which, in contrast to the KBIR algorithms, is a much longer phrase. The phrase is easy to create, use and understand by the human, the example could be "the sunrise in the mountains". The textual descriptions are not used directly, but they are mapped onto semantic features and then a query is performed [17].

All groups of algorithms need efficient data storage methods. One of the most often used structure is a cell or a tree, because it can store the relations between similar images [8]. There are also approaches that joins both data structures. One of the example is [18] which is based on gathering similar records in the same cells. Additionally, some biological processes are added, like mitosis, when the similarity between items in the same cells is below the specified level.

## III. The NoSQL data stores

The multimedia databases very often stores millions of photos or images. Because there has to be stored and processed very huge amount of data, with high availability and workload management, the traditional SQL-based databases may not be sufficient. Much better results are obtained using NoSQL databases. The NoSQL databases may be classified into the following groups: Graph databases, Key-Value data stores, Document data stores and Column-based data stores [19].

The Graph databases provides similar features as relational databases but they are not well-prepared to store huge amount of data because of the problems with scalability [19]. Key-Value data stores use an unique single key (e.g. a number) for storing data (value). Most often they have only two operations: inserting and retrieving data [20]. The examples of such data stores are Dynamo data store by Amazon which uses single integer key [20] or Apache Hbase which identifies the data by a timestamp, column name and row name [21]. The Document data stores use textual data interchange formats, like JSON or xml to store data. The most popular document data store is MongoDB [19].

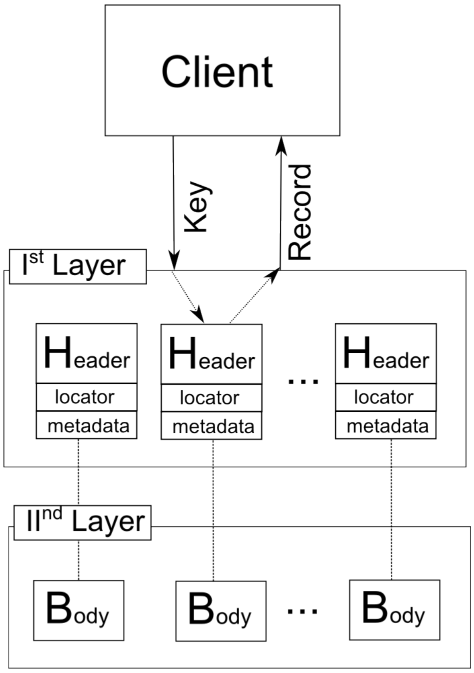


Fig. 1. The SD2DS structure overview.

The NoSQL data stores were successfully adapted to store images. As an example the Apache Hbase may be given, which was used to store Google Earth images [21].

## IV. The SD2DS data stores

Scalable Distributed Two-layer Data Structures (SD2DS) are one of variants of Scalable Distributed Data Structures (SDDS) [2], which may be classified as an distributed NoSQL data store. They stores data using a key and a value in so-called "buckets". Each bucket has storage limit and when it is reached a split is performed. Buckets may be distributed through many machines, e.g. nodes on the multicomputer. The SD2DS stores data in two separate layers. The first layer contains only data headers which consists of the data locator and metadata. The data locator is used to locate the stored data in the second layer. The metadata is an additional information connected with the actual data and may contain optional information e.g. the data length, a checksum, insertion date or data description. The second layer contains only the actual data, called body [2]. That structure maintains data more efficiently, because both layers are independent of each others and may be stored on different machines. Moreover the data store is highly scalable, without theoretical limitations and may contain as many data as is needed. The SD2DS allows also to use as a second layer other solutions, even without buckets. The SD2DS structure is presented in the Fig. 1.

The SD2DS may be easily extended e.g. by adding throughput adjustment [22] which highly improves the data access time.

## V. Motivation

The problem of multimedia database querying is very complicated. As a continuation of the previous research [1], we would like to propose the Content Based Image Retrieval Database incorporating Query by Shape Method. The database has to store a very huge amount of data. Simple records table, a tree or a cell would be not sufficient. Because of that, more advanced data stores, especially NoSQL-based should be used. In order to obtain very high scalability and availability, as a base for our system the Scalable Distributed Two-Layer Data Structures should be used. As a result of our research we would like to create the system which will fulfil the following requirements:

- queries using a graph of features e.g. primitives,
- similarity control for graphs comparison, the similarity should be set by the client,
- client feature: graphical queries easily drawn by a human (without drawing skill), e.g. using predefined shapes,
- client feature: automatic image transformation into a graph, used for a query,
- very fast data access,
- scalability without turning off the system,
- easy addition of the new hardware used for storage,
- good performance for the very huge data workload stored in the database.

## VI. The system overview

The proposed system extends the SD2DS data store structure. The classical approach consists of two layers with the first layer containing record's headers, the second with their bodies and the client which uses single key to retrieve the single record. In order to provide features mentioned in the previous section, some modifications have to be made. The system overview is shown in the Figure 2.

### A. The Client

The client should query the data store by a graph, instead of a standard key. Moreover it has to be able to receive the result of a query which may consists of many records. Therefore, the following messages are defined:
Messages send by the client:

- record (graph and image) insertion (GI) - a message used to insert a new record containing a graph and an image into the database,
- query (SQ) - a query after which all result records are sent by the database in a one message,
- distributed query (DQ) -a query after which result records are sent gradually with one message per one record, allowing e.g. to show progressively results for a user,
- get record (GR) - a message used to retrieve the record using its key.

Messages received by the client:

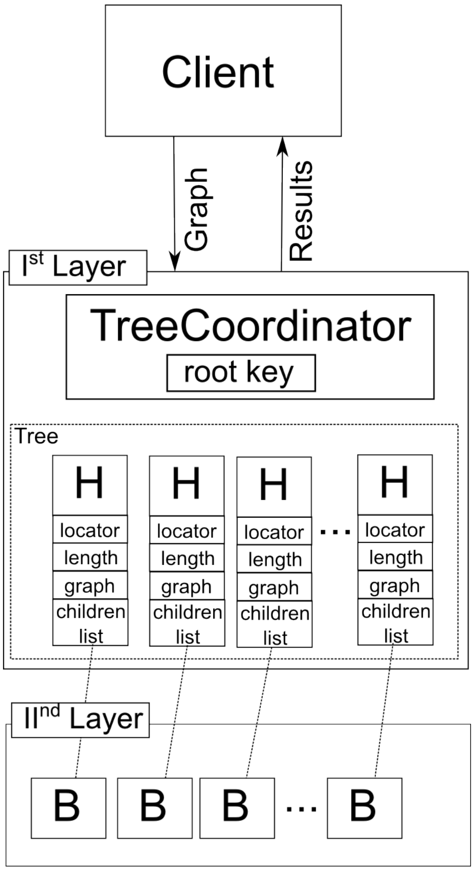


Fig. 2. The proposed system overview.

- result of the insertion (RI) - sent by the database as a response to GI message, returning the failure code or new node's record key,
- not found (NF) - sent by the data store when there are no similar graphs/images in the data store, returned for both SQ and DQ messages,
- results (RR) - contains list of similar records, returned only as a response to SQ message,
- single result (SR) - contains only one similar record, returned only as a response to DQ or GR messages.

All sent messages contain the query graph and the minimal similarity which is used during comparisons. Moreover the messages may define if as a result full record or only a header should be returned.

### B. The First Layer

The first layer of the data store was redesigned into two sub-layers: the Tree Coordinator and the Tree. The Tree sublayer consists of the standard SD2DS record headers. Each record header is treated as a one tree node, containing as a metadata a graph and a list of children nodes (a list of records keys). The Tree Coordinator is responsible for communicating with the clients and the logic of the tree: storing the tree root record key, adding new nodes, traversing tree and making queries. Moreover it is able to execute queries in two ways, analogously to the SQ and DQ messages. In order to improve the performance, some of the Tree Coordinator features, like comparing graphs and tree traversing may be implemented as a part of buckets. Also the $DQ$ message could be directly sent by the bucket.

*1) The tree structure:* The tree nodes used in the database are stored in the first layer using standard SD2DS headers records. The graphs are inserted into the tree hierarchically, storing its common parts in parent nodes. Because of that, the root node may contain a graph with only one shape or an empty graph. The example tree structure with scooter, bicycle and car objects is shown in the Fig. 3. Moreover, because there are indirect nodes which stores only graphs without connection to any images, the SD2DS modification allows records headers in the first layer with empty bodies.

*2) Record Insertion:* after receiving the GI message the Tree Coordinator compares the graph with graphs stored in the tree, starting from the root node using the Query by Shape comparison algorithm (QS) [1] and the minimal similarity parameter extracted from the GI message. The algorithm is presented in the Alg. 1. The Tree Coordinator retrieves the record using similar procedures as the client in the standard SD2DS. Then children nodes are extracted from the record and used for the next records retrieval.

*3) Querying:* The querying algorithm is very similar to the insertion of a node. The algorithm is shown in the Alg. 2. During the query, the results could be sent immediately to the client or stored in the Tree Coordinator temporary buffer before completion and then sent.

### *C. The Second Layer*

The second layer has not been modified. Because of the tree structure, there may be less bodies than records in the first layer. The images are retrieved during querying, but the client in the $SQ$ or $DQ$ could force the database to only send headers from the first layer or even only the keys.

## VII. Experimental results

The presented tree-based structure has been evaluated using experimental implementation written in C++. The results were compared with the linear SD2DS QS system and are shown in the Tables I and II. The linear SD2DS QS system uses also first SD2DS layer to store graphs, but without tree structure. During the query, all elements has to be compared with the query graph. In order to evaluate algorithms performance, the precision and recall coefficients were used [1]:

$$precision = \frac{number\ of\ relevant\ results\ images}{total\ number\ of\ results\ images} \tag{1}$$

$$recall = \frac{number\ of\ relevant\ results\ images}{total\ number\ of\ relevant\ images\ in\ the\ database} \tag{2}$$

As a number of relevant results images we assume the number of images which are from the same class as the query image. For the tests, about 117 real life images of different objects classes were used.

**Algorithm 1** Inserting new node to the tree

```
Require: $graph$ and $image$ extracted from the $GI$ message
  if tree is empty then
      add record as a root node, store its key as a Root Key;
      send $RI$ message with the key to the client;
  else
      add Root Key to the FIFO queue;
      while FIFO queue is not empty do
          $key \leftarrow$ pop first element from FIFO;
          $treeNode \leftarrow$ get record with key == $key$;
          compare $graph$ with $treeNode$ using QS;
          if $treeNode$ is not a root node then
              if graphs similarity $\leq$ similarity of $graph$ and $treeNode$'s parent then
                  add $graph$ as a child of $treeNode$'s parent;
                  send $RI$ message with the $graphs$'s key to the client;
              end if
          else
              add $graph$ as a new root node;
              add $treeNode$ to the $graph$'s children list

          end if
          if graphs similarity $\geq$ highSimilarity then
              add record as a child to $treeNode$;
              send $RI$ message with the record's key to the client;
          else
              if $treeNode$ does not have any children then
                  add $graph$ to the tree;
                  add the common part of $graph$ and $treeNode$ as their parents;
                  if $treeNode$ is a root node then
                      the Root Key $\leftarrow$ common part's key
                  end if
                  send $RI$ message with the $graph$'s key to the client;
              else
                  add each $treeNode$ children to the FIFO queue;
              end if
          end if
      end while
  end if
```

The test results shows that for most queries the tree structure increased the precision comparing to the previous, linear structure. This is due to the additional steps during the query algorithm (Alg. 2) which omits the whole sub-tree with too low precision and check if the precision is increasing in children nodes. Moreover, the lower precision in the automated queries was caused by some failures of shape detection algorithms.

The recall measurements shows that for the manual queries more relevant results were returned by the tree algorithm version. The automatic query was problematic for same cases for tree version because of occurrence of many unconnected

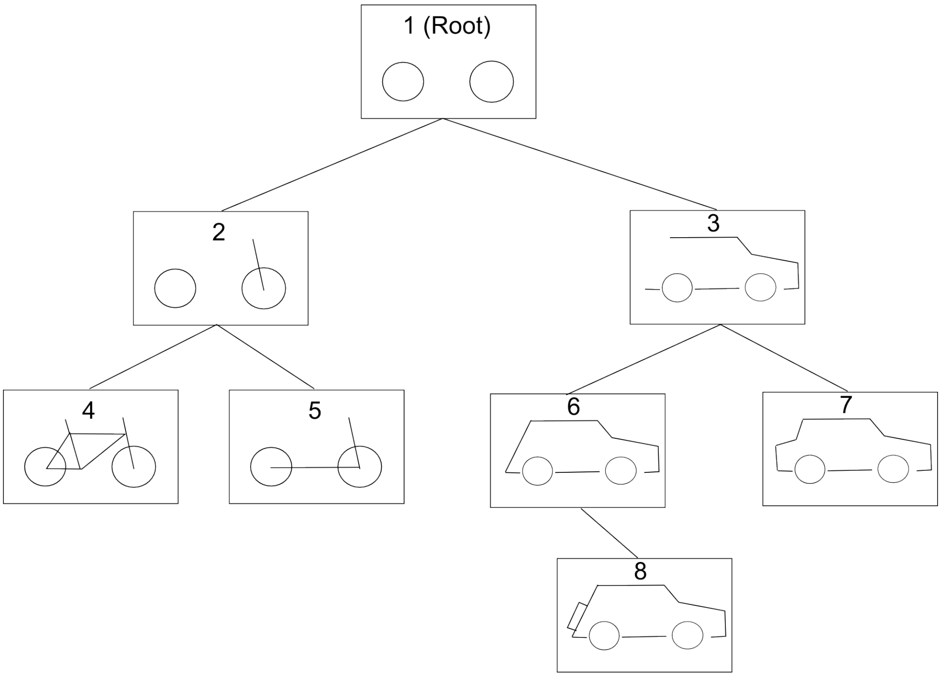


Fig. 3. The example of the tree with bicycle (node 4), scooter (node 5) and 3 cars graphs (nodes 6-8). The shape sizes were omitted during tree construction in order to show the main idea and simplify the structure.

TABLE I
THE PRECISION RESULTS FOR THE EXAMPLE BICYCLE AND CAR QUERIES

| | **Manual queries** | | **Automated queries** | |
|---|---|---|---|---|
| | **Tree-based** | **Linear** | **Tree-based** | **Linear** |
| bicycle no. 1 | 0.7708 | 0.6065 | 1 | 0.9 |
| bicycle no. 2 | 0.8529 | 0.5714 | 1 | 0.8 |
| car no. 1 | 1 | 0.7059 | 0.4182 | 0.5658 |
| car no. 2 | 1 | 0.8 | 0.5902 | 0.5428 |

TABLE II
THE RECALL RESULTS FOR THE EXAMPLE BICYCLE AND CAR QUERIES

| | **Manual queries** | | **Automated queries** | |
|---|---|---|---|---|
| | **Tree-based** | **Linear** | **Tree-based** | **Linear** |
| bicycle no. 1 | 0.9737 | 0.9737 | 0.3023 | 0.2093 |
| bicycle no. 2 | 0.7636 | 0.8421 | 0.2558 | 0.0930 |
| car no. 1 | 0.5814 | 0.5581 | 0.5 | 0.86 |
| car no. 2 | 0.7209 | 0.4651 | 0.72 | 0.76 |

nodes in skeletons. This caused problems with inserting them in the proper sub-tree.

The performance of the structure is dependent on the number of elements. During our experiments using SD2DS, to up to 1000 elements there were very small difference between tree and linear version. For higher number of elements, the tree structure was becoming more and more faster. We also implemented version which uses as a storage vector array. During experiments the tree version occurred to be up to 2x faster than linear version.

## VIII. CONCLUSION

In this paper a new Content Based Image Retrieval database structure was presented. The main idea of our research is to apply our previous research results, Query by Shape method [1], into Scalable Distributed Two-layer Data Structure. The modification of data store included redesigning it to store records in a tree. Also a Tree Coordinator was added to the first layer in order to communicate with the client and perform tree queries.

The future research includes implementing the version of the algorithm with comparison algorithm moved to buckets. We expect that it should improve the querying time as well as allows to use buckets and nodes more efficiently. Moreover the throughput adjustment may be added in order to improve the image data access time.

Another direction of research will concern the Query by Shape method in order to improve the results precision. This may include e.g. adding other primitive types and color features. Moreover more advanced graph matching algorithms should be applied, as well as some optimization methods should be evaluated, e.g. [23].

## ACKNOWLEDGMENT

The research used equipment funded by the European Union in the Innovative Economy Programme, MOLAB - Kielce University of Technology.

## REFERENCES

[1] S. Deniziak and T. Michno, “Query by shape for image retrieval from multimedia databases,” in *Beyond Databases, Architectures and Structures*, ser. Communications in Computer and Information Science, S. Kozielski, D. Mrozek, P. Kasprowski, B. Malysiak-Mrozek, and D. Kostrzewa, Eds. Springer International Publishing, 2015, vol. 521, pp. 377–386. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-18422-7_33

**Algorithm 2** Querying the database with a graph

```
Require: $graph$ extracted from the $SQ$ or $DQ$ query message
if tree is empty then
    send $NF$ message to the client;
else
    add Root Key to the FIFO queue;
    while FIFO queue is not empty do
        $key \leftarrow$ pop first element from FIFO;
        $treeNode \leftarrow$ get record with key == $key$;
        compare $graph$ with $treeNode$ using QS;
        if graphs similarity $\geq$ highSimilarity then
            if the client sent $DQ$ message then
                send $RI$ message with the record's key to the client;
                send $RI$ message with all keys of the sub-tree consisting of record's children to the client;
            else
                add record to the results;
                add all keys of the sub-tree consisting of record's children to the results;
            end if
        end if
        if $treeNode$'s parent similarity $- minSimilarity >$ graphs similarity then
            continue;
        end if
        if $treeNode$ has children then
            add each $treeNode$ children to the FIFO queue;
        end if
    end while
end if
```

[2] K. Sapiecha and G. Lukawski, "Scalable distributed two-layer data structures (sd2ds)," *IJDST*, vol. 4, no. 2, pp. 15–30, 2013. [Online]. Available: http://dx.doi.org/10.4018/jdst.2013040102

[3] C.-Y. Li and C.-T. Hsu, "Image retrieval with relevance feedback based on graph-theoretic region correspondence estimation," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 447–456, April 2008.

[4] H. H. Wang, D. Mohamad, and N. A. Ismail, "Approaches, challenges and future direction of image retrieval," *CoRR*, vol. abs/1006.4568, 2010.

[5] M. Mocofan, I. Ermalai, M. Bucos, M. Onita, and B. Dragulescu, "Supervised tree content based search algorithm for multimedia image databases," in *6th IEEE International Symposium on Applied Computational Intelligence and Informatics*, May 2011, pp. 469–472.

[6] T. K. Shih, "Distributed multimedia databases," T. K. Shih, Ed. Hershey, PA, USA: IGI Global, 2002, ch. Distributed Multimedia Databases, pp. 2–12. [Online]. Available: http://dl.acm.org/citation.cfm?id=510695.510697

[7] H.-P. Kriegel, P. Kroger, P. Kunath, and A. Pryakhin, "Effective similarity search in multimedia databases using multiple representations," in *12th International Multi-Media Modelling Conference Proceedings*, 2006, pp. 4 pp.–.

[8] C. Lalos, A. Doulamis, K. Konstanteli, P. Dellias, and T. Varvarigou, "An innovative content-based indexing technique with linear response suitable for pervasive environments," in *International Workshop on Content-Based Multimedia Indexing*, June 2008, pp. 462–469.

[9] R. Jakubowski, "Extraction of shape features for syntactic recognition of mechanical parts," *IEEE Trans. on Systems, Man and Cybernetics*, vol. SMC-15, no. 5, pp. 642–651, Sept 1985.

[10] M. Bielecka and M. Skomorowski, "Fuzzy-aided parsing for pattern recognition," in *Computer Recognition Systems 2*, ser. Advances in Soft Computing, M. Kurzynski, E. Puchala, M. Wozniak, and A. Zolnierek, Eds. Springer Berlin Heidelberg, 2007, vol. 45, pp. 313–318.

[11] A. Sluzek, "On moment-based local operators for detecting image patterns," *Image and Vision Computing*, vol. 23, no. 3, pp. 287 – 298, 2005. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2004.03.003

[12] H.-C. Lee and K.-S. Fu, "Generating object descriptions for model retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 5, pp. 462–471, Sept 1983.

[13] J. R. Ullmann, "An algorithm for subgraph isomorphism," *J. ACM*, vol. 23, no. 1, pp. 31–42, Jan. 1976. [Online]. Available: http://doi.acm.org/10.1145/321921.321925

[14] G. Aggarwal, T. Ashwin, and S. Ghosal, "An image retrieval system with automatic query modification," *IEEE Transactions on Multimedia*, vol. 4, no. 2, pp. 201–214, Jun 2002.

[15] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A sketch retrieval method for full color image database-query by visual example," in *11th IAPR International Conference on Pattern Recognition, Vol.I. Conference A: Computer Vision and Applications*, Aug 1992, pp. 530–533.

[16] A. Singh, S. Shekhar, and A. Jalal, "Semantic based image retrieval using multi-agent model by searching and filtering replicated web images," in *Information and Communication Technologies (WICT), 2012 World Congress on*, Oct 2012, pp. 817–821.

[17] C.-Y. Li and C.-T. Hsu, "Image retrieval with relevance feedback based on graph-theoretic region correspondence estimation," *Multimedia, IEEE Transactions on*, vol. 10, no. 3, pp. 447–456, April 2008.

[18] S. Kiranyaz and M. Gabbouj, "Hierarchical cellular tree: An efficient indexing scheme for content-based retrieval on multimedia databases," *Multimedia, IEEE Transactions on*, vol. 9, no. 1, pp. 102–119, Jan 2007.

[19] P. J. Sadalage and M. Fowler, *NoSQL distilled : a brief guide to the emerging world of polyglot persistence*. Upper Saddle River, NJ: Addison-Wesley, 2013. [Online]. Available: http://opac.inria.fr/record=b1135051

[20] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: Amazon's highly available key-value store," *SIGOPS Oper. Syst. Rev.*, vol. 41, no. 6, pp. 205–220, Oct. 2007. [Online]. Available: http://doi.acm.org/10.1145/1323293.1294281

[21] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *ACM Trans. Comput. Syst.*, vol. 26, no. 2, pp. 4:1–4:26, Jun. 2008. [Online]. Available: http://doi.acm.org/10.1145/1365815.1365816

[22] K. Sapiecha, G. Lukawski, and A. Krechowicz, "Enhancing throughput of scalable distributed two – layer data structures," in *Parallel and Distributed Computing (ISPDC), 2014 IEEE 13th International Symposium on*, June 2014, pp. 103–110.

[23] P. Sitek and J. Wikarek, "A hybrid framework for the modelling and optimisation of decision problems in sustainable supply chain management," *International Journal of Production Research*, 2015.

DOI: 10.15439/2015F197
ACSIS, Vol. 5

# Exploring users’ experience with e-reading devices

Chrysoula Gatsou
Faculty of Fine Arts and Design,
TEI of Athens
Athens, Greece
Email: cgatsou@teiath.gr

Anastasios Politis
Faculty of Fine Arts and Design,
TEI of Athens
Athens, Greece
Email: politisresearch@techlink.gr

Dimitrios Zevgolis
School of Applied Arts
Hellenic Open University,
Patra, Greece
Email: zevgolis@eap.gr

***Abstract*—New e-book readers and multifunctional mobile tablet devices are currently emerging, so bringing about a transition from printed to the electronic books. It is important to learn how usable these mobile devices are, by testing them on real users from various backgrounds. The paper presents a study which explores the perceived usability of two electronic reading devices, one dedicated reader and one multifunctional device. More specifically, the study employs eight tasks which users were required to complete within a specific time with two devices. Our results show that users functioned better on the multifunctional device in terms of performance measures, such as navigation, task difficulty and satisfaction.**

## I. Introduction

Electronic reading devices, including e-readers and tablet computers, are becoming popular so rapidly today, that it is vital to understand how the users themselves perceive the usability of such devices. An e-reader is designed for displaying electronic books, magazines, and periodicals. Various e-readers use various formats. For example, ibooks are produced for the iPad, whilst AZW, TXT and KF8 formats are mostly used by the Kindle. E-books are frequently available in PDF format, which is common for popular computer operating systems.

New opportunities for e-reading appeared, when lighter devices with better screens, such as the Amazon Kindle, appeared in late 2000 and the content available to devices increased at the same time. A vital question is why one should use an electronic reader, when one can use a net-book, tablet or a smart phone. The fundamental advantage of the electronic reader lies in its display. Thanks to the materials from which the display is manufactured, its appearance resembles that of paper, which means that the user does not have to strain his or her eyes when reading. Making objects more usable and accessible is part of the larger discipline of User-Centered Design, which employs various methods and techniques [1]. Usability testing is a method employed to evaluate a product by testing it on representative users. Greenberg and Buxton point out that “Usability evaluation is valuable for many situations, as it often helps validate both research ideas and products at varying stages in its lifecycle” [2]. The purpose of our study was to explore user experience with two e-reading devices. One was dedicated e-reader device, the Cybook Odyssey and the other was a multifunctional device, the Apple iPad. The main criterion for selecting these mobile devices was their availability on the Greek market at the time, November 2013. Despite the best efforts of designers, new technologies often fail to meet basic human needs and desires [3].

We start our paper with a review of the literature, which establishes the theoretical background to our study. We then describe the research methodology employed, discuss the results and offer some conclusions.

## II. Background

### A. E-readers

An e-reader, also called an e-book reader or e-book device, is a mobile electronic device designed primarily for the purpose of reading digital e-books. New opportunities for e-reading emerged when new lighter devices, such as the Amazon Kindle, that had better screens, appeared at the end of 2000 and the available content for devices increased at the same time. E-readers reproduces the appearance of a printed book.

Thus an e-reader should ideally offer readability, be able to host extensive texts, be portable, allow one to read anywhere and possess a long-lasting battery. It should also offer the ability to create bookmarks, to add notes on the book and to highlight passages as desired. The most important part of an e - reader is the screen. Unlike the majority of displays with which we interact on daily basis, paper is a reflective medium. As Zehner notes, a reflective display has an inherent advantage in terms of readability, because the brightness of the display naturally adapts to the ambient lighting conditions [4]. E-paper displays are expected to provide the user with a more paper-like reading experience that does not cause eye strain and that possess the contrast and reflection of real paper [5].

978-83-60810-66-8/$25.00©2015, IEEE

### *B. E-Ink technology*

E-Ink is a bistable display technology that creates a near-paper-like reading experience and requires minimal battery power [6]. Electronic ink is made up of millions of tiny microcapsules only 100 microns in width. Every microcapsule contains positively-charged white particles and negatively-charged black particles suspended in a clear fluid. When a positive or negative electric field is applied, the chips will either rise to the top or be pulled to the bottom, where they become visible to the viewer [7]. This makes the surface appear either white or black at the spot in question. Patterns of white and dark can then be created to form words and sentences (Fig.1). E-Ink displays do not need any background lighting and are easy to read, even in direct sunlight. Although E Ink technology devices require energy for turning pages, they do not consume much battery power. This means that a device, when fully charged, can be used for several thousand pages or several weeks.

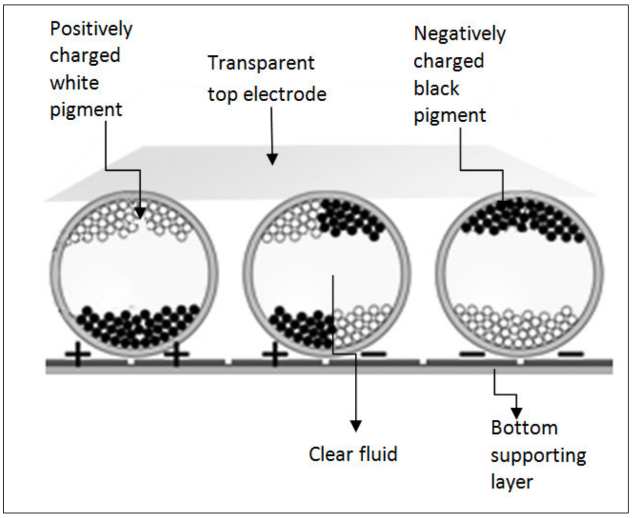

Positively charged white pigment
Transparent top electrode
Negatively charged black pigment
Clear fluid
Bottom supporting layer

Fig. 1 Scheme of electronic ink technology.

### *C. iPad*

The iPad is a tablet computer. Its basic technology differs from that of dedicated e-readers, in that it has a colour screen which does not employ the eye-friendly E-Ink technology and it is multi-functional. The iPad has been found to compete well with dedicated e-readers. Probably the most comprehensive iPad usability studies are those carried out by by Nielsen and Budiu in 2010 and 2011 [8,9]. Their findings are summarized below:

*Read-tap asymmetry*
Content was large enough to read, but too small to tap.
*Accidental activation*
This was a particular problem in apps lacking a back button.
*Too small touchable areas too close together*
This lead to accidental activation.
*Users disliked typing*
They thus avoided the registration process
*Splash screens*
A compulsory introduction screen bothers users.
*Information squeezed into too small areas*
This made the content harder to perceive and manipulate.
*Too much navigation*
The large number of navigation options gives one less space.

### *D. E-readers and Usability*

Nielsen did a within-subjects study employing 32 competent adult readers. The text was a Hemingway short story in several formats, namely printed book, personal computer, iPad and Kindle. The iPad gave a 6.2% lower reading speed than the printed book, whereas the Kindle gave a speed 10.7% slower than print [10]. However, the difference between the two devices was not statistically significant, because of the fairly high variability of the data.

Clark et al., reported that 36 Kindle users at the University of Texas thought that the limited content availability, the inconsistent pricing of titles, poor graphics resolution and other functions were barriers to the wider acceptance and use of the device [11].

### *B. Usability testing*

The term *"usability"* is frequently employed in the field of human-computer interaction (HCI). Nielsen describes usability as an issue related to the broader issue of acceptability [12]. In his view, *"Usability is a quality attribute that assesses how easy user interfaces are to use"*. Usability is a significant part of the user experience and therefore of user satisfaction. A formal definition of usability is given in the ISO standard 9241–11 : *"...the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction, in a specified context of use"*. Effectiveness is defined as the accuracy and completeness with which users achieve specified goals and efficiency as the resources expended in relation to the accuracy and completeness with which users achieve goals. Satisfaction is defined as the freedom from discomfort, and positive attitude to the use of the product, whilst the context of use is defined as users, tasks, equipment and the physical and social environments in which a product is used [13].

Usability testing is a method employed in user-centered design to evaluate product design by testing it on representative users. Such users thus yield quantitative and qualitative data in that they are real users performing real tasks [14].

Dumas & Redish argue that usability testing is a *"a systematic way of observing actual users trying out a product and collecting information about the specific ways in which the product is easy or difficult for them"* [15].

## III. RESEARCH METHODOLOGY

To examine how users conceptualize interaction with e-readers, we created a user test involving two e-reading devices. We compare one dedicated e-reader with one multifunctional device (Fig.2).

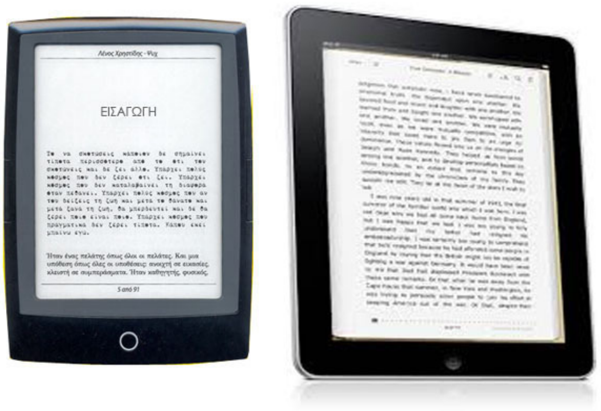


Fig. 2 The two mobile devices compared in the study: The Cybook Odyssey and the Apple iPad.

Nielsen [16] argues that five participants will discover 80% of the problems in a system. In any case, a small quantity of users, that is, generally fewer than 10 subjects, is sufficient, for any formative evaluation of usability [17]. On the other hand, Spool and Schroeder [18] state that five users identified only about 35% of the problems in a website. The research by Turner et al. implies that a group size of seven may be optimal, even when the study is fairly complex [19].

According to Sauro and Lewis “the most important thing in user research, whether the data are qualitative or quantitative, is that the sample of users you measure represents the population about which you intend to make statements” [20].

### A. Participants

Our session was designed specifically to include a representative pool of the potential users of e-readers. Twelve participants (N=12) aged between 18 - 65 (mean age = 36.08, SD = 14.47, years), seven of whom were males and five females, participated in the session. Using either a skilled participant or an under-qualified one will bias the outcomes of usability testing. All participants were novices as regards e-reading devices, but were fond of reading. On the other hand, all of them had used a digital device (e.g. PC or mobile phone) prior to this study, to read news, information or academic content online. The participants received short instructions on how to turn their device on and off, but were not instructed in how to operate it. This was designed to test the usability of the devices and to examine how intuitive the interface was for the participants. Participants did not suffer from any visual or cognitive impairment and were educated to at least high school level. Participants gave written informed consent prior to participation.

### B. Material

Two mobile devices, the Apple iPad, a multifunctional e-reader and the Cybook Odyssey, a dedicated e-reader, were compared in the study. Devices were chosen on grounds of anticipated availability on the Greek market. A Panasonic HDC-SD40 digital camera was used to create a complete record of all user interactions with the e-readers. Technical specifications for the mobile devices are provided in Table I.

TABLE I.
TECHNICAL SPECIFICATIONS OF THE TWO COMPARED DEVICES.

| **Thechnical Specifications** | **Cybook Odyssey** | **Apple iPad (3rd generation)** |
|---|---|---|
| Intro year | 2013 | 2012 |
| Display size | 6" | 9.7″ |
| Display type | I-ink HD | LCD screen |
| Rresolution | 758x1024 | 2048 x 1536 |
| Weight | 180 g | 652 g |
| Touch screen | Multi touch | Multi touch |
| Memory | 2GB iNAND+ MicroSDHC up to 32GB | 16GB |
| Content ebook formats | EPUB, PDF, Adobe DRM, HTML, TXT, FB2 | Ibooks, ePub, PDF |
| Content picture formats | JPEG, PNG, GIF, BMP, ICO, TIF, PSD | JPEG, GIF, TIF, PSD |
| Battery life | Up two weeeks | Up to 10 hours |

### C. User Tasks

For the usability test, the participants were required to read a segment of the text from an e-book on Greek mythology on the two e- reading devices and to complete eight tasks given in Table II. The tasks were chosen as being representative and as covering as many as possible of the features of the e-reading devices. Task success (whether or not a participant successfully completed a task) was recorded. Participants were allowed up to two minutes to complete each task.

TABLE II
PARTICIPANTS TASKS

| **Tasks** | **Task Description** |
|---|---|
| **Task 1** | **Open the book** “Mythology” |
| **Task 2** | **Go to page** 26 |
| **Task 3** | **Change the page** to landscape format |
| **Task 4** | **Highlight** the first two sentences of the page |
| **Task 5** | **Delete the highlighted** sentences |
| **Task 6** | **Make a note** of the first paragraph of the page |
| **Task 7** | **Increase the font size** |
| **Task 8** | **Add bookmark** |

### D. User Performance

User performance was recorded in terms of the effectiveness, efficiency and ease of use of e-reading devices. In order to evaluate task effectiveness, we measured the percentage of tasks successfully completed within the time limit. Task completion time refers to the time needed to accomplish the task. To evaluate efficiency, we recorded the time needed to process a task. To measure user satisfaction, we asked users to complete a post-test questionnaire SUS (system usability scale). A SUS questionnaire gives ten statements regarding different aspects of usability. Users mark their agreement with the statements on a five-point Likert scale.

### *E. Post -test Questionnaire*

The aim in administering a written questionnaire after the test (post-test questionnaire) is to record participants' preference, in order to identify any potential problems with the product. Information collected usually includes opinions and feelings regarding any difficulties encountered in using the product. Our questionnaire was based on the System Usability Scale (SUS) developed by Brooke [21], since this is the most precise type of questionnaire for a small number of participants, as is shown by Tullis and Stetson's study [22]. SUS employs a "quick and dirty" approach in evaluating the overall subjective usability of a system (Appendix A). While the SUS was originally intended to be used for measuring perceived usability, i.e. measuring a single dimension, recent research shows that it provides an overall measure of satisfaction of the system [22],[23],[24]. In addition to these advantages over other systems, the SUS is a powerful and multifunctional instrument [25].

### *F. Test protocol*

Participation in the study lasted approximately 40 minutes for each participant and was conducted in an isolated room in our Faculty. Participation consisted of the series of tasks mentioned above. Participants were informed for the process of the test and all participants were tested individually.

After being welcomed by the experimenter, participants were told that they were to take part in a usability test and were to interact with two e-reading devices. Participants were reminded to note the instruction for gestures on the desk on their left. In addition participants gave their permission to be recorded on video. Subsequently participants completed our eight tasks. Finally there was a question about what readers considered the most important features in eReaders

## IV. Results And Discussion

The main factors to be examined when testing usability are effectiveness, efficiency and user satisfaction. Effectiveness refers to how "well" a system does what it supposed to do. In order to evaluate task effectiveness, we measured the percentage of steps successfully negotiated within the time limit (2min). Efficiency refers to how quickly a system supports the user in what he wants to do. To evaluate efficiency, we recorded the time needed to process the task. Satisfaction refers to the subjective view of the system on the part of the user [1]. Qualitative and quantitative data were collected from each participant. Qualitative data included the participants' verbal protocol as recorded in video recording.

Problems of usability were identified and categorized. We also collected comments on e-reading devices and preference data and evaluations in the form of the SUS data questionnaire completed by the users after the test. Any user action that did not lead to the successful completion of a task we defined as error.

Table III
Tasks completion rates

| | Cybook | iPad |
|---|---|---|
| **Task1** | 11/12 | 12/12 |
| | **91%** | **100%** |
| **Task2** | 9/12 | 11/12 |
| | **75%** | **91%** |
| **Task3** | 8/12 | 9/12 |
| | **66%** | **75%** |
| **Task4** | 7/12 | 8/12 |
| | **58%** | **66%** |
| **Task5** | 7/12 | 11/12 |
| | **58%** | **91%** |
| **Task6** | 2/12 | 7/12 |
| | **16.7%** | **58%** |
| **Task7** | 9/12 | 10/12 |
| | **75%** | **83.3%** |
| **Task8** | 10/12 | 11/12 |
| | **83.3%** | **91%** |

### *A. Effectiveness.*

The percentage of users that manage to complete a task successfully thus becomes a measure of the effectiveness of the design. The number of errors made on the way to completing a task is an example of a performance measure [1],[12].

Errors were classified into two main categories, navigation errors and comprehension errors. Navigation errors occurred when partcipants did not move as expected. Comprehension errors occured when participants did not understand the design of the interface. (Table IV).

Table IV
Types of errors by e-reader device

| **Type of error** | **Cybook** | **iPad** |
|---|---|---|
| **Navigation** | 3 | 4 |
| **Comprehension** | 5 | 3 |
| **Total** | 8 | 7 |

### *B. .Efficiency - Task Completion Time*

Efficiency is a measure that is highly dependent on the time spent on completing the task. We recorded the total amount of time required to complete each task on mobile devices. Table V shows information on the mean time spent by the participants. The results thus indicate that participants spent more time on task completion when using the Cybook.

TABLE V
AVERAGE TIME IN SECONDS FOR COMPLETING ALL GIVEN TASKS

| Tasks | Task Description | Cybook | iPad |
|---|---|---|---|
| **Task 1** | **Open the book** "Mythology" | 29,7 | 24,2 |
| **Task 2** | **Go to page** 26 | 63,7 | 48,3 |
| **Task 3** | **Make a note** of the first paragraph of the page | 78,2 | 65,1 |
| **Task 4** | **Highlight** the first two sentences of the page | 67,3 | 52,6 |
| **Task 5** | **Delete the highlighted** sentences | 42,3 | 38,2 |
| **Task 6** | **Change the page** to landscape format | 27,2 | 14,1 |
| **Task 7** | **Increase the font size** | 35,6 | 17,2 |
| **Task 8** | **Add bookmark** | 48,7 | 29,7 |
| | **Average Time** | **49,1** | **36,2** |
| | **Standard Deviation** | **18,8** | **18,1** |

However with the respect of the amount of time that participants spend on Cybook was higher than iPad. The expectation was, however, that participants would interact more efficiently with a dedicated e-reader than with a multi-functional device when reading a book (Fig.3).

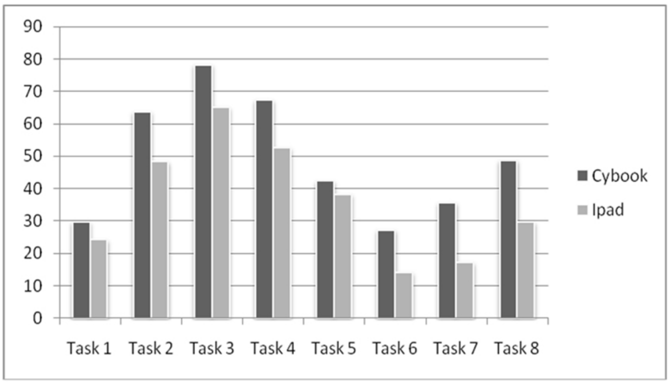


**Fig.3** Task completion time per e-reader device.

### *C. User satisfaction*

We. are aware that time-on-task measures can be useful for collecting data on the efficiency of a system. On the other hand, such data does not give any information on overall satisfaction on the part of the user. User satisfaction may be an important factor in motivating people to use a product and may affect user performance. Thus, as a final point, we decided participants should complete the System Usability Scale (SUS) questionnaire which has ten item attitude Likert-scale which measures the view of subjective assessments of usability and explore users' experiences with the two mobile devices. A crucial feature of the SUS lies in the fact that asks the user to evaluate the system as a whole, rather than specific aspects.

All 10 questionnaire statements having been processed, the overall SUS score for each prototype is that given in Table VI. To calculate the SUS score, first we summed the score contributions of items 1, 3, 5, 7 and 9 (Appendix A). The score contribution of these items are their scale position minus one. We then summed the score contributions of the other items: five minus their scale position. Finally, we multiplied the sum of the scores by 2.5, to obtain the overall score with a range between 0 to 100.

The study results showed the overall level of satisfaction. Sauro reports that a mean value over 74 is level B, value above 80.3 is level A [25]. An average value of below 51 is level F (fail). The iPad, which was given an average value of 79.3, is to be placed on level B, and Cybook, with an average value of 70.8, level B. It can be remarkable to notice that none of the e-Readers had an extremely high satisfaction score, meaning that users preferences were ambiguous.

TABLE VI
OVERALL SUS SCORE

| Participants | Cybook | iPad |
|---|---|---|
| P1 | 80.0 | 80.0 |
| P2 | 70.0 | 87.5 |
| P3 | 82.5 | 90.0 |
| P4 | 77.5 | 90.0 |
| P5 | 55.0 | 82.5 |
| P6 | 70.0 | 77.5 |
| P7 | 75.0 | 65.0 |
| P8 | 70.0 | 75.0 |
| P9 | 72.5 | 82.5 |
| P10 | 70.0 | 82.5 |
| P11 | 72.5 | 70.0 |
| P12 | 55.0 | 70.0 |
| **Mean** | **70.8** | **79.3** |

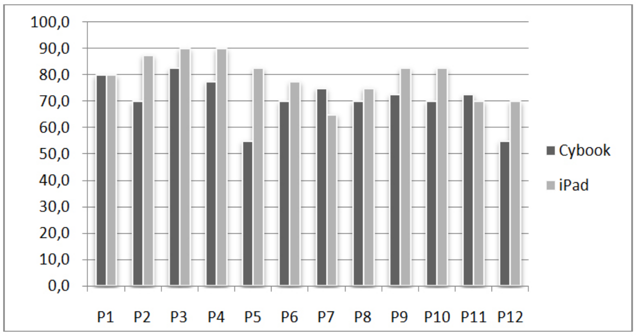


**Fig.4** Overall SUS score.

### *C. Overall user experience*

Overall users liked the process and regarded their interaction with the devices positively. Nevertheless, in some cases, the participants were apprehensive. Uncertain in their selections, they demanded greater confirmation and reassurance about the actions they were to take. In such cases, it is important for the researcher to motivate participants, encouraging them discreetly to investigate alternative directions, while simultaneously recording any mistakes made. Participants reported that the task in which one had to make a note was difficult to perform on the Cybook, as the keypad was complicated in comparison to that of the iPad.

Overall, participants had a preference for the iPad when highlighting and making notes. Additionally, the ability to change font size on the Cybook means that the device uses location numbers, rather than page numbers. The following conclusions can be drawn from the results derived from our group of participants, but one should avoid making any generalizations based on them. (Fig.4)

Participants complained of

- Poor page navigation
- Difficulty in turning the page
- Slowness in adapting to non-linear reading and
- The environment

On the other hand, participants reported that :

- In the case of the dedicated device (the Cybook), the menu allowed users to find some functions easily,
- older participants liked the ability to adjust font size,
- the portability and lightness of the devices was appreciated and that,
- when users read a narrative on an e-book sequentially from beginning to end, they found e-ink technology more friendly than the corresponding technology employed on the other device *(iPad)*.

### *D. Limitations*

The results of the study should not be generalized to all e-readers mobile devices. The small sample size (n = 12 ) limited the ability to acquire a more comprehensive view of the effects of various reading formats. Some e-readers are very similar in format to traditional print books. Reading on a very small screen device, however, like reading on a smartphone or online reading in a personal computer, with its links, multiple pages, and sometimes distracting graphics, raises various issues [26].

Another important reason was that participants read a narrative book, rather than a text offering information.

## V. Conclusion

The aim of our study was to explore user experience with two e-reading devices. One was a dedicated e-reader device, the Cybook Odyssey and the other was a multifunctional device, the Apple iPad. We tested our empirical methodology on twelve individuals, all of them novices in terms of e-reader use. The goal of our user study was to gain knowledge on the readability and usability of two different e-readers devices in a specific context. As a consequence of the small sample size (n=12), we evaluated our data on the basis of descriptive statistics. Eight tasks were selected in order to elicit user experience.

The results of the study show that both in terms of usabillity and overall impression, the iPad was the preferred device. This was somewhat surprising, as the lightness, long battery life and portability of the Cybook was greatly liked by the participants in the study. Additionally the testing material was a black and white book on Greek mythology and not a multi color magazine which is better for iPad use.

However, we feel that our paper, which focuses more on the users and their cognitive abilities, offers a new insight into how users perform tasks with e-readers devices and conveys their overall impressions. Long battery life and portability are advantages for any use, but the inability to facilitate easy browsing and navigation make the devices slow to use for any non-linear reading. In addition, from users' comments there emerged additional issues regarding digital rights management (DRM), and storage which should be explored in future studies. We believe that successful e-reading use depends on the integration between reading device, content providers and service platform.

## Appendix

Appendix **A** *System Usability Scale*

| | Strongly agree | | | | Strongly disagree |
|---|---|---|---|---|---|
| 1. I think I would like to use this device frequently | 1 | 2 | 3 | 4 | 5 |
| 2. I found the device unnecessarily complex. | 1 | 2 | 3 | 4 | 5 |
| 3. I thought the device was easy to use. | 1 | 2 | 3 | 4 | 5 |
| 4. I think that I would need the support of a technical person to be able to use this device. | 1 | 2 | 3 | 4 | 5 |
| 5. I found the various functions in this device were well integrated. | 1 | 2 | 3 | 4 | 5 |
| 6. I thought there was too much inconsistency in this device. | 1 | 2 | 3 | 4 | 5 |
| 7. I would imagine that most people would learn to use this device very quickly. | 1 | 2 | 3 | 4 | 5 |
| 8. I found the device very cumbersome to use. | 1 | 2 | 3 | 4 | 5 |
| 9. I felt very confident using the device. | 1 | 2 | 3 | 4 | 5 |
| 10. I need to learn a lot of things before I could get going with this device. | 1 | 2 | 3 | 4 | 5 |

## ACKNOWLEDGMENT

We express our gratitude to the participants, who enthusiastically gave their time to allow us to understand the value of their experiences.

## REFERENCES

[1] J.Rubin, and D. Chisnell, *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests* (2nd Ed.). Indianapolis, IN: Wiley Publishing, 2008.

[2] S. Greenberg and B. Buxton, Usability evaluation considered harmful (some of the time). Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human factors in Computing Systems, Florence, 2008, 111-120.

[3] M. Hassenzahl, The effect of perceived hedonic quality on product appealingness. International Journal of Human-Computer Interaction, 13(4), pp. 479-497, 2002.

[4] R. Zehner, "Electronic Paper Displays." In Mobile Displays: Technology and Applications (eds A. K. Bhowmik, Z. Li and P. J. Bos), John Wiley & Sons, Ltd, Chichester, UK , 2008.

[5] H. Heikkil␡a. eReading User Experiences: eBook Devices, Reading Software & Contents. Technical Report 54, NextMedia, 2011.

[6] DeJean, D. "The future of e-paper: The Kindle is only the beginning."Computerworld, June, 1, 2008. Retrieved from http://www.computerworld.com/article/2535080/ Accessed on 12/4/2014

[7] E Ink: http://www.eink.com. Accessed on 12 April 2014.

[8] R. Budiu and J. Nielsen. "Usability of iPad Apps andWebsites: 1st edition". Technical report, 2010.

[9] R. Budiu and J. Nielsen. "Usability of iPad Apps and Websites: 2nd edition". Technical report, 2011.

[10] J. Nielsen, "iPad and Kindle reading speeds", 2010 Retrieved from http://www.useit.com/alertbox/ipad-kindle-reading.html

[11] D. T. Clark, S. P. Goodwin, T. Samuelson, C.Coker, "A qualitative assessment of the Kindle e-book reader: results from initial focus groups", Performance Measurement and Metrics, Vol. 9 Iss: 2, pp.118 – 129, 2008.

[12] J. Nielsen and R. L. Mack. Usability inspection methods. Wiley, New York, NY, USA, 1994.

[13] ISO 9241-11 Ergonomic requirements for office work with visual display terminals (VDTs)-Part 11, Guidance on usability, ISO, 1998.

[14] A.Cooper, R. Reimann, and D. Cronin, About Face 3: The Essentials of User Interface Design. John Wiley & Sons, Inc. 2007.

[15] J.S Dumas and J.C Redish, A Practical Guide to Usability Testing (revised Ed.). Portland, Oregon: Intellect Books, 1999.

[16] J.Nielsen, "Why You Only Need to Test With 5 Users". Jakob Nielsen's Alertbox, March 19, 2000.

[17] H. Petrie and N. Bevan. The evaluation of accessibility , usability and user experience *In: The Universal Access Handbook, C Stepanidis (ed), CRC Press*, pp 299–315, 2009.

[18] J. Spool and W. Schroeder " Testing web sites: Five users is nowhere near enough", In: Proceedings of the Conference extended abstracts on Human Factors in Computing Systems, CHI'2001. New York: ACM Press, 2001.

[19] C. W. Turner, J. R. Lewis, and J. Nielsen, "Determining usability test sample size". In *W. Karwowski (ed.), International Encyclopedia of Ergonomics and Human Factors* Boca Raton, FL: CRC Press, 2006, pp. 3084-3088, 2006.

[20] J. Sauro and J.R. Lewis, *Quantifying the user experience: Practical statistics for user research*. Burlington, MA: Morgan Kaufmann, 2012.

[21] J. Brooke, SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), Usability Evaluation in Industry (S. 189 -194). London: Taylor and Francis,1996.

[22] T. Tullis, and J. Stetson, "A comparison of questionnaires for assessing website usability," In Proc. of the Usability Professionals Association (UPA), pp. 7–11, 2004.

[23] J. R. Lewis, and J. Sauro, "The factor structure of the system usability scale", Proc. Human Computer Interaction International Conference (HCII 2009), San Diego, CA, pp. 94–103, 2009.

[24] J. Sauro, "Does prior experience affect perceptions of usability?" Available:http://www.measuringusability.com/blog/prior-exposure.php, January 19, 2011 [Nov.15, 2012]

[25] J. Sauro, *A practical guide to the System Usability Scale (SUS): Background, benchmarks & best practices.* Denver, CO: Measuring Usability LLC, 2011.

[26] Leu, D. J., McVerry, J. G., O'Byrne, W.I ., Kiili, C., Zawilinski, L., Everett-Cacopardo, H., & Forzani, E.. The new literacies of online reading comprehension: Expanding the literacy and learning curriculum. Journal of Adolescent & Adult Literacy, 55, 5–14, 2011.

[27] P Lam, SL Lam, J Lam, C McNaught, "Usability and usefulness of eBooks on PPCs: How students' opinions vary over time,".*Australian Journal of Educational Technology*, 25(1), pp. 30–44, 2009.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 841–847
DOI: 10.15439/2015F202
ACSIS, Vol. 5

# Augmented Reality Using Optical Flow

Konrad Koniarski
System Research Institute,
Polish Academy of Science,
ul. Newelska 6, 01-447 Warsaw, Poland
Email: konrad.koniarski@gmail.com

***Abstract*—The paper deals with the application of Lucas-Kanade optical flow algorithm to develop an augmented reality (AR) system. Merging of a live view of the physical real world with context-related computer-rendered images to create a mixed image is a challenging problem. A virtual object has to be located in the correct pose and position in real time and perspective. Besides the occlusion problem need to be taken into consideration.**

**In the paper a computer-vision based method for AR systems based on the fiducial marker matching is proposed. For simplicity black square was used as the marker. This method consists of two main steps. The initial step uses Hough's transformation to detect the marker initial position and to select the marker tracked points. In the second step for each image frame these selected points are being tracked using Lucas-Kanade optical flow method. The positions of the selected points are used for calculating the pose and position of a virtual object. Unlike existing method proposed system using optical flow to increase speed performance. The examples of AR applications using the proposed algorithm are provided and discussed.**

## I. Introduction

THE Augmented Reality (AR) is a technology supplemented people's perception of real world by using virtual world. It consists in the integration of the computer generated virtual information such as sound, video or graphics with the real-world environment. This technology is characterized by using real-virtual combinations, 3D registration as well as the real-time interaction. AR interfaces superimpose in real-time textural or pictorial virtual information onto real-world scenes registered in 3D and allow users to interact with real and virtual element simultaneously. The technical challenges in AR technology lie in determining in real-time which image elements should be shown, where and how. Since AR proposes that the user has not to be able to distinguish between real and virtual information it requires that the virtual elements show both geometric as well photometric consistency with the real part of the image. Geometric consistency includes correct placement and size of the virtual element as well as the identification of occlusions. Photometric consistency means suitable shadowing, mutual reflections, chromatic adaptation to scene illumination. The solution of these problems even under simplified assumptions is not trivial.

AR technology appeared in the 1990s. It finds a wide variety of applications like education, entertainment, architecture, medical, machinery manufacturing and maintenance, military, regional planning and many other fields. The implementation of AR technology significantly depends on the used devices and/or user interfaces. The main devices for AR are head mounted or handheld displays, wireless wristbands or gloves as input devices, digital cameras or wireless sensors as tracking devices and powerfull CPU with considerable amount of RAM memory to process camera images. The appropiate techniques for intuitive interaction between the user and the virtual content of AR applications are one of the most important aspects of AR. These techniques include tangible, collaborative, hybrid or multimodal AR interfaces.

Correctly positioning of virtual objects relative to the real enironment is called registration. AR image registration uses different computer vision methods consisting usually from two steps: tracking and reconstruction or recognizing. Tracking is based on the detection in the camera images fiducial markers, optical images or other interest points. To interpret camera images tracking can make use of feature or edge detection methods or other image processing methods. Most of the available tracking techniques can be separated into two classes: feature-based and model-based methods [18]. Feature based methods consists in discovering the connection between 2D image features and their 3D world frame coordinates. Model based methods make use of model of the tracked object's features such as CAD models or 2D templates of the items based on distinguishable features. Once a connection is made between 2D image and 3D world frame it is possible to find the camera pose by projecting the 3D coordinates of the feature into the observed 2D image coordinates and by minimizing the distance to their corresponding 2D features. The constraints for camera pose estimation are most often determined using point features. The reconstruction or recognizing stage uses the data obtained from the firts stage to reconstruct a real world coordinate system. There are different tracking technologies available, including among others optical, magnetic, movemnent, ultrasound sensors or thermal imaging. They capture features from the real world and based on this information the AR system determines when, where and how the virual scene should be merged with a real one.

Due to low cost, accuracy as well as robustness requirements optical tracking is mostly used in AR applications. Two types of optical tracking are usually used: marker based and markerless. The first approach makes use of known artificial patterns, called markers, placed along the real environment in order to perform camera pose estimation. Markers are designed to be easy to recognize and the camera matrix can be recovered from the detected marker pose. The position of the virtual object is

978-83-60810-66-8/$25.00©2015, IEEE

computed using a camera matrix and a detected marker pose for correct virtual object projection. Marker based tracking is a more established approach for registration [15]. Marker based AR consists in the identification of their location with respect to camera as well as in the calculation of the projection and transformation matrices used to correct positioning and integrating of the virtual object with the real environmemt. The position of the virtual object is calculated based on the marker information. On the other hand markerless tracking differs from the former one by the method to place virtual objects in the real scene. In markerless AR natural objects in the real scene are used as markers. This approach exploits natural features of the real environment to perform tracking. Since markerless AR is based on natural features rather than on fiducial markers there are no ambient intrusive markers that are not really part of the world. Moreover markerless AR counts on specialized and robust trackers as well as provides possibility of extracting from the surroundings characteristic information that may be later used by the AR system for other purposes. Markerless AR is subject of intensive research in the last years and still presents important challenges to be overcome [18].

## II. Related Work

In this paper marker based AR is adopted to estimate the pose and position of the virtual object. Marker based AR has been investigated by many authors [3], [4], [17], [18], [19], [20], [21], [22], [23], [24]. The most frequently used method has been proposed by Kato in [3] where the matching technique for detection and tracking the square marker is developed. Kato's method has been implemented in the form of computational library ARToolKit. In ARToolKit marker position is computed for every frame of video sequence and that lower real time performance of Kato's method. Since Kato's method appeared sensitive to perturbations it has been improved in [4] where vision-based corner tracker has been used. Paper [23] presents many different types of fiducial markers in AR applications designed for having many image features that are easy to track by optical flow alghoritms [4]. Eleven criterias for robustness and usefulness of marker are indicated where the most important are: false positive recognition, intermarker confusion, immunity to lighting connditions, immunity to occlusion and speed performance. New types of markers used for AR are proposed in [24]. For details concerning the construction and the application of markerless methods see [18]. Different aspects of using AR are indicated in [19], [20], [21], [22]. Paper [14] is concerned with the developement of AR environment without specialized hardware or pre-calibration. AR environment is built using SURF method combined with bi-directional optical flow alghoritm. In [14] Herakleous estemate SURF method as not sufficent for real time performance giving only 10 frames per second. Applying bi directional optical flow Herakleouts improve performance of SURF for 50 percent. Paper [14] is not dealing with immunity of linght conditions and immunity of occlusion.

This paper deals with the development of a marker based approach to AR. A new hybrid algorithm combining camera calibration, the selection of characteristic points on the marker surface as well as tracking the movement of these points using the optical flow technique and the estimation of the position pose of the virtual object is proposed and tested. As a initial step the camera calibration process is performed using moving plane techniques [8]. That step is done only once. Next step is to find marker position. Hough transformation [16] is used for the detection of the marker pattern features. For simplicity marker containt only one black square, but any recognizable pattern could be applied. Next block of presented method is tracking loop. A key elememt of the algorithm is the application of Lucas Kanade optical flow method [2] to track the movement of the selected marker points. Finally the pose of the virtual object is estimated using the updated camera projection matrix. The proposed algorithm has been programmed in C++ enviroment and tested. The obtained results are provided and discussed. They indicate that the proposed algorithm is robust and efiicient tool in realizing AR applications.

## III. Proposed Algorithm

The proposed approach consists from two steps. At initial step camera is calibrated and marker is detected. In main part selected marker points are tracked using Optical Flow method and virtual object pose and position are calculated. Furthermore virtual object is rendered in front of image from camera. System overview diagram is presented on Fig. 1.

### A. Camera calibration

The aim of camera calibration is to determine the extrinsic and intrinsic parameters of the transformation between an object in real 3D space and the 2D image observed by the camera based on visual data acquired from the images. The extrinsic transformation parameters includes the rotation and the translation matrices of the camera. The coordinates of the principal point, the scale factors along the axes and the skew of the two image axses are called intrinsic camera parameters. These parameters, i.e., six extrinsic and 5 intrinsic are assembled in the form of the $3 \times 4$ camera projection matrix.

In this paper Zhang method based on moving planes technique is used [8] to calibrate and rectify the camera. This algorithm consist of the following key steps:

1) The printed pattern is attached to a planar surface.
2) A few images is of the model plane is taken under different orientations due to the movement either the camera or the plane.
3) The feature points in the images are detected.
4) Estimate all the parameters using the homography between the model plane and its image.
5) Estimate the coefficients of the radial distortion by solving the linear least-square problem.

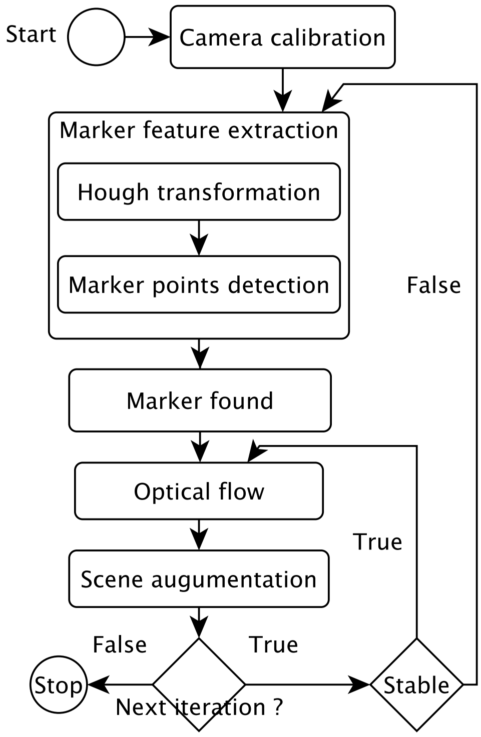


Fig. 1: System overview diagram.

6) Refine all parameters including the lens distortion parameters by minimizing the distances between the image points and their predicted positions.

Remark the pattern could be anything as long as we know the metric on the plane. If pixels are square the minimum number of orientations is two but for better quality it should reach 4 or 5 different orientations. Either camera or the planar pattern can be moved. There is no need to know the motion but a pure translation should be excluded.

Figure 2 presents detection of marker pattern while Figure 4 shows rectified image.

### B. Tracking points selection

Since we confine to the fiducial marker based AR let us choose a pattern of a marker. In literature [3], [17] there exists many examples of marker selection.A marker that is simpe black square was chosen. The black square corners are perfect candidates for tracking points. For that alghoritm any recognizable marker pattern can be used, but that required different method of selecting tracking point then applied Hough trasformation. For details see Fig. 5. In second step of our algorithm our aim is to select the characteristic points of the fiducial marker allowing to determining its initial position and tracking it in the optical flow step.

First having acquired marker image we determined its edges using method of [10]. Next Hough method [16] is used to detect square edges characterizing the marker. Hough transform is recognized as a powerful tool to extract parametrized curves. The idea behind the method is simple: parametric shapes in an image are detected by looking for accumulation points in the parameter space. If a particular shape is present in the image then the mapping of all of its points into parameter space must cluster around the parameter values which correspond to that shape. Usually this method consists from the following phases: voting, peaks localization, determining the actual parameters and verification. In our case black square generate 4 lines.I assume that in the initial position of the marker these lines are perpendicular. For our marker these set of lines is displayed in Fig. 5. Having known the parameters of each line I solve the system of linear equations to find their intersection points. Among these intersection points I choose square corners. These points are stable for Lucas-Kanade alghoritm [2], i.e., such points where the image function takes minimum and maximum values with respect to black and white colours, respectively. The selected points are feature points, deployed

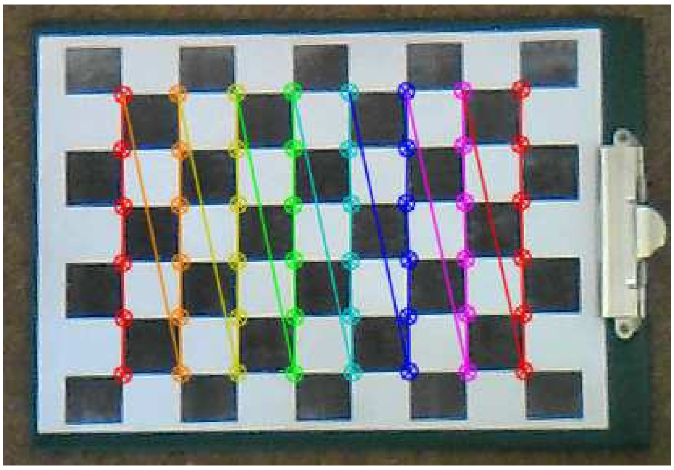


Fig. 2: Camera calibration using marker pattern and its geometrical representation (spherical distorition).

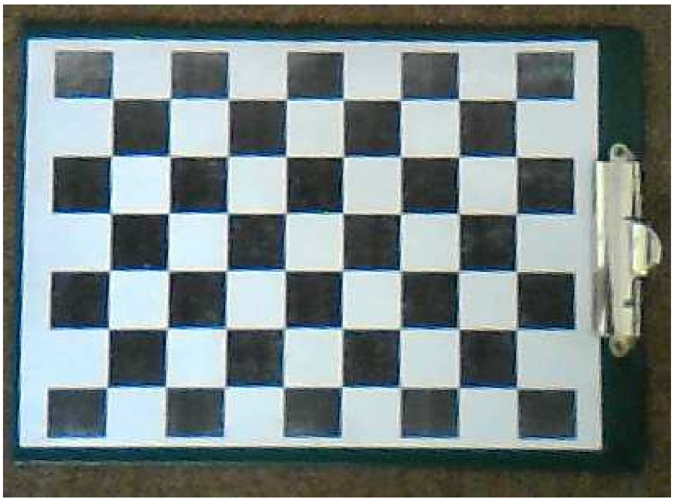


Fig. 3: Calibrated camera image (image without distortion).

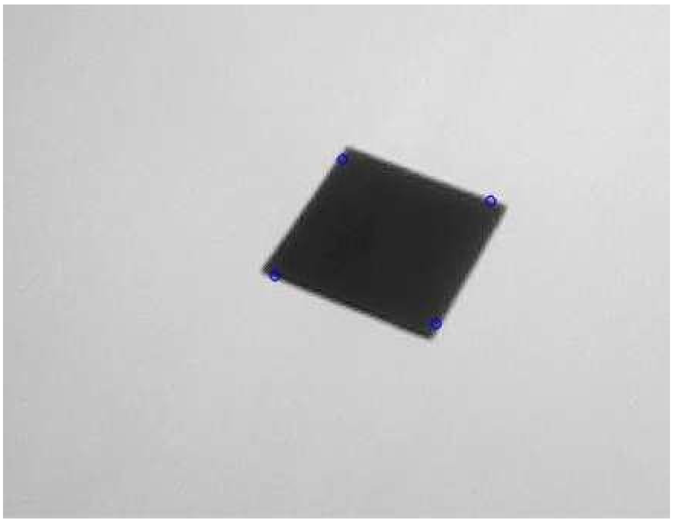


Fig. 4: Marker pattern.

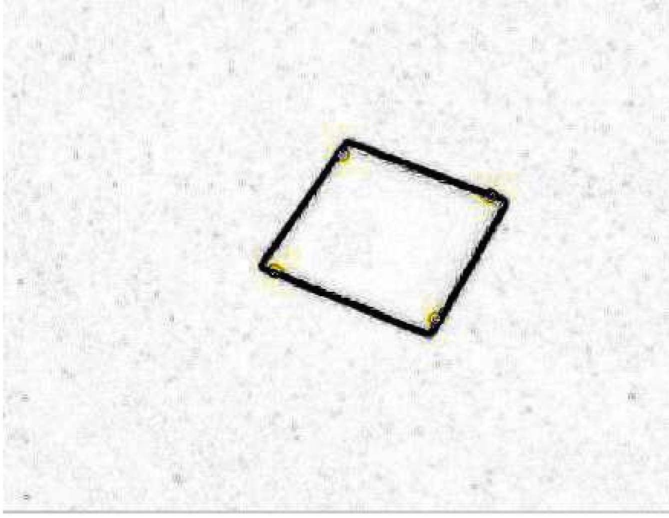


Fig. 5: Marker detection. Black lines are detected by Hough line detector.

at squares corners, which improve tracking stability during the next optical flow step.

### C. Lucas-Kanade Optical Flow

The core area of computer vision is the analysis of the sequence of images to approximate motion of 3D objects. Optical flow field is a convenient and useful image motion representation. It is used to approximate the movement of 3D objects based on a sequence of 2D images [12]. I expect that optical flow is not too different from the motion field. The optical flow calculations are based on two main assumptions: the image brightness constancy and the small motion. The first assumption means that while the object may change the position in short time interval the reflectivity and illumination will remain constant. The second assumption means that the movements less than one pixel are considered. There are numerous methods to calculate optical flow. These methods include differential-based methods, frequency-based methods, correlation techniques, multiple motion methods, feature based method, as well as hierarchical approaches [12]. Optical flow approach is widely used for object tracking, camera stabilization, image mosaics, 3D shape reconstruction or obtaining a structure form motion. Since tracked features points can move in any direction or be obscured by other scene objects, tracking is a chalenging investigation problem.

The optical flow methods are very often used in Augmented Reality applications to track the movement of the extracted features points [2], [14], [15]. Lucas and Kanade's local differential algorithm is one of the most popular method for optical flow computations [13]. This method is based on the assumption that in a given pixel the flow vector will be similar to a small neigbourhood surrounding this pixel. A weighted least square method is used to approximate the optical flow at a given pixel. For each pixel an optical flow vector consistent with the neighbouring spatial and temporal gradients is found. This method is very popular these days, but it still has some drawback. The first disadvantage of this method is challenging computational problem and result image is not dense. Only selected points can be computed for the real time applications. The selected points for tracking can't be placed on edges or low textured regions where this method doesn't give stable results. The best candidates for tracking are points in high textured regions. Lucas-Kanada's optical flow method use correlation window for tracking near pixel movement. For point movement tracking Gaussian pyramid with different scales of images is used [9].

The Lucas-Kanade's Optical Flow algorithm [2] is used for tracking the marker feature points in camera captured images. This method is faster than detection image features analyzing all pixels. Consider small window $\Omega \subset D \subset R^2$ occupied by the pixels $\mathbf{p}$ of an image $D$. The pixel image $p$ is characterized by $p = [x, y]$ where $x, y$ denotes coordinates of a point in $\Omega$.I denote the intensity of the image at position $(x, y)$ and at time $t$ by $I(x, y, t)$ as well as the velocity $v$ of the image pixel $p$ by

$$\dot{p} = v = (v_x, v_y) = (\frac{dx}{dt}, \frac{dy}{dt}) \tag{1}$$

Based on the brightness constancy assumption of the pixel $p$ during the time increament $dt$ I can write

$$I(x + v_x dt, y + v_y dt, t + dt) \approx I(x, y, t) \tag{2}$$

If the brigthness changes smoothly with $x, y, t$ I expand the left-hand side of (2) by a Taylor series and obtain

$$\begin{aligned} I(x + v_x dt, y + v_y dt, t + dt) = I(x, y, t) + \\ \frac{\partial I}{\partial x} v_x dt + \frac{\partial I}{\partial y} v_y dt + \frac{\partial I}{\partial t} dt + O(dt^2), \end{aligned} \tag{3}$$

where $\frac{\partial I}{\partial x}$ denotes the first derivative of the function $I$ with respect to $x$ and $O(dt^2)$ denotes the higher order terms. Due to the small motion assumption I ignore in (3) the higher order terms. Substitutting (3) into (2) I obtain optical flow constraint equation

$$\nabla I \cdot v + I_t = 0 \tag{4}$$

where $\nabla I = (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y})$ denotes a spatial gradient, $I_t = \frac{\partial I}{\partial t}$ is a temporal gradient and $v = (v_x, v_y)$ denotes the optical flow vector.

Remark, for each pixel, the equation (4) provides only a single constraint for the two components of the optical flow vector $v_x$ and $v_y$. Only the component of the vector in the direction of the spatial gradient is provided. It means that I cannot determine optical flow uniquely only from such optical flow constraint equation.We need some other constraints.

Lucas-Kanade's local differential technique is one of the most popular methods to calculate the optical flow. This method assumes that the flow is essentially constant in a local neighborhood of the pixel under consideration and solves the basic optical flow equation (4) for all the pixels in that neighborhood. Therefore we obtain the system consisting from many equations (4) where the number of unknowns is lower than the number of equations. The least square method is used to find a solution to this system of equations. The optical flow is a field which describes the movement of pixels between two frames, often used for image sequences like movies. It often forms the lowest level for motion analysis and is used as input to higher level systems like segmentation and tracking for interpretation. This system follows the same approach. The optical flow field is determined for each frame of the movie and is used to segment the frame into separate regions. These regions are fed as input to an object tracking system which determines information about the regions in the image and matches them with regions seen in previous frames. This allows us to track various objects through the scene.

Let $\Omega$ denotes a surrounding neighbourhood of a pixel $[x, y]$ under consideration of size $N$ where each neighbour pixel is represented as $p_i$ and $\mathbf{p} = (p_1, ..., p_N)$. Let us define a window function $W(\mathbf{p})$, $\mathbf{p} \in \Omega$. We let the weight of the center bigger than others, i.e., the window function favors the center. Usually $W(\mathbf{p})$ is assumed to be Gaussian window. The size of the window $W$ is $5 \times 5$, i.e., usually $N = 25$ pixels are used. Let us introduce the error function for the optical flow vector $v$ for a pixel $[x, y]$

$$E(v) = \sum_{\mathbf{p}\in\Omega} W^2(\mathbf{p})(\nabla I(\mathbf{p}) \cdot v + I_t(\mathbf{p}))^2 \tag{5}$$

This function simply sums the error of applying the flow vector $v$ to the spatial and temporal gradients of all surrounding neighbours usin the optical flow equation (4). The error is higher if the consider flow vector is not consistent with the spatial and temporal gradients for the surrounding pixels. The weights are used to diminish the importance of distant neighbours. We are looking for $v$ minimizing the error function (5), i.e.,

$$\min_v E(v) = E(v^\star) \tag{6}$$

Differentiating the error function (5) with respect to $v_x$ and $v_y$ we get the following necessary optimality condition to the minimization problem (6)

$$\frac{\partial E}{\partial v_x} = \sum_{\mathbf{p}\in\Omega} W^2(\mathbf{p})(\frac{\partial I}{\partial x}(\mathbf{p})v_x + \frac{\partial I}{\partial y}(\mathbf{p})v_y + \frac{\partial I}{\partial t}(\mathbf{p})\frac{\partial I}{\partial x} = 0 \tag{7}$$

$$\frac{\partial E}{\partial v_y} = \sum_{\mathbf{p}\in\Omega} W^2(\mathbf{p})(\frac{\partial I}{\partial x}(\mathbf{p})v_x + \frac{\partial I}{\partial y}(\mathbf{p})v_y + \frac{\partial I}{\partial t}(\mathbf{p}))\frac{\partial I}{\partial y}(\mathbf{p}) = 0 \tag{8}$$

Let us introduce notation:

$$W = diag(W(p_1), ..., W(p_N))_{N\times N}, \quad v = [v_x, v_y]^T \tag{9}$$

$$A = \begin{bmatrix} \frac{\partial I}{\partial x}(p_1) & \frac{\partial I}{\partial y}(p_1) \\ \vdots & \vdots \\ \frac{\partial I}{\partial x}(p_N) & \frac{\partial I}{\partial y}(p_N) \end{bmatrix}_{N\times 2} \tag{10}$$

$$b = -\begin{bmatrix} \frac{\partial I}{\partial t}(p_1) \\ \vdots \\ \frac{\partial I}{\partial t}(p_N) \end{bmatrix}_{N\times 1} . \tag{11}$$

Using this notatiotion the system (7)-(8) can be written in the matrix form

$$A^T W^2 A v = A^T W^2 b \tag{12}$$

So, assuming the matrix

$$A^T W^2 A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}_{2\times 2} \tag{13}$$

$$\begin{aligned} a_{11} &= \sum_{\mathbf{p}\in\Omega} W^2(\mathbf{p})(\frac{\partial I}{\partial x}(\mathbf{p}))^2 \\ a_{12} &= \sum_{\mathbf{p}\in\Omega} W^2(\mathbf{p})\frac{\partial I}{\partial x}(\mathbf{p})\frac{\partial I}{\partial y}(\mathbf{p}) \\ a_{21} &= \sum_{\mathbf{p}\in\Omega} W^2(\mathbf{p})\frac{\partial I}{\partial x}(\mathbf{p})\frac{\partial I}{\partial y}(\mathbf{p}) \\ a_{22} &= \sum_{\mathbf{p}\in\Omega} W^2(\mathbf{p})(\frac{\partial I}{\partial y}(\mathbf{p}))^2 \end{aligned}$$

is invertible, the flow $v$ for the image pixel $p$ equals

$$v = (A^T W^2 A)^{-1} A^T W^2 b \tag{14}$$

### D. Pose estimation of the augmented object

In the last step of the proposed method the projection matrix of the augmented object is computed using the approach from section III-A. However here this matrix projects [8] 3D virtual object points positions on the 2D image plane in global coordinates. First based on the intrinsic camera parameters and the homography matrix the estimations of the rotation and translation matrices as well as the projection matrix are calculated. The homography matrix is calculated using

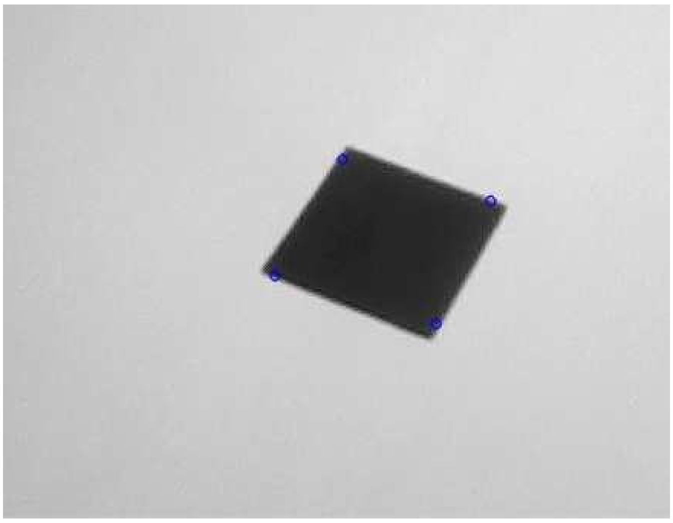


Fig. 6: Blue circles marked tracked marker points.

maximum likelihood criterion and the coordinates of the model and image points. For details see [8]. Next the components of the projection matrix are refined by minimizing the square of the Euclidean norm distance between the positions of the geometric reference points on the marker and the positions of the tracked marker points.

## IV. Results

The algorithm has been implemented in C/C++ and MinGW environments. Moreover OpenCV 2.4 development library was used for image processing and manipulations [11]. Then OpenGL library is used for rendering 3D objects and composing it into real scene. All tests were run on 8GB RAM machine equipped with Intel i7 2640M 2.80GHz processor. The processed images have 640 by 480 pixels resolution.
The marker presented in Fig. 4 was printed and displayed on flat moving surface. Webcam was static. Results are presented on Fig. 7. Right part of images show tracked marker, left part present marker and the virtual object - standard openGL teapot. All of those images presents marker from different perspective and proper projection of virtual object.

## V. Evaluation

For evaluation of proposed method five criterias were used: false/positive recognition, marker confusion, immunity of light conditions, immunity of occlusion and speed performance.

1) False/positive recognition: This criteria indicate sensitiveness for not recognize marker. First step of the presented method detects marker tracking points and passing marker points as input for second stage. If second stage loose it's stability in tracking marker points methods fallback to detection stage. This robust solution works stable until brightness constraint is violate or view angle is to acute.
2) Marker confusion: indicates sensitiveness if marker is recognized correctly. In presented case marker have four symmetric points any of them can be used as first feature to track.
3) Immunity of light conditions: Lucas-Kanade method use brightness constraint to track pixel movement between two frames. This constraint is also required for presented method. Selected marker has big contrast, so small changes in image brightness not disturb stability of presented method.
4) Immunity of occlusion: Optical flow is sensitive on occlusion. When tracking point is vanished by to acute angle or by other item occlusion Lucas-Kanade method is not stable and observed point usually move to other local minimum. Proposed method using only 4 points to track marker and if any of them is not tracked correctly methods loose stability. That triggers step back to marker detection block.
5) Speed performance: One of the most important criteria

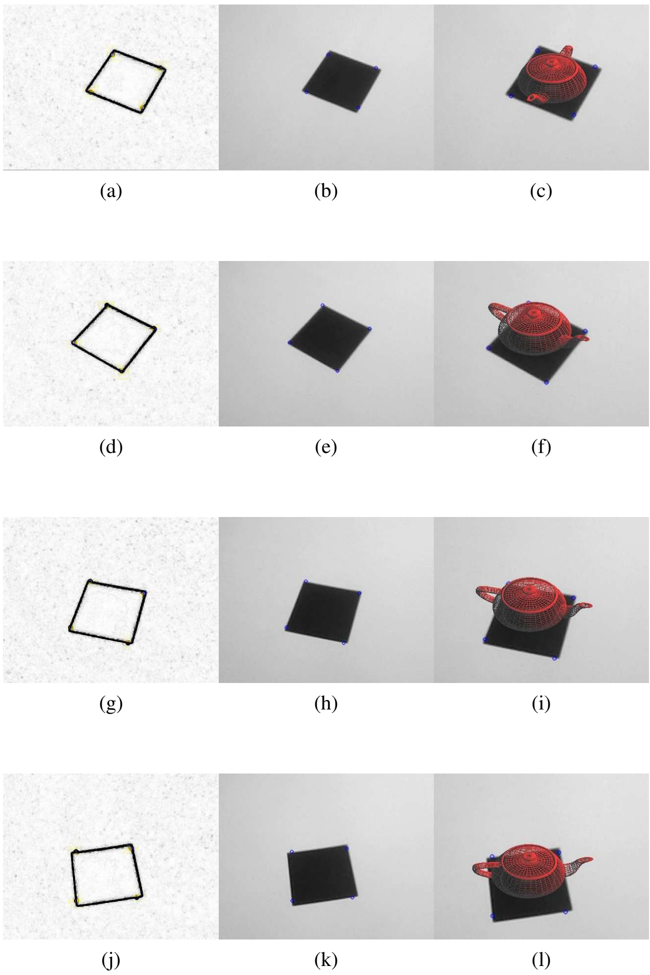


Fig. 7: Augmented virtual image - First column - marker detection. Middle column marker points tracking. Right column augumented virtual image

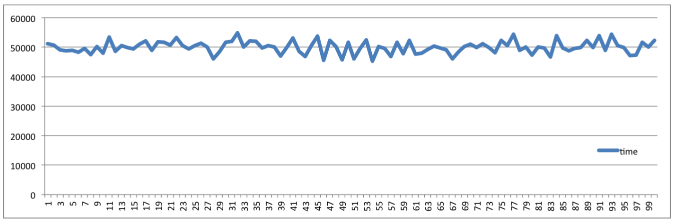


Fig. 8: Time in microseconds.

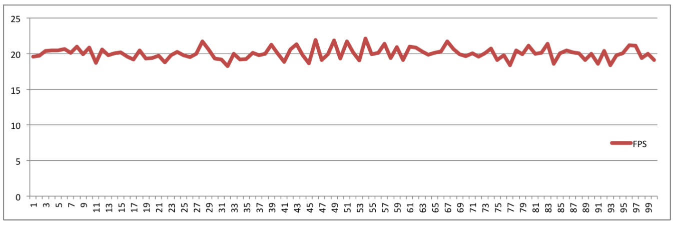


Fig. 9: Frames per second.

for AR is real time performance. Figure 8 presents time in microseconds to render each frame for 100 frames video. Figure 9 showing that system works with average 20 frames per second. That is comparable with real time performance.

## VI. Conclusion

This paper presents a new approach for Augmented Reality system based on Optical Flow method. The presented method uses a marker for obtaining the position, scale and pose of the virtual object. The proposed method runs in real time. The marker is detected correctly when there is no brightness change between image frames. All part tracking point from the marker must be visible during tracking otherwise method become unstable. The performed experiment shows that method is working correctly when marker was moved.

Using Optical Flow method for Augmented Reality is considered by the author as the natural step forward in combining marker and markerless method, and it will certainly become much popular in the near future.

## VII. Future work

As presented methods gives good results in Augmented Reality application, the next step should improve alghoritm stability when one of tracked points is obscured. This method is also sensitive to changing light conditions. Future work should improve presented mehtod to be less sensitive for changing light conditions.

## References

[1] Chari V, Singh J M, Narayanan P J, (2008) Augmented Reality using Over-Segmentation, National Conference on Computer Vision Pattern Recognition Image Processing and Graphics.

[2] Lucas B D, Kanada T (1981) An Iterative Image Registration Technique with an Application to Stereo Vision, IJCAI, 81, 674–679.

[3] Kato H, Billinghurst M (1999) Marker tracking and HMD calibration for a video-based augmented reality conferencing system, Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR'99), 85–94.

[4] Malik S, Roth G, McDonald C (2002) Robust Corner Tracking for Real-Time Augmented Reality, Vision Interface 2002, Calgary, Alberta, Canada, May 2002, 399-406. (National Research Council of Canada, Report No 45860).

[5] Koniarski K (2011), Image features detection methods in multiframe analysis(in Polish), in: Techniki informacyjne: teoria i zastosowania, eds: J. Hołubiec, 1(13), 68–82.

[6] Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: An efficient alternative to SIFT or SURF, , 2011 IEEE International Conference on Computer Vision (ICCV), 2564–2571.

[7] Klein G, Murray D (2009) Parallel Tracking and Mapping on a Camera Phone, 8th IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2009), 83-86.

[8] Zhang Z (2000) A Flexible New Technique for Camera Calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11), 1330-1334.

[9] Bouguet J Y (2001) Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm, Intel Corporation, 5.

[10] Sun D, Roth S, Lewis J P, (2008) Learning Optical Flow, Computer Vision - ECCV 2008, Springer, 83-97.

[11] Computer Vision Library: OpenCV (2012) http://opencv.willowgarage.com version 2.4.

[12] Beauchemin S S, Baron J L (1995) The computation of optical flow, ACM Computing Surveys (CSUR), 27(3), 433–466.

[13] Baker S, Scharstein D, Lewis J L, Roth S, Black M, Szeliski M (2011) A database and Evaluation Metodology for Optical Flow, International Journal of Computer Vision, 92(1), 1-31.

[14] Herakleous K, Poullis C H (2013) Improving augmented reality applications with optical flow, IEEE International Conference on Image Processing 2013, 3403-3406.

[15] Li H, Qi M, Wu Y (2012) A Real-Time Registration Method of Augmented Reality based on SURF and Optical Flow, Journal of Theoretical and Applied Information Technology, 42(2), 281–286.

[16] Ji J, Chen G, Sun L (2011) A novel Hough transform method for line detection by anhancing accumulator array, Pattern Recognition Letters, 32(11), 1503-1510.

[17] Hirzer M (2008) Marker detection for augmented reality applications, Institut for Computer Graphics and Vision, Graz University, Technical Report, ICG-TR-08/05.

[18] Fuhrt B (2011) Handbook of Augmented Reality, Springer Science+Business Media, New York, New York.

[19] Lee T, Hollerer T (2009) Multithreaded Hybrid Feature Tracking for Markerless Augmented Reality, IEEE Transactions on Visualization and Computer Graphics, 15(3), 355–368.

[20] Gedik O S, Alatan A A (2013) 3-D Rigid Body Tracking Using Vision and Depth Sensors, IEEE Transactions on Cybernetics, 43(5), 1395–1405.

[21] Cheok A D, Qiu Y, Xu K, Kumar G K (2007) Combined Wireless Hardware and Real-Time Computer Vision Interface for Tangible Mixed Reality, IEEE Transactions on Industrial Electronics, 54(4), 2174–2189.

[22] Chen Z, Li X (2010) Markless Tracking based on Natural Feature for Augmented Reality, IEEE International Conference on Educational and Information Technology (ICEIT 2010), 2, 126-129.

[23] Fiala M (2010) Designing highly reliable fiducial markers, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(7), 1317–1324.

[24] Demuynck O, Menendez J M (2013) Magic Cards: A New Augmented Reality Approach, IEEE Computer Graphics and Applications, 33(1), 12–19.

DOI: 10.15439/2015F341
ACSIS, Vol. 5

# Word detection in recorded speech using textual queries

*Łukasz Laszko*
Cybernetics Faculty,
Military University of
Technology,
ul. Gen. S. Kaliskiego 2,
00-908 Warsaw, Poland
Email:
lukasz.laszko@wat.edu.pl

***Abstract*—The paper presents unsupervised method for word detection in recorded spoken language signal. The method is based on examining signal similarity of two analyzed media description: registered voice and a word (textual query) synthesized by using Text-to-Speech tools. The descriptions of media were given by a sequence of Mel-Frequency Cepstral Coefficients or Human-Factor Cepstral Coefficients. Dynamic Time Warping algorithm has been applied to provide time alignment of the given media description. The detection involved classification method based on cost function, calculated upon signal similarity and alignment path. Potential false matches were eliminated in the algorithm by comparing costs of the path subsequences to a threshold value. The results of the work could provide incentives to build affordable commercial or non-commercial solutions for specific and multilingual applications.**

***Index Terms*—Speech processing, speech analysis, pattern matching, keyword search, audio information retrieval**

## I. Introduction

CURRENTLY one can observe increasing use of methods and techniques of digital processing of sound information for simple daily tasks. Many of these methods and techniques have been implemented in various types of mobile devices, having as a matter of principle, relatively small memory resources and being not efficient enough to fulfill their requirements to full extent. Two common tasks in this field are connected with speech recognition and speech synthesis. Due to the limited resources two models, especially in recognition, are being observed: low quality local processing for specific usage, high quality remote (Internet service-oriented) processing[1] for wider usage. Regarding the described trend, in this paper the combination of both approaches were being adopted to word detection in recorded speech task.

Word detection relates to problem of searching for a given word in a speech medium (container or stream). In this paper only the solid container, such as: WAV, MP3 is concerned. The detection is usually given by the two [1] coupled values: position in medium (i.e. time code) and quality ratio. This problem is also recalled in contemporary literature as "keyword spotting[2]" (KWS) [2], [3] or "spoken term detection" [4] and usually refers to continues, unconstrained vocabulary speech.

Classical solutions to this problem address two-step-supervised approach where models such as hidden Markov model (HMM) or support-vector machine (SVM) are trained like in a typical automatic speech recognition (ASR) system, using Large Vocabulary Continuous Speech Recognition (LVCSR) methods [5] at the first step [6]. At this step the speech signal is divided into segments of equal-size, from which speech features are extracted. At the second step, appropriate algorithm is employed to determine the type of signal present in each segment.

Different approaches like this described in [1] base on the fact that for some applications it is not possible to have model trained, either due to lack of relevant training data or due to time-specific limitations. Moreover as maintains [7] by exploiting the structure of repeating patterns within the speech signal, unsupervised recognition task is made possible directly from an un-transcribed audio stream.

Under the concept of the unsupervised matching process lay suitable speech features and a classification strategy. Speech signal cepstrum-based features like Mel-Frequency Cepstral Coefficients (MFCC) are those used extensively in nowadays ASR [8]. Interesting study on MFCC and its two siblings can be found in [9]. However that work hasn't considered Human-Factor Cepstral Coefficients (HFCC), which had been introduced in [10] one year later. In [1] HFCCs are described as the closest to human perception system and therefore seen as more robust in this task.

Concerning classification strategy mainly two are considered: HMM with Viterbi algorithm [2] and Dynamic Time Warping (DTW) [1]. Advantages and disadvantages of

This work was supported by Cybernetics Faculty (Military University of Technology, Warsaw, Poland) under the grant no. RMN/953/2014

[1] See: www.nuancehealthcaredeveloper.com, as an example of speech recognition/synthesis service and www.shazam.com, as an example of music excerpts recognition service. For legal notice see footnote 3.

[2] Spotting task is strictly connected with pattern matching but usually without explicit requirement for further (a posteriori) verification. The word "keyword" usually occurs within a clause (not alone) or/and describes its context.

978-83-60810-66-8/$25.00©2015, IEEE

these approaches are discussed in [11]. Further in this paper the approach based on Dynamic Time Warping (DTW) for pattern matching has been chosen, as it does not require any modeling or training as compared to the HMM, but still enables one to mitigate temporal differences.

## II. PROBLEM STATEMENT

### A. Scenario and speech features

In the paper the following scenario is considered. To an operator (a user) a container with recorded speech is provided. The operator has to search the speech content for the existence of specific words. The words are either not known in advance or are changing frequently. Moreover the recorded voice origin and its language is out of operator's knowledge. The sound quality is low and its characteristics (especially the environment) are changing during analyzed period like in live telephone conversation. Still the detecting process is time-sensitive as the existence of specific words will result in more detailed examination and perhaps in appropriate human activity (like calling police or fire brigade).

The presented scenario provides considerably limited usage of classical LVCSR methods because of too little knowledge of speech signal. The approach speculated here is directed to unsupervised methods, resulting in coarse detection, connected either with user interaction (i.e. hearing) or involving precise detecting methods.

Proposed approach applied to the scenario has been taken from [1] but the innovative reference queries strategy has been proposed in this paper. The approach assumes at this point the choice of appropriate speech signal features. In the research two types of feature vectors have been used:

- Mel-Frequency Cepstral Coefficients (MFCC),
- Human-Factor Cepstral Coefficients (HFCC).

MFCCs has been computed according to the following algorithm:

1) given signal $S$ has been windowed by Hamming window resulting in $N$ segments, $s_1 ... s_N$ ;

2) each segment has been processed by short-time Fourier transform (STFT) with length of 51 ms and step size of 10 ms;

3) then the triangular filter bank has been developed with 40 equally spaced mel-scale center frequencies $f_i$, $i = 1,...,40$ and with uniform bands controlled by the neighbor center frequencies $f_{i \pm 1}$ (see Fig. 1);

4) in this step the actual filtering (or spectral smoothing) has been done, by multiplication of each STFT segment (representing magnitude spectrum) with magnitude spectrum of bands for MFCC;

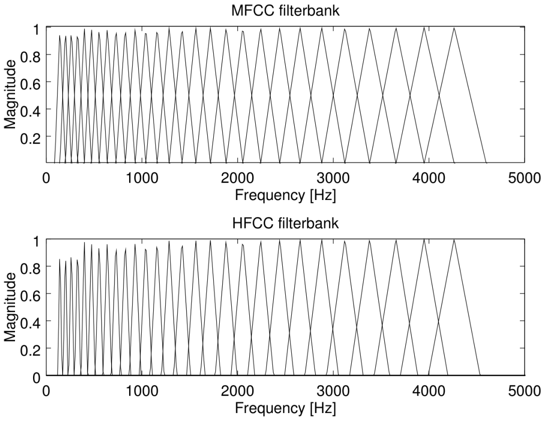


Fig. 1. Exemplary filter banks for MFCC and HFCC. HFCC filters corresponding to MFCC filters have narrower bandwidth, determined by (1).

5) the result has been then decorrelated using Discrete Cosinus Transform (DCT). Finally only 15 the most decorrelated vectors (MFCC coefficients) have been kept.

The concept of Human-Factor Cepstral Coefficients introduced in [10] employs congenial algorithm to the abovementioned. The essence of HFCC lays in filter design stage. In MFCC filter bands are determined by the spacing of the center frequencies which are equally distributed in mel frequency scale. In HFCC design filter center frequencies are still equally spaced in mel frequency scale, but unlike in MFCC filter bandwidth is treated as a parameter. This parameter determines filter bands' endpoints (cut-off frequencies) using the following measure, called Equivalent Rectangular Bandwidth (ERB) [10]:

$$ERB(f) = 6.23 f^2 + 93.39 f + 28.52 \text{ Hz} \qquad (1)$$

where $f$ states for filter center frequency, expressed in kHz.

As a result of using the approach, HFCCs are perceived as a better approximation of human auditory model [1], [10] than MFCCs and by incorporating appropriate scaling to ERB, by some factor, HFCC-based speech recognition can be under some circumstances treated as robust speech features to chosen applications [12].

### B. Textual query

For the recalled scenario two reasonable approaches are emerging:

1) Profile (or design and implement new) ASR system: write down a set of words likely to be searched for, gather relevant training set, then prepare, train and verify model of each word, according to chosen pattern learning method. Set the system to recognize only selected words from the trained set. In case of positive recognition register the time of related speech segment.

2) Exploit Text-to-Speech (TTS) system to generate synthetic voice from the text (query). Transform the produced speech query to chosen speech feature's space (the query becomes a pattern). Read a chunk of speech signal from the given source, transform it to the same speech feature space and apply appropriate classification strategy.

In case of detection (pattern matched) of the given word register the time of related speech segment.

These approaches shall be amplified (or duplicated) to reflect language variations assumed in the scenario. Implementation of the selected approach could then take advantage of parallel processing techniques and search for the same word, translated to several languages, in the given speech signal.

The second approach makes impression of being much more innovative and promising than the first one. While using TTS there is rather no limitation in producing new query, in ASR-based approach this task will be probably the most demanding, and resource consuming. Supporting argument is the existence of publicly available, free of charge online translators with speech synthesis features and accessible APIs, e.g.[3]: Google Translate, Bing Translator, Yandex Translate, etc. Based on this argument several queries can be created in less than a second.

### C. Similarity and alignment path

Dynamic Time Warping is a known and still used with success speech classifier [3]. It is popular for readability of implementation and analysis as well as for relatively high recognition accuracy similar to HMM.

In the overall look on DTW used in speech recognition it is to compare two feature vectors of different length (analyzed voice and the reference pattern) and to find an optimal alignment path $P$ of both by stretching them with respect to time. $P$ is usually calculated upon the local distance matrix (similarity matrix) from lower left corner to upper right corner of the matrix. Optimal means here the lowest cost path $P$ for passing from one point of matrix to another within given constraints. Application of DTW to exemplary speech features vectors is presented in Fig. 2. The similarity refers to speech signal of the same phrase, spoken by the same speaker. Double reduced utterance tempo is observed in analyzed voice part.

Building similarity matrix $D_{A,R}$ where $A$ stands for analyzed voice feature vector and $R$ stands for referenced pattern feature vector, is the first step considered in speech classification. Feature vector consists for either MFCC or HFCC coefficients computed for segments $s_1 \ldots s_N$. Individual element $d(a,r)$ of similarity matrix, where $a,r$ stands for specific element of vector $A$ and vector $R$ respectively, is given by inner product [6]:

$$d(a,r)=\frac{\langle A_a, R_r\rangle}{\|A_a\|\|R_r\|} \tag{2}$$

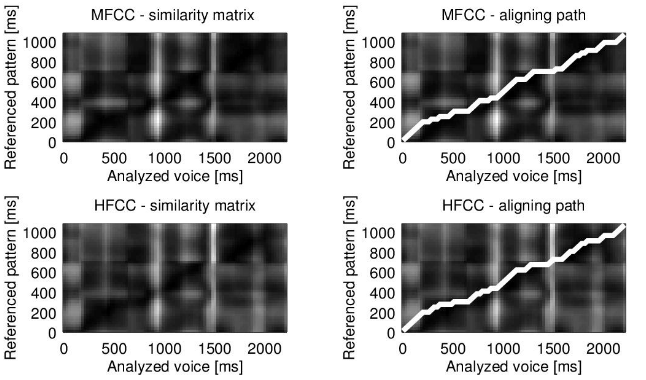


Fig. 2. Exemplary similarity matrices for MFCC and HFCC as well as corresponding aligning paths, received after applying DTW. The difference in image contrast between MFCC and HFCC indicates a higher MFCC features fuzziness.

In the following step DTW algorithm is exploited to perform cost path computation. The algorithm is two-staged[4]. At the first stage the calculation of an accumulator $C_{A,R}$ is performed (where $C$ is of size $D$). The accumulator is a structure that contains at each of its point $c(a,r)$ the value of accumulated lowest transition cost to this point from its neighbors, including the cost of lowest transition to the neighbors from theirs consequent neighbors until the starting point $c(1,1)$, retaining directional constraints, according to the recursion:

$$c(a+1,r+1)=d(a+1,r+1)+\min\begin{cases} c(a-1,r) \\ c(a,r) \\ c(a,r-1) \end{cases} \tag{3}$$

where: $a,r \geq 1$ and $c(1,1)=d(1,1)$.

At the second stage the optimal aligning path $P$ is created. Its creation is based on accumulator traceback, starting from its last point $c(N_A,N_R)$ and ending in point $c(1,1)$ recursively by searching across all allowable predecessors to each point. Because each point holds the value of the lowest transition cost to itself, the actual calculation of the path is based on choosing the next point upon the minimal value.

Presented algorithm is a type of conventional Dynamic Time Warping and it is regarded as slow and memory consuming [11], especially for aligning large sequences, therefore a practical usage could engage its specific modifications, like a multiscale approach [14]. Nevertheless in this for DTW has been used in the research described in the paper.

### D. Classification and verification

DTW presented in preceding paragraph refers to global alignment of one feature vector to another (in time domain), such as these presented in Fig. 2 (on the right). In general this is not enough to classify the part of analyzed voice as detected word, with regard to referenced pattern. Moreover

[3] Author of the paper would like to strongly accent having nothing in common with the companies whose products were listed. The author is far from rating, comparing and criticizing these products. Names of the products have been presented in this paper only in relation to the contemporary, publicly available technology, not for marketing purposes.

[4] The DTW algorithm used in the reported research, is based on the code originated with the project described in [13].

as commonly $N_A >> N_R$ holds, additional matching procedure shall be applied. The matching procedure is presented in figure Fig. 3.

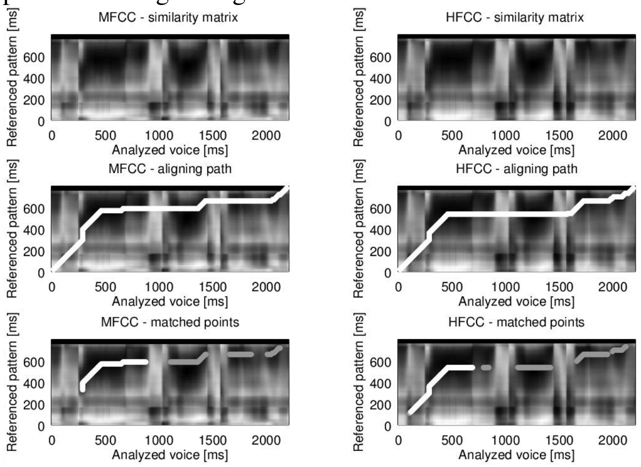


Fig. 3. Pattern matching procedure. Upper images present computed similarity between the pattern (the word “school” - synthesized woman’s voice) and analyzed voice (recorded man’s voice “school is closed today”). Images in the middle present global alignment path. Bottom images present resultant match: white strip is for the best match, gray strips are for remaining matches.

After computation of aligning path $P$, the points $p_1 ... p_{N_P}$ lying on the path are being assigned weight values $v$ based on referring points of matrix $D_{A,R}$ and a path threshold $T_P$, satisfying inequality (4).

$$T_P \leq v := 1 - d \leq 1 \tag{4}$$

$T_P$ controls the number of points suspected to indicate detected words. In Fig. 3 they are presented in the bottom images (as gray strips). As there could be several alleged word occurrences detected, indicated by subsequences: $p_{k_1}^{(l)} ... p_{k_{N_P}}^{(l)}$, $l = 1,2...$, the verification step is executed. This step consists of computing Longest Common Subsequence (LCS) with maximization criterion of cumulative weights of subsequences, i.e. the longest subsequence with the highest weights sum wins. It is worth noting that the minimal cumulative cost for a subsequence should be restricted by the second threshold value (called here sequence threshold), controlling the number of possible word reoccurrence.

As a result of matching procedure, assuming only one occurrence of the searched word, the best match is projected to the analyzed voice time domain (see white strips in the bottom images of Fig. 3) and, according to the scenario assumed in research, presented to the operator as a speech signal.

## III. UNSUPERVISED KEY DETECTION

### A. Algorithm

On the basis of literature search, especially [1], [2], [3], [11], [12] and performed experiments, the following algorithm has been proposed to unsupervised key detection.

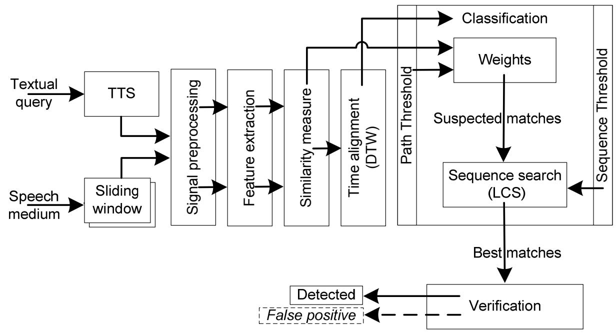


Fig. 4. Unsupervised key detection algorithm.

Description of respective processing blocks, presented in Fig. 4 is the content of the previous paragraphs, except of the sliding window and signal preprocessing. Sliding window block represents the abstract of feeding the model (or the program etc.) with respectively short signal, which the model is able to process. The size of the window is to be determined before the start of examination. It is assumed to be the function of size of the query.

Signal preprocessing is understood as applying standard digital signal processing techniques, at least but not limited to: silence cancellation, resampling and pre-emphasis filtering.

### B. Experiments

A series of preliminary experiments have been conducted on the basis of methods and techniques of speech signal processing addressed in the paper, with regard to the presented algorithm. The target was to detect a word in a given speech medium by examining signal similarity of two analyzed media description: registered voice and referenced pattern, where the descriptions were given by a sequence of MFCC or HFCC coefficients.

Dedicated research material has been prepared, which consists of five short (from 1 – 5 seconds) sentences in English language: spoken by one man (natural speech) and synthesized by six (free of charge) TTS systems with fourteen men voices, and nine women voices. This material has been stored on a hard drive in the WAV containers.

The queries have been produced online by three TTS systems, different from these used in preparing research material. One TTS was the part of local operating system while two others were available through the World Wide Web via HTTP protocol. Nevertheless query generation was taking less than one second.

The research material as well as the queries were of the following (different[5]) properties: PCM (lossless) codec, one channel, 16 bits per sample, sampling frequency: 8000 – 22050 Hz, bit rate: 256 – 352 kbps. During examination all used sounds were resampled to 8000 Hz. The model (Fig. 4) was configured according to the guidelines from paragraph II.A. For HFCC ERB scaling factor of value 3 was used.

[5] During material acquisition (except of the natural voice recording) there was not possible to influence sound properties.

Experiments have been conducted according to the following strategy: selected word to find (textual query) has been sent to the TTS system to obtain speech signal, the signal then has been read by program and compared with the entire research material according to the algorithm (Fig. 4). This has covered both situations: the existence and inexistence of the word in the examination set.

### C. Results

Overall results have been presented in Table 1. "No detection" phrase used in the results means the percent of false negatives (lack of detection, when the word actually existed in the analyzed signal). Higher detection rate for HFCC, as well as lack of false negatives for these features is noticeable. Nevertheless MFCC present lower false detections.

TABLE 1. OVERALL RESULTS BY SPEECH FEATURES

| | Detected words | No detection | False positive |
|---|---|---|---|
| MFCC | 82,43% | 4,05% | 13,51% |
| HFCC | 85,14% | 0,00% | 14,86% |

The results showed that the unsupervised detection of word in a given set is possible with relatively high detection rate. It is worth noting the lack of "no detection" results when using HFCC.

In Table 2 the discrimination on the basis of speech source has been presented. The results give the overview that either male or female synthesized voice can be with success use to detect words in speech.

TABLE 2. MEAN RESULTS BY SPEECH SOURCE

| | Detected words | No detection | False positive |
|---|---|---|---|
| Real speech | 95% | 0% | 5% |
| TTS (male) | 80% | 15% | 15% |
| TTS (female) | 70% | 12,5% | 17,5% |

HFCC based detection gives better overall results in comparison with the MFCC method, but the results of the research don't allow to make general statement for this.

## IV. CONCLUSION

Results of the work presented in this paper are satisfactory, but still far from industrial standard. It is difficult to make direct comparisons with other related works as this work has been conducted partly on synthesized material. The following work shall be considered: (1) applying multiscale approach (like presented in [13]) for bypassing fully different signal parts, and analyzing these initially consonant parts on higher resolution, (2) parallel processing: for searching the multilingual queries, as well as increasing detection reliability based on information fusion.

Moreover additional study on less speaker-dependent features should be done, to limit the number of false positives.

The main problem in the approach (especially while applying DTW and LCS) that cannot be fully circumvent is the determination of the threshold values. The mitigation of the problem could employ computation of the values based on query or signal properties.

In general view of this work, its results could provide incentives to build affordable multilingual commercial or non-commercial solutions for specific applications.

## ACKNOWLEDGMENT

Author would like to thank the team (of the year 2010) of KOM department, Fraunhofer FKIE, Wachtberg, Germany, especially Dirk, Frank and Paul (NUWC), who gave me the inspiration of the topic, as well as the leader of the department Dr. Markus Antweiler, who enabled my scientific internship at FKIE.

## REFERENCES

[1] D. von Zeddelmann, F. Kurth, and M. Müller, "Perceptual audio features for unsupervised key-phrase detection," Proc. ICASSP2010, 2010, pp. 257-260, DOI:10.1109/ICASSP.2010.5495974.

[2] S. Tabibian, A. Akbar, B. Nasersharif, "A fast search technique for discriminative keyword spotting," Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on, pp.140-144, 2-3 May 2012, DOI:10.1109/AISP.2012.6313733.

[3] M. Sigmund, "Search for Keywords and Vocal Elements in Audio Recordings", Elektronika ir elektrotechnika, ISSN 1392-1215, vol. 19, no. 9, pp. 71-74, 2013

[4] V. Mitra, J. van Hout, et. al., "Feature Fusion for High-Accuracy Keyword Spotting", Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 7143-7147 2014

[5] A. Manos and V. Zue, "A segment-based word spotter using phonetic Filler models," in proceeding of ICASSP, pp. 899–902, 1997.

[6] W. Kwiatkowski, "Methods of automatic pattern recognition" (in polish: „Metody automatycznego rozpoznawania wzorców"), Instytut Automatyki i Robotyki, WAT, ISBN 83-912747-7-2, Wydanie I, Warszawa 2001, pp. 185-191; 118-121 .

[7] A. S. Park and James R. Glass, (Cited in [1]) "Unsupervised pattern discovery in speech," IEEE Trans. on Audio, Speech and Language Processing, vol. 16, no. 1, pp. 186–197, 2008.

[8] D. Eringis, G. Tamulevičius. "Modified Filterbank Analysis Features for Speech Recognition", Baltic J. Modern Computing, Vol. 3 (2015), No. 1, 29-42.

[9] B. J. Shannon, K. K. Paliwal, "A Comparative Study of Filter Bank Spacing for Speech Recognition", Microelectronic Engineering Research Conference 2003.

[10] M. D. Skowronski and J. G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," The Journal of the Acoustical Society of America (JASA), vol. 116, no. 3, pp. 1774–1780, 2004.

[11] M. S. Barakat, C. H. Ritz , D. A. Stirling, "Keyword spotting based on the analysis of template matching distances", Signal Processing and Communication Systems (ICSPCS), 2011 5th International Conference on, pp.1-6, 12-14 Dec. 2011, DOI:10.1109/ ICSPCS.2011.6140822.

[12] R. Wielgat, T. P. Zieliński, T. Potempa, A. Lisowska-Lis, D. Król, "HFCC based recognition of bird species", In: Signal Processing: Algorithms, Architectures, Arrangements, and Applications, ISBN-13 978-83-913251-8-6, pp. 129–134, Poznań 2007.

[13] R. Turetsky and D. Ellis, "Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses", 4th International Symposium on Music Information Retrieval ISMIR-03, pp. 135-141, Baltimore, October 2003.

[14] A. Zinke and D. Mayer, "Iterative Multi Scale Dynamic Time Warping", In: Computer graphics technical reports, CG-2006/1, ISSN 1610-8892, Universität Bonn, 2006.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 855–860

DOI: 10.15439/2015F53
ACSIS, Vol. 5

# Using the Generalised Viterbi Algorithm to Achieve a Highly Effective Stegosystem for Images

Valery Korzhik
(Member of IEEE)
State University
of Telecommunications
Saint-Petersburg, Russia
Email: val-korzhik@yandex.ru

Guillermo Morales-Luna
Computer Science
CINVESTAV-IPN
Mexico City, Mexico
Email: gmorales@cs.cinvestav.mx

Ivan Fedyanin
State University of Telecommunications
Saint-Petersburg, Russia
Email: ivan.a.fedyanin@gmail.com

***Abstract*—The HUGO project, published at 2010, can be considered as one of the most promising direction in the design of highly undetectable steganography. The main idea of that approach is to minimise the embedding impact from the steganalysis point of view. This goal is achieved by using trellis codes in the embedding procedure, the Viterbi algorithm (VA) and the SPAM features. But the optimality of VA was kept still unclear because a generic purpose of VA is to correct errors with trellis codes instead of embedding secret information. The first goal of the current paper is to prove the optimality of VA application in its generalised form proposed about 30 years ago by one of the authors of this paper. The second goal is to optimise the parameters of the trellis code check matrix for better undetectability of stegosystems.**

***Index Terms*—stegosystem, images, HUGO project, trellis codes, generalised Viterbi algorithm.**

## I. Introduction

STEGANOGRAPHY (SG) is the information hiding technique that embeds the hidden information into an innocent *cover object* (CO) under conditions that CO is not corrupted significantly and that the presence of the additional information in CO may not be detected. This goal entails an obvious requirement: the embedding impact has to be minimised from the steganalysis point of view. Moreover, it does not mean that the number of changes into CO just after embedding should be minimised because the changes “weights” may differ, hence a minimisation of changes into CO does not necessary results in the minimisation of SG detectability.

The assumed most effective stegoanalytic method is the so called *blind steganalysis* based on the transition probabilities of the *Subtractive Pixel Adjacency Matrix* (SPAM) features model between neighboring pixel of the image and CO along 8 different directions [1], [2].

Let $X \in \mathbb{R}^{n_1\times n_2}$ be a grey scale cover image and let $Y \in \mathbb{R}^{n_1\times n_2}$ be the resulting image after embedding using some stego-algorithm. Let $D(X,Y)$ be the distortion of the stegoimage and the cover image, in the sense of ability of the SPAM stegoalgorithm to distinguish $X$ and $Y$. Up to an enumeration of the pixel array, any image $X$ can be regarded as an array $\mathbf{x} = (x_i)_{i=1}^n$, with $n = n_1 n_2$.

The additive distortion function chosen for SPAM and *Highly Undetectable steGO* (HUGO) [3] is

$$(X,Y) \mapsto D(X,Y) = \sum_{i=1}^{n} \rho_i \left|x_i - y_i\right| \tag{1}$$

where, $\rho_i \in \mathbb{R}^+ \cup \{+\infty\}$ is the weight coefficient (the cost of changing the $i$-th pixel), $x_i, y_i$ are the values of the $i$-th pixel at $X$ and $Y$ respectively. All changes are restricted to $\pm 1$ increments, so that the following inequality holds after $\pm 1$-embeddings in LSB:

$$\forall i \text{ with } 1 \leq i \leq n: \ |x_i - y_i| \leq 1.$$

The additive form of (1) means that detectability of SG does not depend on the correlation between the embedded bits. That assumption holds when the changed pixels are located sufficiently far from each other, which in turn holds when the embedding rate is relatively low.

Two immediate problems arise in the design of effective SG:

1) How to choose adequately the weights $(\rho_i)_{i=1}^n$?
2) What is the best coding method for changing pixels according to their weights?

In order to find the pixel weights to be used at (1), for each $i \in \{1,\dots,n\}$, let $Y^{(i)}$ be the image $X$ with the $i$-th pixel changed. Then, $\forall j \in \{1,\dots,n\}: \ |x_j - y_j^{(i)}| \leq 1$. Let us pick $\rho_i = D(X, Y^{(i)})$.

The weight $D(X,Y)$ can be calculated as proposed in [3], as the addition of two sums:

$$\sum_{d_1,d_2=-T}^{T} w(d_1,d_2) \left| \sum_{k\in U} \left(C^k_{d_1 d_2}(X) - C^k_{d_1 d_2}(Y)\right)\right| + \\ \sum_{d_1,d_2=-T}^{T} w(d_1,d_2) \left| \sum_{k\in V} \left(C^k_{d_1 d_2}(X) - C^k_{d_1 d_2}(Y)\right)\right| \tag{2}$$

where the map $C^k_{d_1 d_2}$ is calculated, in line with the SPAM features, as the Markov transition probabilities for the eight directions at $U \cup V$, with

$$U = \{\leftarrow, \rightarrow, \uparrow, \downarrow\} \text{ and } V = \{\searrow, \nwarrow, \nearrow, \swarrow\}.$$

978-83-60810-66-8/$25.00©2015, IEEE

In particular, for the case of the horizontal direction (→), for any integers $d_1, d_2 \in [-T, T]$:

$$C^{\rightarrow}_{d_1 d_2}(X) = \Pr\left(D^{\rightarrow}_{i,j} = d_1 \ \& \ D^{\rightarrow}_{i,j+1} = d_2\right) \tag{3}$$

where for $1 \le i \le n_1$ and $1 \le j \le n_2 - 1$,

$$D^{\rightarrow}_{i,j} = X_{i,j} - X_{i,j+1}.$$

As in [3], we consider a weight function of the form:

$$w(d_1, d_2) = \left[\sqrt{d_1^2 + d_2^2} + \sigma\right]^{-\gamma} \tag{4}$$

for some optimised parameters $\sigma > 0$ and $\gamma > 0$. It is well known [1] that in order to minimise the number of changes among pixels of the CO, a *syndrome coding* may be used:

$$H\,\mathbf{y} = \mathbf{m} \tag{5}$$

where $\mathbf{y}$ is the block of stego-image bits of length $n = n_1 n_2$, $\mathbf{m}$ is a block of information bits of length $k$ that should be embedded into $\mathbf{y}$ and $H$ is a $(k \times n)$-matrix chosen for encoding. Next, among all the solutions $\mathbf{y}$, for given $\mathbf{m}$ and $H$, it is necessary to select those providing minimum number of changes between $\mathbf{y}$ and its CO block $\mathbf{x}$. At [1], a technique to solve this problem was presented, and it is especially simple for the Hamming code with a specific check matrix $H$. But in our case it is necessary to minimise not the number of changes in CO but the distortion function $D : (X, Y) \mapsto D(X, Y)$ given by (1). It seems to be similar to a transition from hard decoding to soft decoding in communications where trellis codes using Viterbi algorithm for decoding can be more favorable than block codes, using some algebraic decoding algorithms. Nevertheless there exists a significant difference in the solution of the matrix equation (5) with respect to $\mathbf{y}$ given $\mathbf{m}$ with minimising the distortion function $D$ given $\mathbf{x}$ as a CO and the solution of the error correction procedure equation $H\,\mathbf{m} = \mathbf{y}$, given some distance between $\mathbf{y}$ and $\mathbf{x}$. Fortunately, there exists the so called *generalised Viterbi algorithm* (GVA) proposed in 1984 [4], [5] that is able to describe both error correction and also steganographic embedding problem in analytic form without execution on trellis graphs.

This method of stegosystem design based on trellis codes and GVA becomes clearer and gets a strict justification. We describe GVA and its application to steganographic problem design in the Section II.

Results of matrix optimisation, which are obtained by simulation with the use of GVA, are presented in Section III. Section IV concludes the paper.

## II. Application of GVA to the SG system design problem

Let us state a problem that results in a natural solution within GVA [4], [5].

**Problem.** Find

$$\tilde{\mathbf{x}}_N = \arg\min_{\mathbf{x}_N} \Lambda_N(\mathbf{x}_N) \tag{6}$$

where

$$\begin{aligned} \Lambda_N(\mathbf{x}_N) &= \sum_{k=1}^{N} \lambda(\xi_k) \quad \text{with} \\ \xi_k &= \begin{cases} (x_1, x_2, \dots, x_\nu) & \text{if } k \le \nu \\ (x_{k-\nu}, x_{k-(\nu-1)}, \dots, x_k) & \text{if } k > \nu \end{cases} \end{aligned}$$

where $0 \le \nu \le N - 1$, each entry $x_j$ is in the set $X$, $\operatorname{card}(X) = r$, and $\lambda$ is a real function defined on the set of finite length real sequences. Here, the goal at (6) is to minimise, but it can be to maximise, as well.

The stated problem can be solved through an algorithm introduced at [5]:

1. Find $\tilde{x}_1 = \arg\min_{x_1} \Lambda_{\nu+1}(x_1, x_2, \dots, x_{\nu+1})$.
2. Find $\tilde{x}_2 = \arg\min_{x_s} \Lambda_{\nu+2}(\tilde{x}_1, x_2, \dots, x_{\nu+2})$.

$\vdots$

$s$. Find $\tilde{x}_s = \arg\min_{x_s} \Lambda_{\nu+2}(\tilde{x}_1, \dots, \tilde{x}_{s-1}, x_s, \dots, x_{\nu+s})$.

$\vdots$

$N-\nu$. Find $(\tilde{x}_{N-\nu}, \dots, \tilde{x}_N) = \underset{(x_{N-\nu}, \dots, x_N)}{\arg\min} \Lambda_{\nu+2}(\tilde{\mathbf{x}}_N)$, where $\tilde{\mathbf{x}}_N = (\tilde{x}_1, \dots, \tilde{x}_{N-\nu-1}, x_{N-\nu}, \dots, x_{\nu+s})$

It is easy to see that for all steps, except the last one, the number of calculations for every argument is at most $r^{\nu+1}$ and the last step requires at most $r^\nu$ operations.

We note that the conventional VA is a particular case of GVA with $\nu = 1$. Then we get:

$$\Lambda_N(\mathbf{x}_N) = \lambda(x_1) + \lambda(x_1, x_2) + \cdots + \lambda(x_{N-1}, x_N). \tag{7}$$

If we assume now that each $x_k$ in (7) is the state of trellis on the $k$-th step and $\lambda(x_k, x_{k+1})$ is the length of branch from the state $x_k$ to the state $x_{k+1}$, then the decoding problem for convolutional code presented by trellis diagram is equivalent to a minimisation of (6). But fortunately, GVA can be used also for a situation of embedding problem for SG given by (5) that seems to be at a single glance completely different than correction of errors by convolutional codes.

Let us consider a matrix $H$ in (5) that has the "step-wise" sliding form based on a submatrix $\tilde{H}$ of order $t \times w$ pictured at Fig. 1.

In order to simplify further the description, let us consider the particular case of $t = 2$, $w = 2$, and $\tilde{H} = [h_{ij}]_{1 \le i,j \le 2}$. Let us assume also $k > 0$, and $n = 2k$. Then equation (5) can be written as follows:

$$\begin{bmatrix} h_{11} & h_{12} & 0 & 0 & \cdots & 0 & 0 \\ h_{21} & h_{22} & h_{11} & h_{12} & \cdots & 0 & 0 \\ 0 & 0 & h_{21} & h_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & h_{11} & h_{12} \end{bmatrix} \mathbf{y} = \mathbf{m} \tag{8}$$

with

$$\begin{aligned} \mathbf{y} &= [y_1 \ \cdots \ y_{2k}]^T \\ \mathbf{m} &= [m_1 \ \cdots \ m_k]^T \end{aligned}$$

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1w} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ h_{21} & h_{22} & \cdots & h_{2w} & h_{11} & h_{12} & \cdots & h_{1w} & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & h_{21} & h_{22} & \cdots & h_{2w} & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ h_{t1} & h_{t2} & \cdots & h_{tw} & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & h_{t1} & h_{t2} & \cdots & h_{tw} & & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & h_{11} & h_{12} & \cdots & h_{1w} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & h_{21} & h_{22} & \cdots & h_{2w} & h_{11} & h_{12} & \cdots & h_{1w} \end{bmatrix}$$

Fig. 1. $H$ as a step-wise sliding matrix.

determining the equation system

$$\begin{array}{rcl} h_{11}y_1 + h_{12}y_2 & = & m_1 \\ \forall j = 2, \ldots, k-1 : & & \\ h_{21}y_{2j-3} + h_{22}y_{2j-2} + h_{11}y_{2j-1} + h_{12}y_{2j} & = & m_j \\ h_{11}y_{2k-1} + h_{12}y_{2k} & = & m_k \end{array} \tag{9}$$

It is possible to apply GVA to solve the system (9) given the column vector $M$ and $X$ providing a minimisation of the weight

$$D(X, Y) = \sum_{j=1}^{2k} \rho_j \, |x_j - y_j|.$$

This algorithm can be performed through the following steps:

1) Build the Table of variants $(\hat{y}_1, \hat{y}_2)$ for all possible tuples $(y_3, y_4)$ satisfying equation (9), for $j = 2$, and minimise the function
$$\Lambda_1 = \rho(x_1, \hat{y}_1) + \rho(x_2, \hat{y}_2),$$
where $\rho(x_i, y_i) = \rho_i |x_i - y_i|$.
2) Build the Table of variants $(\hat{y}_3, \hat{y}_4)$ for all possible tuples $(y_5, y_6)$ satisfying equation (9), for $j = 3$, and minimise the function
$$\Lambda_2 = \rho(x_1, \hat{y}_1) + \rho(x_2, \hat{y}_2) + \rho(x_3, \hat{y}_3) + \rho(x_4, \hat{y}_4),$$
where $\rho(x_i, y_i) = \rho_i |x_i - y_i|$.

Proceed similarly up to the last equation at (9).

**Example.** Let $k = 3$, $\rho_i = 1$ for $i = 1, \ldots, 6$, $\mathbf{x} = 101110$ and $\mathbf{m} = 100$. Then the matrix $H$ is

$$\begin{array}{rcl} H & = & \begin{bmatrix} h_{11} & h_{12} & 0 & 0 & 0 & 0 \\ h_{21} & h_{22} & h_{11} & h_{12} & 0 & 0 \\ 0 & 0 & h_{21} & h_{22} & h_{11} & h_{12} \end{bmatrix} \\ & = & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}. \end{array}$$

In line with (8) we have

$$H\mathbf{y} = \mathbf{m} \tag{10}$$

where

$$\begin{array}{rcl} \mathbf{y} & = & [y_1 \;\; y_2 \;\; y_3 \;\; y_4 \;\; y_5 \;\; y_6]^T \\ \mathbf{m} & = & [1 \;\; 0 \;\; 0]^T \end{array}$$

| $y_1$ | $y_2$ | $y_3$ | $y_4$ | (*) | $\Lambda_1$ |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | – |
| 1 | 1 | 0 | 0 | 0* | 1 |
| 1 | 0 | 0 | 1 | 1 | – |
| 1 | 1 | 0 | 1 | 0* | 1 |
| 1 | 0 | 1 | 0 | 0* | 0 |
| 1 | 1 | 1 | 0 | 1 | – |
| 1 | 0 | 1 | 1 | 0* | 0 |
| 1 | 1 | 1 | 1 | 1 | – |

(*) Left side of the second equation in (9)
Asterisk means that left side of the second equation of (9) coincides with $m_2 = 0$.

TABLE I
VALUES OF $\Lambda_1$.

| $y_3$ | $y_4$ | $y_5$ | $y_6$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1 | | |
| 0 | 0 | 0 | 1 |
| 1 | 1 | | |
| 0 | 1 | 1 | 0 |
| 1 | 0 | | |
| 0 | 1 | 1 | 1 |
| 1 | 0 | | |

TABLE II
ADMISSIBLE TUPLES $(y_1, y_2, y_3, y_4)$.

1) In order to satisfy the first equation in (9) we get two possibilities $(y_1, y_2) = (1, 0)$, and $(y_1, y_2) = (1, 1)$.
2) For all possible pairs $(y_3, y_4)$ and $(y_1, y_2)$ which are obtained in the Step 1) we get the possibilities to satisfy the second equation in (9) shown at Table I.

In Table I there are presented also the calculation results of the values $\Lambda_1$ for "survived" strings (with an asterisk). Next in Table II we present all possible pairs $(y_3, y_4)$ satisfying to the corresponding equation in (9) for every pair $(y_5, y_6)$.

By combining the Tables I and II, we get Table III presenting all possible$(y_1, y_2, y_3, y_4)$ tuples for every pair $(y_5, y_6)$ and corresponding to them, the values $\Lambda_3$.

Now it is possible to minimise $\Lambda_3$ by selecting a pair $(y_5, y_6)$. This gives two optimal tuples $\mathbf{y} = 101100$ and $\mathbf{y} = 101010$. It is easy to see that each of these tuples requires one change of the tuple $\mathbf{x}$ and satisfies the equation (10).

This approach can be extended to any "step-wise" matrix generated by shifting the $(t \times w)$-submatrix $\tilde{H}$. Then the

| $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $\Lambda_3$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 4 |
| 1 | 0 | 1 | 1 | | | 1 |
| 1 | 1 | 0 | 0 | 0 | 1 | 5 |
| 1 | 0 | 1 | 1 | | | 2 |
| 1 | 1 | 0 | 1 | 1 | 0 | 2 |
| 1 | 0 | 1 | 0 | | | 1 |
| 1 | 1 | 0 | 1 | 1 | 1 | 3 |
| 1 | 0 | 1 | 0 | | | 2 |

TABLE III
VALUES OF $\Lambda_3$.

complexity for GVA will be of the order $O(kwt\,2^{wt})$ with respect to operations.

## III. OPTIMISATION OF $\tilde{H}$ BY SIMULATION WITH GVA

Let us consider initially the case with $\rho_i = 1$, for $i = 1, \dots, n$. This means that we try to minimise the number of changes into the image pixels after embedding against the sizes $t$ and $w$ of a randomly chosen submatrix $\tilde{H}$. The results of simulation are shown in Fig. 2 where relative changes

$$\nu = \frac{\text{card}\left(\{i|\ x_i \neq y_j\}\right)}{n}$$

are presented on the vertical axis.

From Fig. 2 it can be seen that it is sufficient to bound the sizes of the submatrix $\tilde{H}$ as $t \leq 15$, $w \leq 10$, at least for the case $\rho_i = 1$, for $i = 1, \dots, n$.

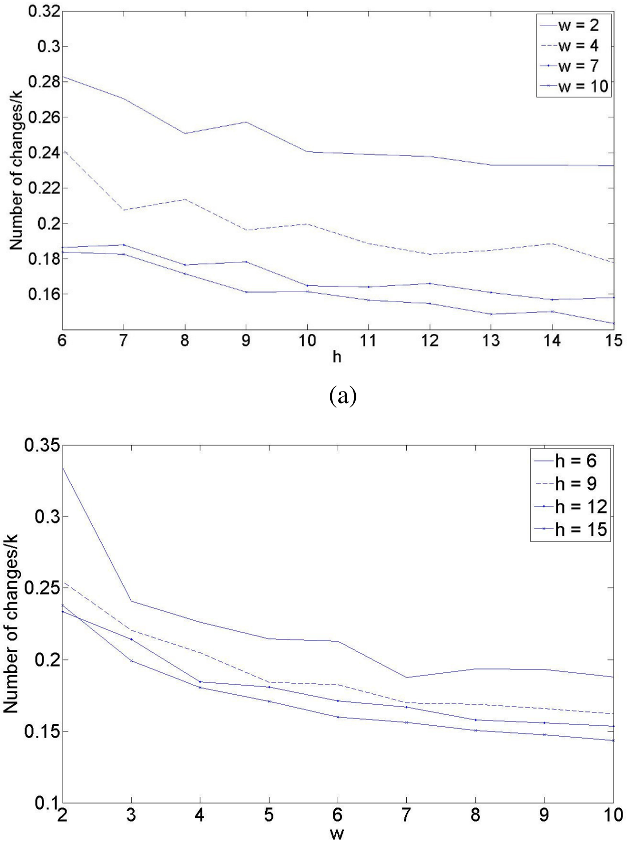
But it is more interesting to investigate the undetectability of SG based on syndrome embedding under the condition that the SG detection is performed by blind SVM-based steganalysis with the use of SPAM features (see (2)-(4)). The experiment has been arranged as follows: The image base was taken similar to those considered in [6]. Since the embedding procedure is very time-consuming the images were reduced to lesser sizes. The weight coefficients $\rho_{ij}$ in (1) were calculated by (2), (3) with the weight function (4) after optimization of parameters $\sigma$ and $\gamma$. At the training SVM stage both 500 images with and without embedding were used. During the testing SVM stage, there were executed 500 images with and without embedding. This HUGO-based algorithm was compared with conventional ±1 embedding algorithm taken with the same embedding rate $R = \frac{k}{n}$, where $k$ is the number of embedded bits and $n$ is the number of image pixels. As a criterion of stegosystem undetectability, it was used (in line with a recommendations [1]) the averaged error probability

$$P_e = \frac{1}{2}\left(P_m + P_{fa}\right)$$

where $P_m$ is the probability of SG missing and $P_{fa}$ is the probability of SG false alarm. The value $P_e$ has been minimised at the cost of SVM threshold optimization.

The results of simulations are shown in Tables IV and V.

We can see from these Tables that the use of the HUGO trellis code-based SG offers some advantages in undetectabilities against a conventional ±1 in LSB embedding SG.



(a)

(b)

Fig. 2. Relative changes of image pixels after embedding based on trellis codes with $n_1 n_2 = 10^5$. (a) against the matrix parameter $t$ given $w$, and (b) against the matrix parameter $w$ given $t$.

| Image size | Embedding rate $R = \frac{k}{n}$ | $P_e$ |
|---|---|---|
| $16 \times 16$ | 0.4 | 0.27 |
| $64 \times 64$ | 0.4 | 0.176 |
| $128 \times 128$ | 0.2 | 0.1780 |
| $128 \times 128$ | 0.4 | 0.123 |
| $256 \times 256$ | 0.4 | 0.099 |

TABLE IV
THE PROBABILITIES OF INCORRECT ±1 SG SYSTEM DETECTION ($P_e$) FOR DIFFERENT IMAGE SIZES AND EMBEDDING RATES $R$.

It is worth to note that the undetectability of SG, even in the case of a trellis code-based embedding (used with the HUGO project), depends significantly on the image texturing. We have found that the greater is the degree of texture, the most frequent undetectability of the corresponding image. Qualitatively, image texturing means that the image has a presence of precise contours, while not texture images have sliding luminance changing and noisy background. Numerically image texture can be estimated by the parameter [1]

$$t_n = \frac{1}{n_1 n_2} \sum_{ij} \left( \max_k B_{ij}^k - \min_k B_{ij}^k \right) \qquad (11)$$

| Image size | $R = \frac{k}{n}$ | SPAM parameters | | | $P_e$ |
|---|---|---|---|---|---|
| | | $T$ | $\sigma$ | $\gamma$ | |
| $16 \times 16$ | 0.4 | 10 | 10 | 4 | 0.337 |
| $64 \times 64$ | 0.4 | 10 | 10 | 4 | 0.24 |
| $128 \times 128$ | 0.4 | 10 | 10 | 4 | 0.162 |
| $128 \times 128$ | 0.2 | 10 | 10 | 4 | 0.269 |

TABLE V
THE PROBABILITIES OF INCORRECT HUGO SG SYSTEM SPAM-BASED DETECTION ($P_e$) FOR DIFFERENT IMAGE SIZES AND EMBEDDING RATES $R$

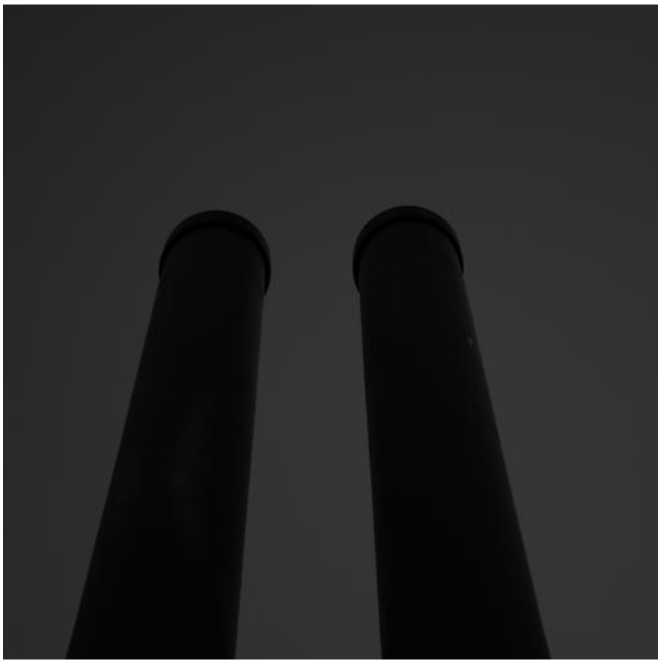


Fig. 3. Example of a low texture image.

where $B_{ij}^k$ is a $(2 \times 2)$-pixel block with $(i, j)$-coordinates and $k$ is the $k$-th pixel of this block.

On Fig's. 3 and 4 there are shown examples of images with low and high texture coefficient $t_n$, respectively.

As it can be seen by (11), in order to calculate a measure of image texture $t_n$ it is necessary to divide the image on disjoint $(2 \times 2)$-blocks and next to calculate for every block the difference between maximum and minimum pixel luminance of this block.

In Table VI there are presented the averaged error detecting probabilities for HUGO-based SG after embedding of messages into the images with different texturing. For both SVM training and testing phases 500 images from different image sets were used.

We can see from this Table that, in fact, the level of image texturing affects very significantly on SG system detectability. It seems to be recommended to select for SG embedding such images, which have large texture level. But on the other side it can be looking suspiciously with point of steganographic usage view. Maybe it is better to embed the amount of secret bits depending on the level of image texture.

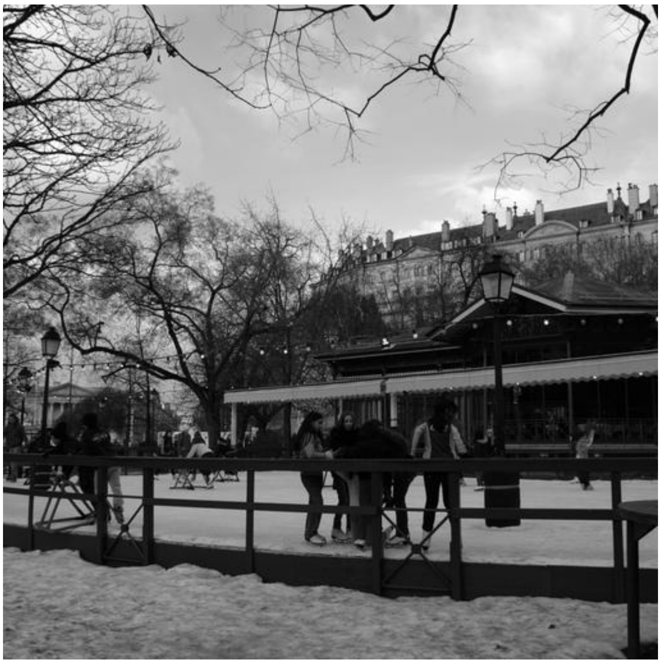


Fig. 4. Example of a high texture image.

| Image size | $t_n$ | $R = \frac{k}{n}$ | SPAM parameters | | | $P_e$ |
|---|---|---|---|---|---|---|
| | | | $T$ | $\sigma$ | $\gamma$ | |
| $64 \times 64$ | $< 0.194$ | 0.4 | 3 | 10 | 4 | 0.012 |
| $64 \times 64$ | $> 5.28$ | 0.4 | 3 | 10 | 4 | 0.391 |

TABLE VI
THE ERROR DETECTING PROBABILITIES FOR HUGO-BASED SG AFTER EMBEDDING OF MESSAGES INTO THE IMAGES WITH DIFFERENT TEXTURING.

## IV. CONCLUSION

A new generation of stegosystems with the use of trellis code-based embedding is very promising because this approach minimises an embedding impact with the point of view blind SVM-SPAM based steganalysis. This is the so called HUGO project developed recently [3].

But our main contribution into this direction is to make more clear and to prove rigorously that the use of generalised Viterbi Algorithm is in fact the optimal embedding procedure jointly with trellis codes.

With the use of this method we have shown experimentally that HUGO-based SG has significant advantage against simple stegosystems (like LSB or ±1 algorithm). We have found also that the level of image texturing is very important in a choice of images intended for embedding of hidden messages with high undetectability.

We agree with the importance given to “Open Problem 1” in [7], namely the design of effective coding schemes for non-additive distortion function.

# References

[1] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*, 1st ed. New York, NY, USA: Cambridge University Press, 2009.

[2] T. Pevny, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *Information Forensics and Security, IEEE Transactions on*, vol. 5, no. 2, pp. 215–224, June 2010.

[3] T. Pevný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Proceedings of the 12th International Conference on Information Hiding*, ser. IH'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 161–177.

[4] V. Korzhik, "A generalization of Viterbi algorithm on the case of channel model described by additive Markov channel," in *Proc. of IV Intern'l Symp on Information Theory, Part II*, 1984, pp. 109–111.

[5] V. Korzhik and Y. Lopato, "Optimal decoding of convolutional codes in channels with additive markov noises," in *Proc. of IV Intern'l Symp on Information Theory, Part II*, 1984, pp. 35–40.

[6] P. Bas, T. Filler, and T. Pevný, ""Break our steganographic system": the ins and outs of organizing BOSS," in *Proceedings of the 13th international conference on Information hiding*, ser. IH'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 59–70. [Online]. Available: http://dl.acm.org/citation.cfm?id=2042445.2042452

[7] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevný, "Moving steganography and steganalysis from the laboratory into the real world," in *Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&#38;MMSec '13. New York, NY, USA: ACM, 2013, pp. 45–58. [Online]. Available: http://doi.acm.org/10.1145/2482513.2482965

DOI: 10.15439/2015F258
ACSIS, Vol. 5

# Automatic Classification of Fruit Defects based on Co-Occurrence Matrix and Neural Networks

Giacomo Capizzi*, Grazia Lo Sciuto†, Christian Napoli‡, Emiliano Tramontana‡, Marcin Woźniak§
*Department of Electrical and Informatics Engineering, University of Catania, Viale A. Doria 6, 95125 Catania, Italy
Email: capizzi@dieei.unict.it
†Department of Electronic Engineering, University of Roma Tre, Via della Vasca Navale 84, 00146 Roma, Italy
Email: glosciuto@dii.unict.it
‡Department of Mathematics and Informatics, University of Catania, Viale A. Doria 6, 95125 Catania, Italy
Email: napoli@dmi.unict.it, tramontana@dmi.unict.it
§Institute of Mathematics, Silesian University of Technology, Kaszubska 23, 44-100 Gliwice, Poland
Email: marcin.wozniak@polsl.pl

***Abstract*—Nowadays the effective and fast detection of fruit defects is one of the main concerns for fruit selling companies. This paper presents a new approach that classifies fruit surface defects in color and texture using Radial Basis Probabilistic Neural Networks (RBPNN). The texture and gray features of defect area are extracted by computing a gray level co-occurrence matrix and then defect areas are classified by the applied RBPNN solution.**

***Keywords*-Co-occurrence Matrix, Texture Analysis, Pattern Recognition, Probabilistic Neural Network.**

## I. Introduction

IN ORDER to ensure the quality standard required in orange production lines, companies need trained people to inspect the fruits while they move in a conveyor belt. These experts classify oranges according to several categories based on visual features. However, for fast and precise quality standards, this approach is not competitive and sometimes unreliable. Therefore is paramount to use of an automatic system based on intelligent methods [1], [2], [3], [4], [5], [6]. The aim of this paper is to investigate the applicability of the Artificial Intelligence (AI) methods based on soft computing for detection of external defects of the fruits. Several methods based on more specific image acquisition are reported in [7], [8], [9], [10], [11], [12], [13]. Wen and Tao developed a near-infrared vision system for automatic apple defect inspection, see [14], While some recent advances in feature extraction for images and biometrics are reported in [15], [16], [17], [18], [19]. Zion et al. introduced a computerised method to detect the bruises of Jonathan, Golden Delicious, and Hermon apples from magnetic resonance images by threshold technique. The algorithm was only able to discriminate between all-bruised and non-bruised apples and was not applicable to on-line detection. Pla and Juste presented a thinning algorithm to discriminate between stem and body of the apples on monochromatic images. However, the task of classifying the calyx and defected parts in real-time was missing. Yang and Marchant used the ‘flooding’ algorithm for initial segmentation and ‘snakes’ algorithm for refining the boundary of the blemishes on the monochromatic images of apples. Miller et al. in [20] compared different neural network models for detection of blemishes of various kinds of apples by their reflectance characteristics.

From the said research, and according to the current literature, we can conclude that Multi-Layer Back Propagation (MLBP) gives proper recognition rates and also that increased complexity of the neural network system did not yield to better results [21]. Leemans, segmented defects of ‘Golden Delicious’ apples by a pixel-wise comparison method between the chromatic (RGB) values of the related pixel and the color reference model. The local and global approaches of comparison were effective, but more research was needed. In his further research work, Leemans [22] used a Bayesian classification method for pixel-wise segmentation on chromatic images of ‘Jonagold’ apples. Machine vision systems are successfully used for recognition of greenhouse cucumber fruit using computer vision [23]. A method for the classification and gradation of different grains (for a single grain kernel) such as groundnut, Bengal gram, Wheat etc., is described in [24]. The effect of foreign bodies on recognition and classification of food grains is given in [25]. Some researchers have used a neural network approach to the color grading of apples [26]. All these studies show that it is hard to define geometric or spectral properties for the fruit skin.

Most of the industry treatments such as washing and packing are highly automated, while the most important verification steps (e.g., inspection and grading in quality) are still performed manually, around the world. For this reason, we propose an efficient classifier based on fruit skin texture analysis, which can be implemented in manufacturing expert systems [27]. A new automatic classifier of oranges defects based on co-occurrence matrix and probabilistic neural networks is presented in the following sections. Orange is an important fruit of Mediterranean countries and it is well-known for its considerable anti-microbial, anti-viral, potent anti-oxidant properties. High-quality products are the basis for the success in today’s highly competitive market. Currently, manual inspection is being used in order to determine the orange quality. The increasing demand for quality assurance

978-83-60810-66-8/$25.00©2015, IEEE

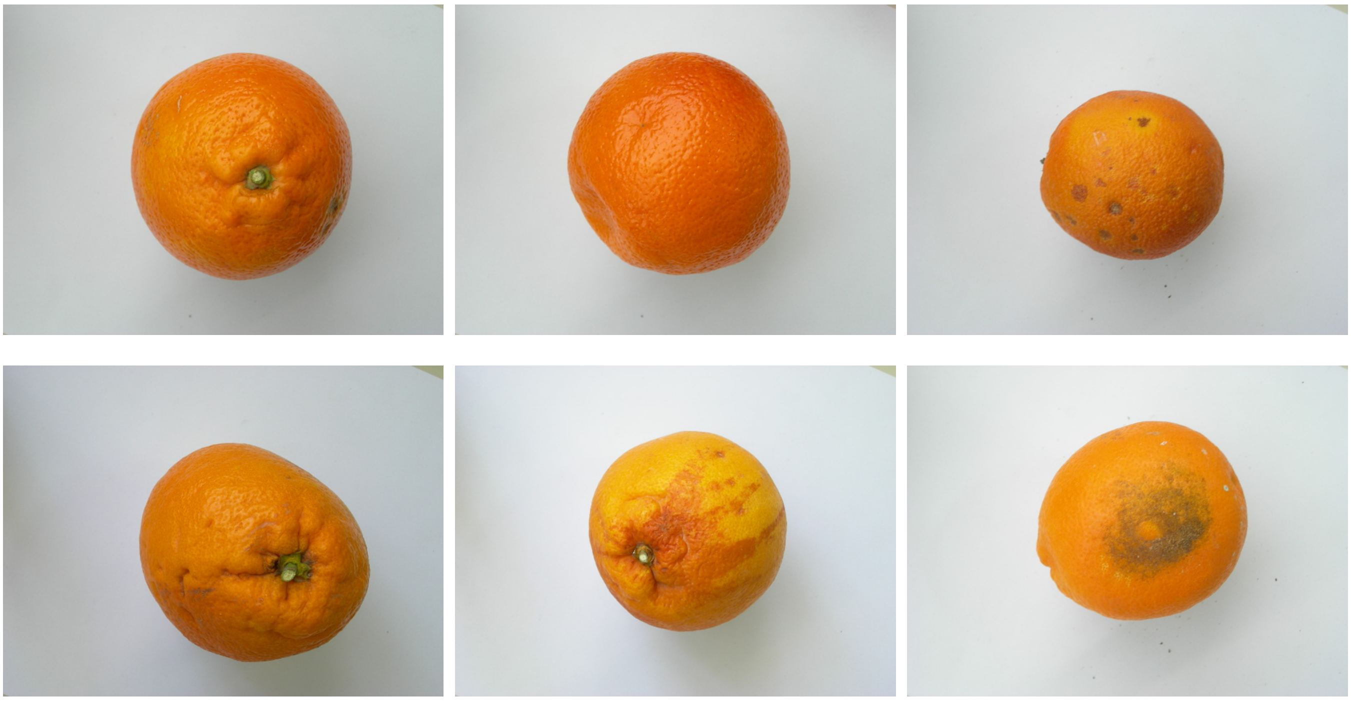


Fig. 1. Different sample of oranges used for this work. From left to right and top to bottom: Two normal oranges, an orange with several surface defects, a morphology defect, an orange with color defect, and an orange affected by a black mould.

requires simple and reliable sorting methods. The use of computer vision systems enables the detection of external quality defects. The objective of this study was to investigate the applicability of a method for detection of external defects and an automatic orange's classification system. For this purpose, the development of a classifier based on textural features of images, captured with a digital camera, was evaluated. The proposed technique is very robust and can identify specific defects like: surface defect, morphological defect, color defect, black mould or recognise a normal fruit. In fact, our RBPNN model is able to correctly attribute the samples to the correct defect groups with an overall error of $2.75\%$.

## II. Fruit classification Technique

In previous years, several types of image analysis techniques are applied to analyse the agricultural images such as fruits and vegetables, for recognition and classification purposes. The most popular analysis techniques that have been used for both recognition and classifications of two dimensional (2D) fruit images are color-based and texture-based analysis methods.

### A. Fruit classification based on shape

Shape based classification of fruits is based on various features like area, perimeter, major axis length and minor axis length. For calculating shape features an RGB image is converted into a gray scale image. After conversion into gray scale, the image represents a luminance intensity scale. There is a difference in intensity values for an object to be classified and its background, hence a threshold value is used to separate an object from its background. According to this threshold value, a gray scale image is converted into a binary image in which the value greater than the threshold is $1$ and the value lower than the threshold is $0$. With the help of this binary image different shape features are computed. The most common shape features are computed from the image area, perimeter, major axis length and minor axis length.

### B. Fruit classification based on color

RGB color space is converted into another color space such as HSV and for all the converted color space values, the mean and standard deviation are calculated. Each fruit image gives different values of mean and standard deviation, therefore assisting its classification.

*1) HSV Color Space:* HSI stands for hue, saturation and intensity. Then, for an image the color attribute is given by hue and the amount by which the pure color is diluted by white is given by saturation. The RGB components are separated from the original image, and the Hue (H), Saturation (S) and Intensity (I) components are extracted from RGB components. Equations (1), (2) and (3) are used to evaluate Hue, Saturation and Intensity of the image samples.

$$H = \begin{cases} \theta & B \leq G \\ 360 - \theta & B \geq G \end{cases}$$

$$\theta = cos^{-1}\left\{\tfrac{1}{2}\left(\tfrac{\lfloor (R-G)+(R-B) \rfloor}{\lfloor (R-G)^2+(R-B)\sqrt{G-B} \rfloor}\right)\right\} \tag{1}$$

The saturation component is given by:

$$S = 1 - \left(\frac{3}{R+G+B}\right)[min(R,G,B)] \tag{2}$$

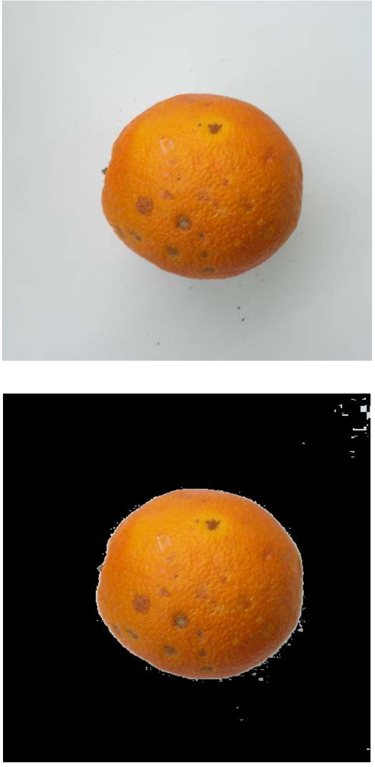


Fig. 2. An orange with some surface defects and the segmentation preprocessing result: the background is removed in order to analyse the orange surface only.

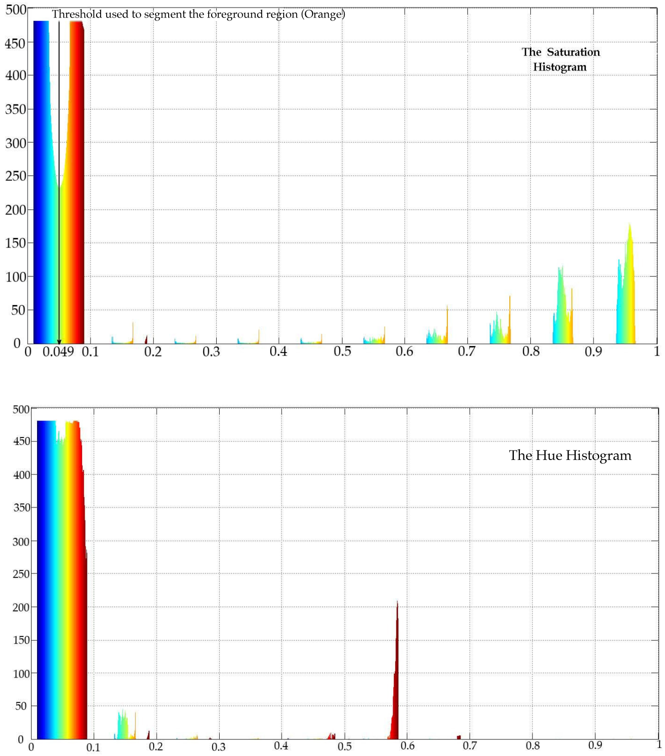


Fig. 3. The saturation (above) and hue (below) histogram related to the orange depicted in Fig. 2. The result shown in Fig. 2 is obtained by thresholding the saturation histogram shown here.

The intensity component is given by:

$$I = \frac{1}{3}(R + G + B) \tag{3}$$

### C. Fruit classification based on texture

Texture is classified by the spatial distribution of gray levels in a neighborood. It also helps in surface and shape determination. Gray level co-occurrence matrix is used to calculate different texture features [28]. There are two methods that can be used to calculate the texture feature of an image. One is the statistical texture analysis; the other is the structure of texture analysis. The former is the most conventional. Statistical texture analysis methods include spatial autocorrelation method, Fourier power spectrum method, Co-occurrence matrix method, gray level difference statistics method and trip length statistics method. Color mapping co-occurrence matrix (CMCM) is used to extract the texture information from a skin image. Gray level co-occurrence matrix (GLCM) is used to extract texture features in an image. It represents the form of tabulation which contains different combinations of pixel brightness values (gray levels) that occurs in an image. To calculate the different texture features like entropy, energy, homogeneity and dissimilarity, a gray level co-occurrence matrix is created.

## III. Classification Process of the Orange Defects

The European Union defines three quality classes (*extra*, *class I* and *class II*) for the fresh oranges with the tolerances of 5% and 10% by number or weight, respectively. The oranges in the *extra* class must be of superior quality with no defects or irregularity in shape, whereas the *class I* and *class II* can contain defects up to 1 cm$^2$ and 2.5 cm$^2$, respectively. All defect types in oranges contribute roughly equally to the final grading decision as local color, structural or textural variation of oranges.

In our classifier the defect detector uses a set of masks to recognise regions on the orange image. The defect is characterised by a discontinuity in the skin pigmentation. The features extracted from the orange images in either spatial or frequency domain can be used for classification. The external surface quality is directly related to the marketing and sales. Our automatic grading system can significantly improve the accuracy and consistency, while at the same time eliminating the subjectivity of manual inspection. The defects that we are able to detect on the external surface are caused by two reasons:

- Pre-harvest and Post-harvest diseases, like: diplodia and phomopsis stemend rot, splitting, pitting, green and blue mold, sour and brown rots, anthracnose, etc.
- Mechanical damages during transportation.

The defects on the orange fruit are characterised by different textures. Among various textures, we categorised defect types like:

- Pitting, which is caused by mechanical damage or reduced gas exchange during transportation. Pits can coalesce to form irregular patches and brown to black blemishes.
- Splitting is caused by the inability of the outer skin to hold the weight of the whole fruit. The outer skin of the citrus fruit splits and the inner pulp gets exposed. The defective region is usually brighter when compared with the normal skin.

For the research we used a data set obtained after processing high quality orange images and defective quality orange images, for a total of $400$ acquired images that are obtained by rotating and rearranging fruit samples. Color values (RGB-channels) as local features, are directly related with the images, so in our classifier they were introduced into the system without any change.

## IV. Proposed technique for orange feature extraction

The foreground region (orange) then can be segmented out from the background by using the Hue and Saturation histograms of the image. In order to extract the features, the pictures have been firstly segmented and the background removed. This latter stage is very fast and its computational weight can be neglected, since it only uses the saturation histogram and a saturation threshold having value $0.049$. Figure 2 shows an orange with several surface defects and the segmentation result, while Figure 3 shows the saturation and hue histogram of the same orange.

In addition to the segmentation using its HSV representation, we use color histogram and GLCM (Gray level co-occurrence matrix) to find quality classes corresponding to the orange qualities. For practical applications, the original input variables are typically preprocessed to transform them into some new space of variables where the pattern recognition problem will be easier to solve. This preprocessing stage is called *feature extraction*, see [29]. Preprocessing is performed in order to speed up computation and improve the classification performance [30]. We then selected useful features that are fast to compute and allow easy discrimination [31]. A feature set suitable for the classification should be insensitive to significant translations and have a very low correlation. Therefore, we base the extraction of features on the array of co-occurrence matrix, see [32]. Co-occurrence matrix is a single level dependence matrix that contains relative frequencies of two coordinate elements separated by a distance $d$. As you move from one pixel to another on the image, entries of the initial and final pixels become the coordinates of the co-occurrence matrix to be incremented, which in the end will represent structural characteristics of the image. Therefore, moving in different directions and distances on the image will lead to different co-occurrence matrices.

For each image we have calculated three co-occurrence matrices: one for each channel Hue, Saturation, Value. The obtained features represent the area of the orange, the background (the contrast, like gray level uniformity), gray level correlation between neighbours, sum average and sum variance. The features were calculated from the normalised co-occurrence matrix for $d = 1$ and four main directions: $0°$, $45°$, $90°$ and $135°$.

Since the use of the co-occurrence matrices leads to a course of dimensionality, we use five statistical descriptors of these matrices which use a co-occurrences gray level matrix. This is to calculate the number of adjacent pixels repetitions with the same gray level in the whole image. The statistical descriptors are the five textural features derived from the co-occurrence matrices.

*Angular Second Moment*: Consists of the sum of the squared elements in the co-occurrences matrix taken by pairs

$$\text{ASM} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left[P_1(i,j)\right]^2 . \tag{4}$$

*Contrast*: The global contrast of the image (also known as variance or *inertia*) measures the contrast intensity between a pixel and its neighbours

$$\text{Contrast} = \frac{1}{(N_g-1)^2}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-j)^2 P_1(i,j). \tag{5}$$

*Correlation*: Measures the relation of a pixel and its neighbours. The degree in which if the gray level of a pixel increases, its neighbour also increases

$$\text{Correlation} = \frac{\sum\limits_{i=1}^{N_g}\sum\limits_{j=1}^{N_g}(1-\mu_x)(j-\mu_y)P_1(i,j)}{\sigma_x\sigma_y}, \tag{6}$$

where the expected values are expressed as

$$\begin{array}{l} \mu_x = \sum_{i=1}^{N_g} i \sum_{j=1}^{N_g} P_1(i,j) \\ \mu_y = \sum_{j=1}^{N_g} j \sum_{i=1}^{N_g} P_1(i,j) \end{array} \tag{7}$$

and $\sigma_x$, $\sigma_y$ which are the variances of all the values of the class that feature belongs to

$$\begin{array}{l} \sigma_x^2 = \sum_{i=1}^{N_g} (i-\mu_x)^2 \sum_{j=1}^{N_g} P_1(i,j) \\ \sigma_y^2 = \sum_{j=1}^{N_g} (i-\mu_y)^2 \sum_{i=1}^{N_g} P_1(i,j) \end{array} . \tag{8}$$

*Gradient Module*: Is a measure of the degree of asymmetry of a distribution around the mean. It is obtained by calculating the third central moment of the distribution. If the obtained value is zero, it means that it is centered (like the normal distribution). If it is positive, it is asymmetrical to the right, and if it is negative, to the left

$$\begin{array}{l} \text{GM} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P_1(i,j) f(i)^2 \\ f(i) = i - N_g + 1 \end{array} . \tag{9}$$

*Intensity symmetry*: Is a measure of the variation of the texture if the grey levels are reversed

$$\text{IS} = 1 - \sum_{i,j=1}^{N_g} \left|P_1(i,j) - P_1(N_g-1-i, N_g-1-j)\right| . \tag{10}$$

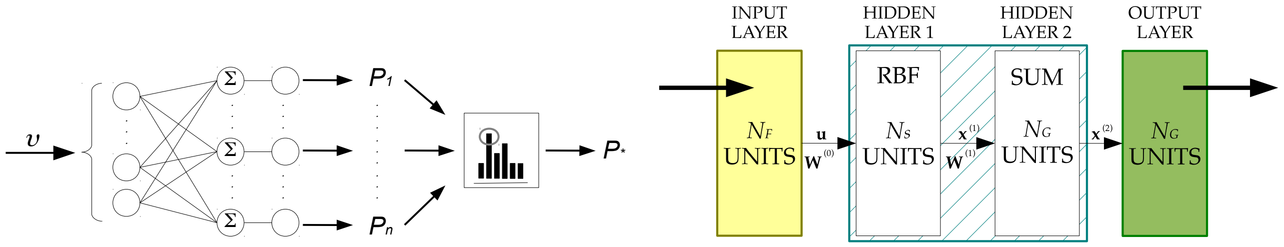


Fig. 4. A representation of our Radial Basis Probabilistic Neural Network (RBPNN) with maximum probability selector module. On the right the RBPNN layers model: $N_F$ is the number of features, $N_S$ is the number of samples and $N_G$ is the desired number defects to identify.

Firstly, we use the information from the co-occurrence matrices [33]. These matrices represent the spatial distribution and the dependence of the gray levels within a local area. Each $(i, j)$ entry of the matrices represents the probability of going from one pixel with a gray level $(i)$ to another with a gray level $(j)$ under a predefined distance and angle. More matrices are formed for specific spatial distances and predefined angles. From these matrices, sets of statistical measures are computed (called feature vectors) for building different texture models. In our case four angles, namely 0°, 45°, 90°, and 135°, are considered as well as a predefined distance of one pixel, in the formation of the co-occurrence matrices. Therefore, we have formed four co-occurrence matrices. A contrast equation (5) is used to calculate five parameters respectively, 0°, 45°, 90°, 135°, at limited distance one and 0° at a limited distance ten. This feature set extraction has a low dimensionality and a good discriminatory power. The features, e.g., mean and standard deviation, are extracted from each of the detail as well as the approximation sub-windows, and then fed into the Artificial Neural Network (ANN) classifier for defect classification. If the edges are not clearly defined, then the features extracted from such images are not useful. Training, testing, and validation of neural networks are performed using sample images.

## V. The Selected RBPNN

The proposed RBPNN (Fig. 4) is an implementation of a statistical algorithm called kernel discriminant analysis, see [34], in which the operations are organised into a multi-layered feed-forward network with four layers: an input layer, a pattern layer (the first hidden layer), a summation layer (the second hidden layer), and an output layer. Basically a RBPNN consists of an input layer, which represents the input pattern or feature vector. The input layer is fully interconnected with the hidden layer, which consists of the example vectors (the training set for the PNN [35], [36]). One other important element of the RBPNN is the output layer and the determination of the class for which the input layer fits. This is done through a winner-takes-all approach [37]. The output class node with the largest activation represents the winning class. While the class nodes are connected only to the example hidden nodes for their class, the input feature vector connects to all examples and therefore influences their activations. It is therefore the sum of the example vector activations that determines the class of the input feature vector. In RBPNN algorithm, calculating the class-node activations is a simple process. For each class node, the example vector activations are summed, which are the sum of the products of the example vector and the input vector, see [38]. Input neurons are used as distribution units that supply the same input values to all the neurons in the first hidden layer (such neurons are called *pattern units*). Each pattern unit performs the dot product $(\cdot)$ of the input pattern vector $\mathbf{u}$ by a weight vector $\mathbf{W}^{(0)}$. Then each pattern unit performs a nonlinear operation on the result. This nonlinear operation gives output $\mathbf{x}^{(1)}$ that is handed to the following summation layer. While the common sigmoid function is used for a standard FFNN, in our BPTA for presented PNN the activation function is exponential. Therefore for the $j$ neuron the output is where $\sigma$ represents the statistical distribution spread

$$\mathbf{x}_j^{(1)} \propto \exp\left(\frac{||\mathbf{W}^{(0)} \cdot \mathbf{u}||}{2\sigma^2}\right). \tag{11}$$

In the proposed model, while preserving the PNN topology [39], in order to obtain RBPNN capabilities, the activation function has been substituted with a Radial Basis Function (RBF). Moreover, there is the equivalence between the $\mathbf{W}^{(0)}$ vector of weights and the centroids vector of RBNN, which in our classifier are computed as the statistical centroids of all the given input sets. We name $\rho$ the chosen RBF, then the new output of the first hidden layer for the $j$ neuron is

$$\mathbf{x}_j^{(1)} \triangleq \rho\left(\frac{||\mathbf{u} - \mathbf{W}^{(0)}||}{\beta}\right), \tag{12}$$

where $\beta$ is the distribution shape control parameter, similar to $\sigma$ used in (11).

The second hidden layer in our RBPNN is a PNN. It computes weighted sums of received values from the preceding neurons. This second hidden layer is called summation layer with the output of the $k$ summation unit

$$\mathbf{x}_k^{(2)} = \sum_j \mathbf{W}_{jk}\mathbf{x}_j^{(1)}, \tag{13}$$

where $\mathbf{W}_{jk}$ represents the weight matrix. Such weight matrix consists of a weight value for each connection from the $j$ pattern unit to the $k$ summation unit. The summation units work as the neurons of a linear perceptron network. The training for the output in the applied model is performed similarly to RBNNs.

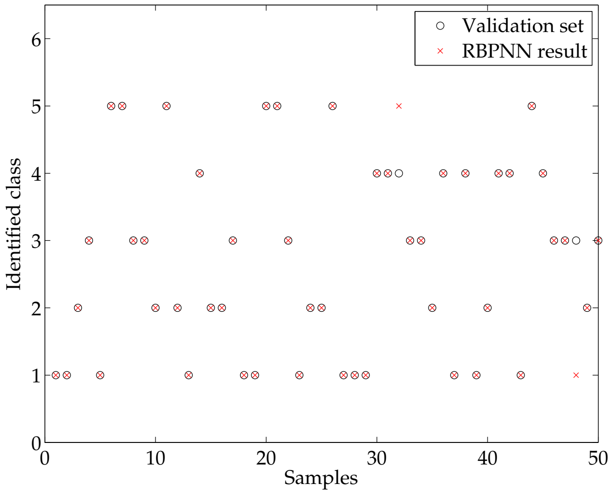


Fig. 5. The results of the identification proposed by our system at validation time: the circles show the effective class of the oranges, while the crosses show the identification output given by the implemented system.

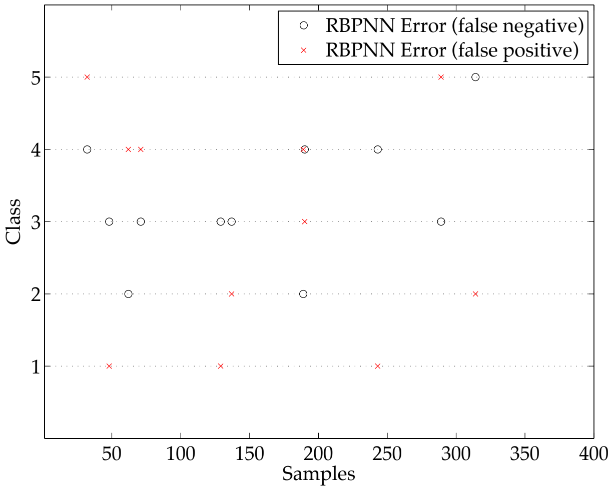


Fig. 6. The errors of identification of the proposed system: the circles identify the false negative results, while the crosses the false positive results.

## VI. Experimental results and conclusions

In our RBPNN classifier, defects were grouped into categories: surface defect as bruises (class 2), morphological defects (class 3), slight color defects (class 4), black mould (class 5) and a category of good fruit (class 1). The proposed classifier has been tested on a large number of image orange samples collected in a database. This section shows 400 samples of different defects of fresh orange surface include stab wounds, bruise, abrasion, sunburn, injury, hail damage, cracks and insect pest damage and good oranges. Fig. 6 shows the different recognitions. Our RBPNN model is able to correctly attribute the orange samples to the correct defect groups with an overall error of 2.75% (see Table I).

The proposed solution enabled us to obtain a fast and efficient automatic classification system for fruits defects. Although in an early stage of development, still such a system could be proposed on large scale for industrial applications, as well as for an implementation into programmable hardware.

TABLE I
A brief statistical analysis of the results by class.

| Class | Fruit condition | Number of samples | False negative | False positive |
|---|---|---|---|---|
| 1 | Normal fruit | 85 | 0 | 3 |
| 2 | Surface defect | 97 | 2 | 2 |
| 3 | Morphological defect | 71 | 5 | 1 |
| 4 | Color defect | 83 | 3 | 3 |
| 5 | Black mould | 64 | 1 | 2 |

| | Number of samples | Correctly classified | Overall error |
|---|---|---|---|
| **Overall classfication** | **400** | **389** | **2.75%** |

## References

[1] M. S. Pukish, P. Rózycki, and B. M. Wilamowski, "Polynet: A polynomial-based learning machine for universal approximation," *IEEE Trans. Industrial Informatics*, vol. 11, no. 3, pp. 708–716, 2015. [Online]. Available: http://dx.doi.org/10.1109/TII.2015.2426012

[2] P. D. Reiner and B. M. Wilamowski, "Efficient incremental construction of RBF networks using quasi-gradient method," *Neurocomputing*, vol. 150, pp. 349–356, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231214012946

[3] A. Horzyk, "How does generalization and creativity come into being in neural associative systems and how does it form human-like knowledge?" *Neurocomputing*, vol. 144, pp. 238–257, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2014.04.046

[4] M. Woźniak, W. M. Kempa, M. Gabryel, and R. K. Nowicki, "A finite-buffer queue with a single vacation policy: An analytical study with evolutionary positioning," *Applied Mathematics and Computer Science*, vol. 24, no. 4, pp. 887–900, 2014. [Online]. Available: http://dx.doi.org/10.2478/amcs-2014-0065

[5] A. Horzyk, "Innovative types and abilities of neural networks based on associative mechanisms and a new associative model of neurons," in *Artificial Intelligence and Soft Computing - 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14-18, 2015, Proceedings, Part I*, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., 2015, pp. 26–38. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-19324-3_3

[6] M. Woźniak and D. Połap, "On some aspects of genetic and evolutionary methods for optimization purposes," *International Journal of Electronics and Telecommunications*, vol. 61, no. 1, pp. 7–16, 2015. [Online]. Available: http://dx.doi.org/10.1515/eletel-2015-0001

[7] C. Napoli, G. Pappalardo, and E. Tramontana, "Improving files availability for bittorrent using a diffusion model," in *23rd IEEE International WETICE Conference*. IEEE, 2014, pp. 191–196. [Online]. Available: http://dx.doi.org/10.1109/WETICE.2014.65

[8] G. Capizzi, F. Bonanno, and C. Napoli, "Recurrent neural network-based control strategy for battery energy storage in generation systems with intermittent renewable energy sources," in *IEEE international conference on clean electrical power (ICCEP)*. IEEE, 2011, pp. 336–340. [Online]. Available: http://dx.doi.org/10.1109/ICCEP.2011.6036300

[9] F. Bonanno, G. Capizzi, S. Coco, C. Napoli, A. Laudani, and G. Lo Sciuto, "Optimal thicknesses determination in a multilayer structure to improve the spp efficiency for photovoltaic devices by an hybrid fem–cascade neural network based approach," in *International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM)*. IEEE, 2014, pp. 355–362. [Online]. Available: http://dx.doi.org/10.1109/SPEEDAM.2014.6872103

[10] C. Napoli, G. Pappalardo, E. Tramontana, and G. Zappalà, "A cloud-distributed gpu architecture for pattern identification in segmented detectors big-data surveys," *The Computer Journal*, p. bxu147, 2014. [Online]. Available: http://dx.doi.org/10.1093/comjnl/bxu147

[11] M. Woźniak, "Fitness function for evolutionary computation applied in dynamic object simulation and positioning," in *IEEE SSCI 2014: 2014 IEEE Symposium Series on Computational Intelligence - CIVTS*

*2014: 2014 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems, Proceedings*. 9-12 December, Orlando, Florida, USA: IEEE, 2014, pp. 108–114. [Online]. Available: http://dx.doi.org/10.1109/CIVTS.2014.7009485

[12] S. Fidanova, M. Paprzycki, and O. Roeva, "Hybrid GA-ACO algorithm for a model parameters identification problem," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, September 7-10, 2014.*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2014, pp. 413–420. [Online]. Available: http://dx.doi.org/10.15439/2014F373

[13] I. Lirkov, M. Paprzycki, M. Ganzha, S. Sedukhin, and P. Gepner, "Performance analysis of scalable algorithms for 3d linear transforms," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, September 7-10, 2014.*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2014, pp. 613–622. [Online]. Available: http://dx.doi.org/10.15439/2014F374

[14] Z. Wen and Y. Tao, "Dual-camera nir/mir imaging for stem-end/calyx identification in apple defect sorting," *Transactions of the ASAE*, vol. 43, no. 2, pp. 449–452, 2000.

[15] M. Woźniak and Z. Marszałek, "An idea to apply firefly algorithm in 2D images key-points search," *Communications in Computer and Information Science - ICIST'2014*, vol. 465, pp. 312–323, 2014.

[16] C. Napoli, G. Pappalardo, E. Tramontana, Z. Marszałek, D. Połap, and M. Woźniak, "Simplified firefly algorithm for 2d image key-points search," in *2014 IEEE Symposium on Computational Intelligence for Human-like Intelligence*. IEEE, 2014, pp. 118–125. [Online]. Available: http://dx.doi.org/10.1109/CIHLI.2014.7013395

[17] M. Woźniak and D. Polap, "Basic concept of cuckoo search algorithm for 2d images processing with some research results - an idea to apply cuckoo search algorithm in 2d images key-points search," in *SIGMAP 2014 - Proceedings of the 11th International Conference on Signal Processing and Multimedia Applications, Vienna, Austria, 28-30 August, 2014*, M. S. Obaidat, A. Holzinger, and E. Cabello, Eds., 2014, pp. 157–164. [Online]. Available: http://dx.doi.org/10.5220/0005015801570164

[18] M. Kubanek, D. Smorawa, and T. Holotyak, "Feature extraction of palm vein patterns based on two-dimensional density function," in *Artificial Intelligence and Soft Computing - 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14-28, 2015, Proceedings, Part II*, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., 2015, pp. 101–111. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-19369-4_10

[19] M. Kubanek, D. Smorawa, and M. Kurkowski, "Using facial asymmetry properties and hidden markov models for biometric authentication in security systems," in *Artificial Intelligence and Soft Computing - 13th International Conference, ICAISC 2014, Zakopane, Poland, June 1-5, 2014, Proceedings, Part II*, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., 2014, pp. 627–638. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-07176-3_55

[20] W. Miller, J. Throop, and B. Upchurch, "Pattern recognition models for spectral reflectance evaluation of apple blemishes," *Postharvest Biology and Technology*, vol. 14, no. 1, pp. 11–20, 1998.

[21] C. Napoli, G. Pappalardo, and E. Tramontana, "A hybrid neuro–wavelet predictor for qos control and stability," in *AI*IA 2013: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2013, vol. 9119, pp. 527–538. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-03524-6_45

[22] V. Leemans, H. Magein, and M.-F. Destain, "Defects segmentation on golden delicious apples by using colour machine vision," *Computers and Electronics in Agriculture*, vol. 20, no. 2, pp. 117–130, 1998.

[23] L. Zhang, Q. Yang, Y. Xun, X. Chen, Y. Ren, T. Yuan, Y. Tan, and W. Li, "Recognition of greenhouse cucumber fruit using computer vision," *New Zealand Journal of Agricultural Research*, vol. 50, no. 5, pp. 1293–1298, 2007.

[24] D. G. Savakar and B. S. Anami, "Recognition and classification of food grains, fruits and flowers using machine vision," *International Journal of Food Engineering*, vol. 5, no. 4, 2009.

[25] B. Anami and D. Savakar, "Effect of foreign bodies on recognition and classification of bulk food grains image samples," *J Appl Comput Sci*, vol. 6, no. 3, pp. 77–83, 2009.

[26] K. Nakano, "Application of neural networks to the color grading of apples," *Computers and electronics in agriculture*, vol. 18, no. 2, pp. 105–116, 1997.

[27] G. Ćwikła, A. Sekala, and M. Woźniak, "The expert system supporting design of the manufacturing information acquisition system (mias) for production management," *Advanced Materials Research*, vol. 1036, pp. 852–857, 2014.

[28] J. M. Keller, S. Chen, and R. M. Crownover, "Texture description and segmentation through fractal geometry," *Computer Vision, Graphics, and Image Processing*, vol. 45, no. 2, pp. 150–166, 1989.

[29] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, 2006, vol. 1.

[30] M. Woźniak, D. Połap, M. Gabryel, R. Nowicki, C. Napoli, and E. Tramontana, "Can we process 2d images using artificial bee colony?" in *Artificial Intelligence and Soft Computing*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, vol. 9119, pp. 660–671. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-19324-3_59

[31] B. Nowak, R. Nowicki, M. Woźniak, and C. Napoli, "Multi-class nearest neighbour classifier for incomplete data handling," in *Artificial Intelligence and Soft Computing*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, vol. 9119, pp. 469–480. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-19324-3_42

[32] V. Arvis *et al.*, "Generalization of the cooccurrence matrix for colour images: application to colour texture classification," *Image Analysis & Stereology*, vol. 23, no. 1, pp. 63–72, 2011.

[33] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

[34] B. D. Ripley, *Pattern recognition and neural networks*. Cambridge university press, 1996.

[35] C. Napoli, G. Pappalardo, E. Tramontana, R. Nowicki, J. Starczewski, and M. Woźniak, "Toward work groups classification based on probabilistic neural network approach," in *Artificial Intelligence and Soft Computing*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, vol. 9119, pp. 79–89. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-19324-3_8

[36] F. Bonanno, G. Capizzi, G. Lo Sciuto, C. Napoli, G. Pappalardo, and E. Tramontana, "A novel cloud-distributed toolbox for optimal energy dispatch management from renewables in igss by using wrnn predictors and gpu parallel solutions," in *Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), 2014 International Symposium on*. IEEE, 2014, pp. 1077–1084. [Online]. Available: http://dx.doi.org/10.1109/SPEEDAM.2014.6872127

[37] C. Napoli, G. Pappalardo, and E. Tramontana, "An agent-driven semantical identifier using radial basis neural networks and reinforcement learning," in *Proceedings of the XV Workshop Dagli Oggetti agli Agenti*, vol. 1260. CEUR-WS, 2014.

[38] M. M. Gupta, L. Jin, and N. Homma, *Static and dynamic neural networks, Hoboken*. NJ: Wiley-Interscience, 2003.

[39] F. Bonanno, G. Capizzi, and C. Napoli, "Some remarks on the application of rnn and prnn for the charge-discharge simulation of advanced lithium-ions battery energy storage," in *International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM)*. IEEE, 2012, pp. 941–945. [Online]. Available: http://dx.doi.org/10.1109/SPEEDAM.2012.6264500

DOI: 10.15439/2015F398
ACSIS, Vol. 5

# Minimum Variance Method to Obtain the Best Shot in Video for Face Recognition

Kazuo Ohzeki
Graduate School of Engineering
and Science, Shibaura Institute of
Technology 3-7-5 Toyosu,
Koutou-ku, Tokyo
Email: ohzeki@shibaura-it.ac.jp

Ryota Aoyama
Graduate School of Engineering
and Science, Shibaura Institute of
Technology 3-7-5 Toyosu,
Koutou-ku, Tokyo
Email: ma15001@shibaura-it.ac.jp

Yutaka Hirakawa
Graduate School of Engineering
and Science, Shibaura Institute of
Technology 3-7-5 Toyosu,
Koutou-ku, Tokyo
Email: hirakawa@shibaura-it.ac.jp

*Abstract*—**This paper describes a face recognition algorithm using feature points of face parts, which is classified as a feature-based method. As recognition performance depends on the combination of adopted feature points, we utilize all reliable feature points effectively. From moving video input, well-conditioned face images with a frontal direction and without facial expression are extracted. To select such well-conditioned images, an iteratively minimizing variance method is used with variable input face images. This iteration drastically brings convergence to the minimum variance of 1 for a quarter to an eighth of all data, which means 3.75-7.5 Hz by frequency on average. Also, the maximum interval, which is the worst case, between the two values with minimum deviation is about 0.8 seconds for the tested feature point sample.**

## I. Introduction

THERE are two major methods for face recognition [1]; one is the feature-based method which uses feature points at the endpoints of facial parts. The other is the holistic method, which processes the whole face region without decomposing regions into feature points. The former method has become unpopular because it is difficult to detect accurate feature points automatically. The latter method is now popular. However, Kathryn Bonnen and Anil K. Jain have recently proposed the effectiveness of a component-based method which uses facial parts in detail, rather than globally recognizing face information [2]. The holistic method conventionally analyzes the whole face region at once to achieve robust recognition. The component-based method utilizes separate regions of the eyebrows, eyes, nose and mouth, and performs dedicated recognition for each separate region, then integrates the results. The component-based method implies that it is now possible to detect facial parts before face recognition. The performance of the component-based method is better than that of the single holistic face recognition method.

The performance of face recognition has improved recently [3], and the results of various contests have been reported [4-6]. In those reports, recognition of frontal faces was considered as an easy task and more complicated conditions with age changes and with expressions are now targeted. Then, the basic face recognition technology is left in popular development activity. The recognition rate is now 97.5% for the frontal face without expressions [7], and it is still a difficult problem to realize 100% recognition for the frontal face without expressions.

One of the disadvantages of feature-based methods using feature points of face parts is that it is difficult to detect feature points correctly [1]. On the other hand, many devices are presented in the holistic methods with additional cases, such as faces with a non-frontal direction, under bad lighting conditions, and with facial expressions. However, for all cases, the recognition rate of the face is up to 99.90%, with a false acceptance ratio (FAR) of 1%. It is now difficult to reduce the FAR value since the recognition rate is at the upper limit.

In this paper, we consider the feature-based method because it provides high-precision results if the feature point detection works well and if its work successfully provides digital precision, while the holistic method views a whole face image with rough ambiguity at most. Recognition improves for a well-conditioned image from the frontal direction and without facial expression. We use moving video as the input and automatically extract the best shot from the frontal direction and without expression, and use the best shot images for registering the face image and matching input and registered face images. To obtain the best shot from an input video sequence, there are two methods using feature points. One involves maximizing the distance between feature points while viewing the input sequence. The other involves minimizing the variance of the distance between feature points. In this paper, the latter method will be described.

One of the methods using feature points of 3D presented by Drira et al. [8] performs 99.2% utilizing distances of 3D curves for faces without expressions. Another method using 3D mesh and the distance function in a wavelet-transformed domain under bad lighting conditions presented by Toderici [9] outperforms the 2D case. However, the recognition ratio is not so high. Guillaumin et al. presented an improvement using a learning method utilizing the nearest neighbor method to feature point distances [10].

978-83-60810-66-8/$25.00©2015, IEEE

## II. Previous Methods And Ideas For Improvement

In contests for face recognition, the recognition ratio for still pictures is reported to be 0.92 for rank 1 [4]. Further contests with age changes and with expressions are reported [5-6]. Also, most of these studies and reports settle some values of FAR (False Acceptance Ratio) or FRR (False Rejection Ratio), real distinguishing ability is not realized. From these results, the number of people that the face recognition system can distinguish without using any other information such as ID numbers is several hundred to a thousand. The so-called face-pass that permits a person to pass an authentication gate with only a face identification system without using an ID card does not make a service for more than thousand people. To improve the basic distinguishing ability of the number of people recognized and increase it above a thousand is the object of this paper.

To improve the recognition ratio, it is important to enhance the processing methods in the first stage of the system to acquire larger distances of feature values between people, not to incorporate additional conditions of age changes, expressions lighting conditions etc.

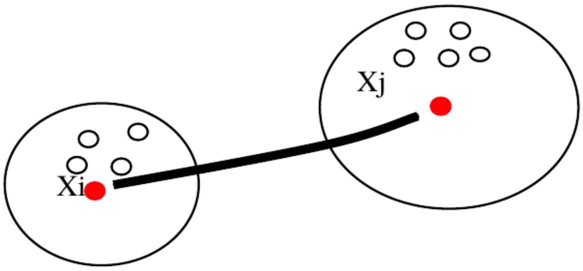


Fig.1 Feature points with the same name neighborhood[10].

## III. Face Recognition Using Feature Point Distance

### *A. Proposed system[11][12]*

Fig.2 shows the proposed face recognition system. Using input N frames, the best shot frame is selected by the method to be described in the following section. The best shot frame is registered in a database before recognition. At recognition, the best shot frame is selected and is verified using data in the database. The feature points of face parts are detected from the input face image. The detection is performed using software developed by Milborrow et al. [13] [14]. The software is based on the Active Shape Model and detects 77 feature points on the frontal face image. After detection of the feature points, two compensation operations of rotation and scale normalization are performed.

*A(1) Rotation compensation[12]*

The rotation compensation is to rotate at an angleθ which is a slope between the edges of both eyes. The rotation operation matrix shown in (1) is applied to the (x,y) coordinates of all feature points.

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (1)$$

After this rotation compensation, both eyes are aligned horizontally.

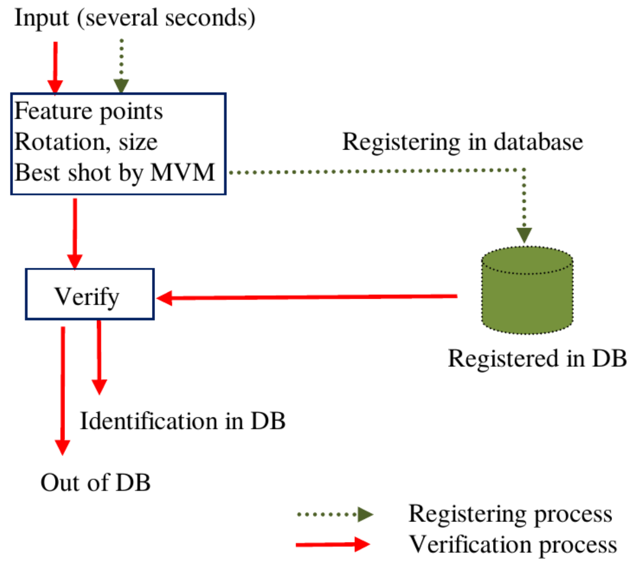


Fig.2 Face recognition system

*A(2) Size and position normalization*

Next, normalization according to the size is performed. This operation is applied to all feature points at the same rate as changing the size of a specified face part to a fixed size. In actual, let NF be

NF=sc/(X coordinate of the left edge of the left eye – X coordinate of the right edge of the right eye) (2)

Then, multiply this NF to all X and Y coordinates.

Position normalization is applied by moving all points in parallel after the size normalization. Let the input X and Y coordinate values of the first feature point be (In(0),In(1)) and the registered values be (Reg(0),Reg(1)), then the parallel movement value for X is Reg(0)-In(0), and for Y is Reg(1)-In(1)).

### *B. Best Shot Detection methods from video*

An advantage of using video for face recognition is that it can utilize well-conditioned data and discard poorly conditioned data. Therefore, video provides temporal continuity, so classification information from several frames can be combined to improve recognition performance [15]. Also, tracking of detected facial regions is possible and the system can be expanded to carry out facial expression detection [16].

Two methods of obtaining the best shot are considered. One is maximizing the length between the two feature points. Another is removing the value with greatest deviation to minimize variance.

*B (1) Maximizing Length Method*

A moving head in the input video shows a three- dimensional rotation pattern. Figure 3 is a face with a frontal direction and without a facial expression, which is the best shot. We will detect this kind of best shot image from all varying data. As for the Y and X axes rotations, the distances between feature points can be reduced by three-dimensional displacement and there is no data for compensation from the single camera environment. The Maximizing Length Method involves selecting a frame in which the distance between feature points is the maximum in all data.

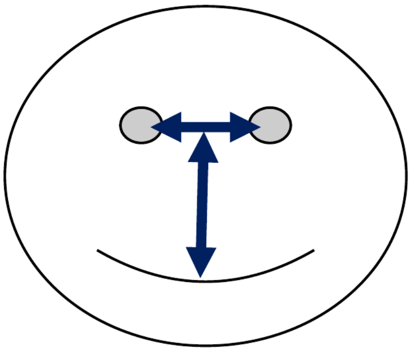


Fig.3 Regular face with frontal direction and without facial expression.

### *B (2) Minimizing Variance Method*

The Minimizing Variance Method is used to remove irregular data from the total data to obtain concentrated data with smaller variance. The method actually removes iteratively the largest value distant from the temporal average value and makes a new data set. Figure 4(a)-(d) shows the iterative removal status for the case of a pre-obtained time sequence data set of feature points. The asterisks in Figure 4(a)-(c) indicate the largest value at each stage. A removed value is replaced by a dotted line. From Figure 4(a) to (c), two values are removed one after another, and the third largest value is marked in Figure 4(c). The same processes are repeated to reach Figure 4(d), which shows a temporally small variance as if the iteration may converge to a fixed average with the smallest variance.

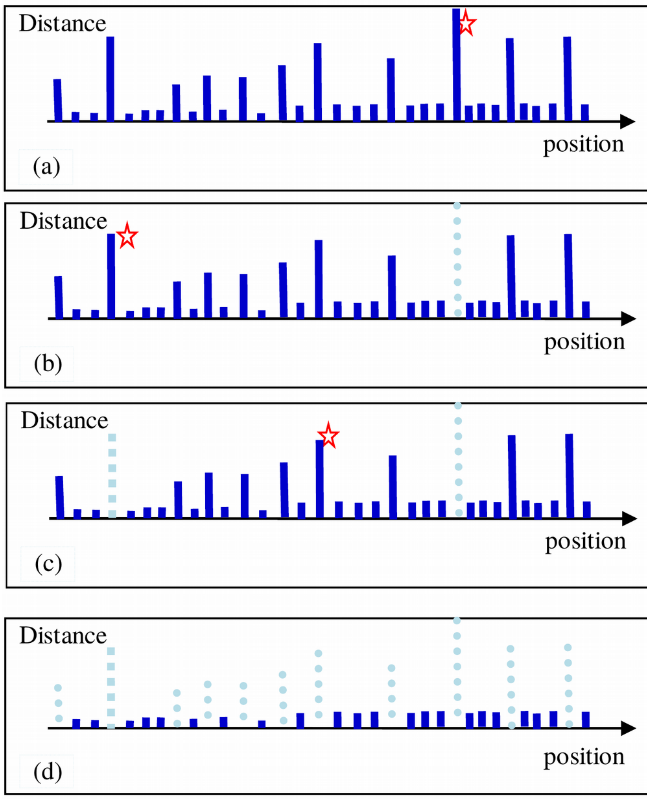


Fig. 4 Variance is reduced by removing the value with the greatest deviation.

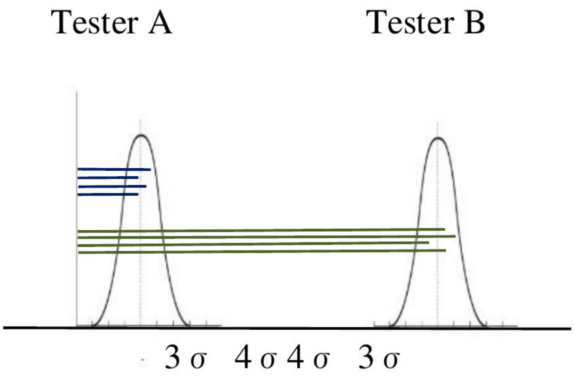


Fig.5 Distribution of two different persons.

### *C Estimation of the number of distinguishable people*

To distinguish one person from all the other persons using the detected feature points, the values should be clearly different to each other. To reduce the FAR and FRR, the detected feature point values should have meaningful distances. Here, we try to estimate the distance distribution for a larger number of testers from a smaller number of testers. The distance between two feature points on a face is a feature value. Let us consider a distribution of this feature values that are length data. We assume this distribution is normal with mean m and standard deviation σ.

Fig.5 shows a one-dimensional distribution for a feature value of a distance between two values. The distribution is made from the data of many testers. When this distribution is normal, the value of a sample taken from the data set randomly is considered to form a random sequence with this distribution.

Fig. 6 shows an example. This shows the minimum distance among data taken one after another from data with a distribution of mean m and standard deviation σ.When this minimum value goes below a threshold, i.e. 1, we cannot distinguish at least one person among the data using the feature value. When the minimum value remains larger than the threshold, we can distinguish all people in the taken samples. Fig.6 shows the case of m=16, σ=4, 8, 16, 32, 64. The horizontal axis is the number of samples taken. This graph shows that the minimum value goes below the threshold for more than 5-20 samples.

## IV. EXPERIMENTS

Four experimental items are carried out in the following sections. The Minimizing Variance Method (MVM) described in III B(2) is introduced in experiments to automati-

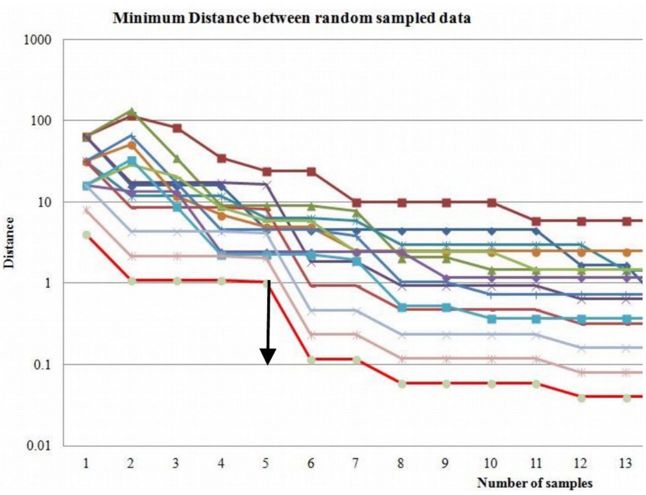


Fig. 6 Minimum distance among sampled random data. m=16, σ=64,32,16,8,4. This graph is based on a model generated by random samples from the normal distribution. Five values can be sampled keeping the distance values larger than 1 for all cases, which will be used in a later section, to allow taking samples five times. The width of windows is restricted to 13.

cally obtain frontal face images without expressions from input moving video of a person speaking for face recognition.

### A. Reducing variance by MVM

First, by adapting MVM described in III B(2) to use with a recorded part of a video with a length of about one minute, the convergence status of variances is investigated. The number of pieces of distances between feature points is the combination of two out of 77 feature points, that is $_{77}C_2=2926$ , where the total number of feature points is 77. For all this length data, let an initial value of the number of frames according to the time direction be “n”. The algorithm involves the following three steps;

a. Obtain the average length for time direction with the total number of the lengths of n.
b. Remove the largest value that is distant from the average value.

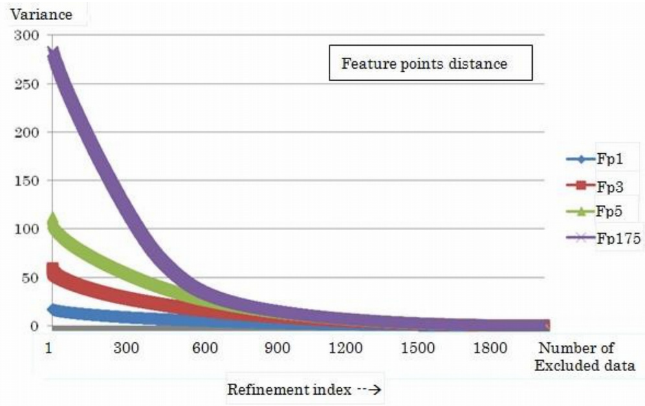


Fig.7 Variances of four distances between feature points converge as deleting iterations.

c. We get a new set of length data in which the number of lengths is subtracted by 1. (n=n-1)

This process is repeated until we get n=1.

Figure 7 shows four data which were randomly chosen from 2926 results with graphs of variances as the process above proceeded. The variances at the position of 1000 times of iteration, which is about half the number of total iterations are greatly reduced. Also, at the position of 1500 times, which is about a quarter of the total number, the variance is almost below 1-2. Other results show nearly the same tendency, as shown in the graphs.

### B. Variation by the positions of feature points

Figure 8 shows all 2926 distance values between feature points with well converging status according to the number of iterations. The numbers of iterations are from 500 to 1935, and their results are displayed overlapped. Figure 8(a) shows variances with the iteration numbers of 500 and above. Fig. 8(b) shows 1000 and above, Fig. 8(c) shows 1500 and above. These Figures show the variances reduce to small values after 1500 iteration. And after 1750, almost all variance data go below 2, and half of them seems to be below 1. A quarter to an eighth of all data seem to be below 1, which the input data are converge to stable values for register and recognition.

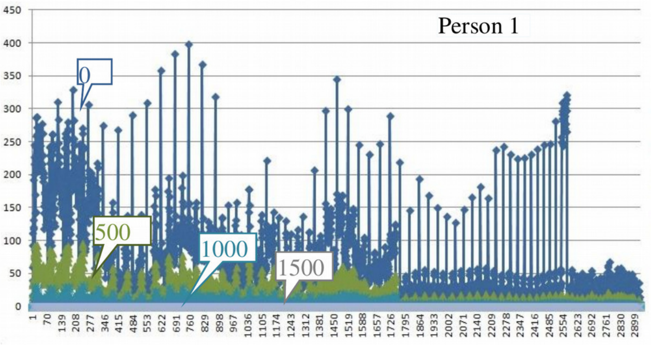


Fig.8 (a) Variances of distances between feature points converge as the number of iterations increases from 500 to 1935.

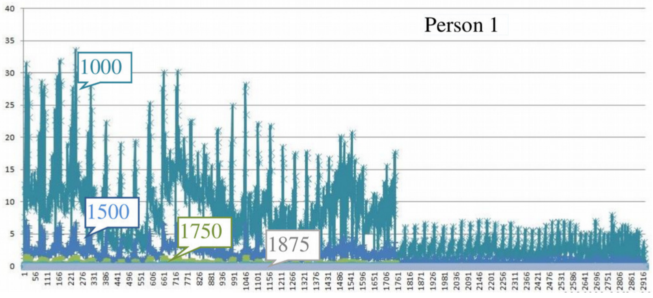


Fig.8 (b) Partial variances of distances between feature points converge as the number of iterations increases 1500, 1750, 1875, 1935. Enhanced in the vertical direction.

### C. Maximum interval between minimum variance data

After adapting the MVM to the length data, a quarter to an eighth of the resulting data forms a set with a small variance of 1-2. This quarter means 3.75-7.5 Hertz in the time

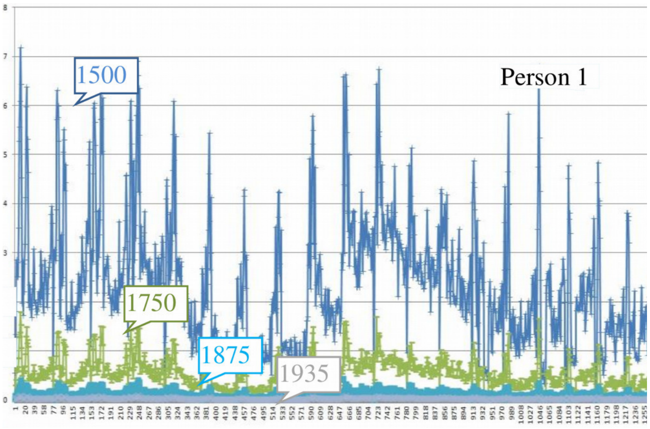


Fig.8 (c) Variances of distances between feature points converge as the number of iterations increases from 1500 to 1935. For the 1750 case, most feature values are below 2 and half of them are below 1.

direction because the original data was 30 Hertz. Face recognition can work at about 5Hz in the best conditions on average. But the value of 5Hz is average and we must consider the possible worst case. Then, the maximum interval between the converged data with small variances, which shows the worst case for detecting the best shot, is searched for. Figure 9 shows the maximum interval between the best shot frames vs. thresholds of deviation between the average value and the length value. The average values are the final values obtained by experiment III-A.

### D Estimation of the number of distinguishable people

For identification by face recognition, the number of distinguishable people will be estimated. The total number of feature values is 2926 as the pieces of distances between feature points. It is important to select feature values whose distance value is large enough to distinguish people. Fig. 10 shows a hundred feature values from the first number. The number of feature values that have a distance larger than 5 is 45. The distance value larger than 4 can be assumed to be the case of a standard deviation larger than 4 in Fig.6. From the curves of Fig.6, the number of valid samples to be distinguishable is 5. From this result, and if we take four more feature values (in total five feature values) we can find the number of distinguishable people is $5^5=3125$ [17]. If we take one more feature value it will be $5^6=15625$ .

We can estimate that there are at least five feature values from the different face parts of an eye, an eye brow, a mouth, a nose, and a contour, though the total 2926 data points may have correlations and are not independent of each other.

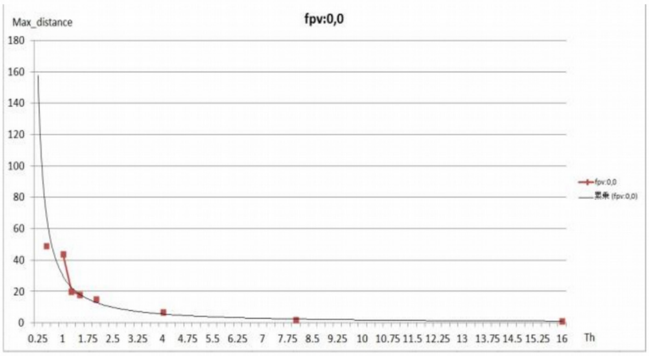


Fig.9 The maximum interval between the best shot frames vs. thresholds of deviation between the average value and length value. Dots are measured intervals. The value belongs to the best shot if it is smaller than the threshold.

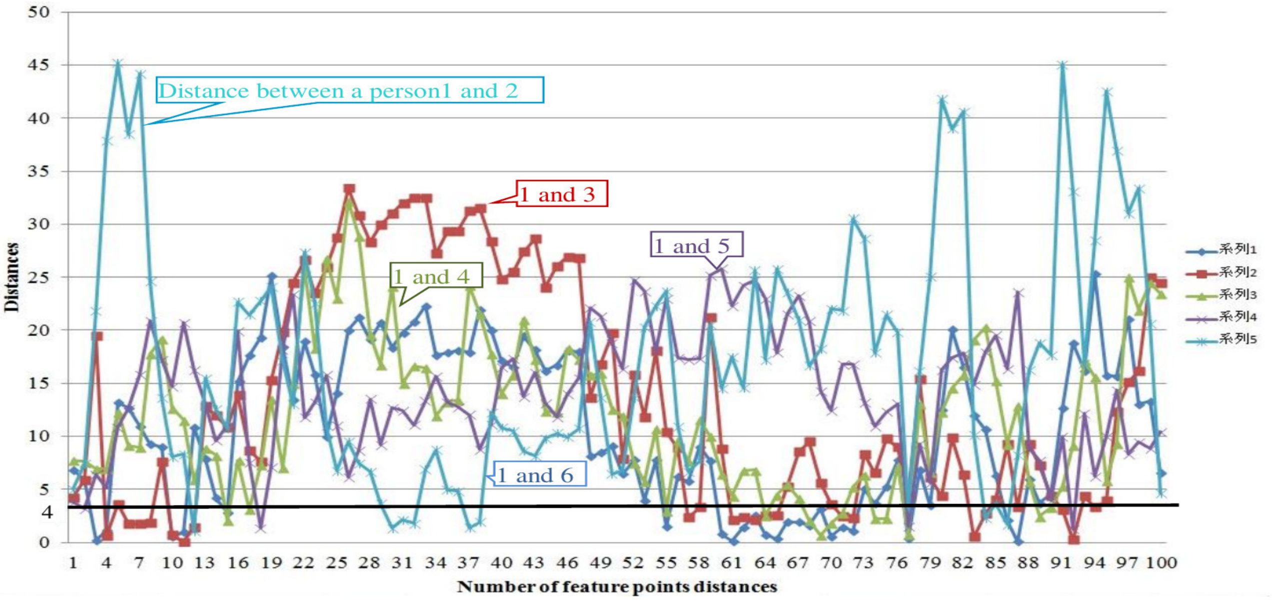


Fig.10 Five distance values obtained by six testers. The horizontal axis is the numbering of 100 feature values out of 2926. The feature values that have a distance of more than 4 are good distinguishable features for recognition.

## V. CONCLUSIONS AND FUTURE WORK

A new method of extracting the best shot from moving video input for face recognition is proposed. The best shot is obtained at the average probability of 1/4 to 1/8, which means 3.75-7.5 Hz in the time direction on average. The best shot can be obtained for all 2926 combinations of feature points. The maximum interval between the best shots is 0.8 second. According to Fig6, feature values larger than 4 can be used five times when sampling people. According to Fig.10, more than five feature values can be used. The number of distinguishable people can be estimated up to 3125 based on normal distribution. If we take six features, it implies the features distribute within 4σ.

In the future, how to select more reliable feature values for 5σ will be studied.

## REFERENCES

[1] Rabia Jafri and Hamid R Arabnia, "A Survey of Face Recognition Techniques", *Journal of information Processing Systems* Volume: 5, No: 2, pp. 41-68, 2009.

[2] Bonnen, K. Klare, B.F. Jain, A.K., "Component- Based Representation in Automated Face Recognition", *IEEE Transactions on Information Forensics and Security,* Vol.8, No.1 pp.239-253, Jan. 2013.

[3] JCB 2014 Conference report http://www.ijcb2014.org/IJCB%20Conference/IJCB14_Conference_Report.pdf

[4] Patrick J. Grother; George W. Quinn; P J. Phillips, "Report on the Evaluation of 2D Still-Image Face Recognition Algorithms", *NIST Interagency/Internal Report (NISTIR)* – 7709 June, 2010.

[5] M. Ngan and P. Grother, "Face Recognition Vendor Test (FRVT) Performance of Automated Age Estimation Algorithms", *NIST Interagency Report* 7995 Mar 2014.

[6] Patrick Grother Mei Ngan, "Face Recognition Vendor Test (FRVT) Performance of Face Identification Algorithms" *NIST Interagency Report* 8009 May 2014.

[7] Carl Gohringer, "Advances in Face Recognition Technology and its Application in Airports", *Allevate Limited.* Pp.1-10., 17 Jul, 2012.

[8] Drira, H. Ben Amor, B. ; Srivastava, A. ; Daoudi, M. ; Slama, R.," 3D Face Recognition under Expressions, Occlusions, and Pose Variations", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 35, ,Issue: 9, pp.2270 – 2283 Feb. 2013.

[9] Toderici, G.; Passalis, G. ; Zafeiriou, S. ; Tzimiropoulos, G., "Bidirectional relighting for 3D-aided 2D face recognition", *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2721-2728, 2010 June 2010

[10] Guillaumin, M. ; Verbeek, J. ; Schmid, C.," Is that you? Metric Learning Approaches for Face Identification", *Proc. of IEEE 12th International Conference on Computer Vision,* 2009 pp. 498–505, Sept. 2009

[11] Kazuo Ohzeki, YuanYu Wei, Yutaka Hirakawa, Toru Sugimoto, "Authentication System using Encrypted Discrete Biometrics Data", *Proceedings of TRUST 2014 Greece Springer LNCS* 8564 pp.210-211 June 30-July2 2014.

[12] ]Kazuo Ohzeki, Masahiro Takatsuka, Masaaki Kajihara, Yutaka Hirakawa, Kiyotsugu Sato, "On the False Rejection Ratio of Face Recognition Based on Automatic Detected Feature Points", *Proc. international workshops on "Pattern Recognition and Image Understanding"OGRW-9* Mo.3-1, ogrw2014_024_Ohzeki.pdf Dec.2014.

[13] Stephen Milborrow, Fred Nicolls, "Locating Facial Features with an Extended Active Shape Model", *Proceeding of ECCV Part IV* pp.504-513, Springer-Verlag Berlin, Heidelberg 2008

[14] S. Milborrow and F. Nicolls, "Active Shape Models with SIFT Descriptors and MARS", *International Conference on Computer Vision Theory and Applications (VISAPP)* pp.380-387. 2014

[15] Howell and H. Buxton, "Towards unconstrained face recognition from image sequences," in *Proceedings of the Second IEEE International Conference on Automatic Face and Gesture Recognition,* 1996, pp.224-229.

[16] L. Torres, "Is there any hope for face recognition?" in *Proc. of the 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2004)*. Lisboa, Portugal, 2004.

[17] Yasuko Tanaka, Eigo Miyazaki, and Kazuo Ohzeki, "Feature Point Analysis Using Facial Parts for Face Recognition", *National Convention, D-12-36* Institute of Electronics, Information, and Communication Engineers Mar. 2011 (in Japanese)

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 875–881

DOI: 10.15439/2015F158
ACSIS, Vol. 5

# Pedestrian tracking in video sequences: a particle filtering approach

Mateusz Owczarek, Przemysław Barański, Paweł Strumiłło
Institute of Electronics, Lodz University of Technology
Email: {mateusz.owczarek, przemyslaw.baranski, pawel.strumillo}@p.lodz.pl

***Abstract*—In this work we study the methods for pedestrian tracking in video sequences and indicate various applications of these methods ranging from surveillance systems to aiding the visually impaired persons. First, we define the general problem of object tracking that comprises the tasks of object detection, identifying the flow of object location in consecutive video images and finally analysis of the tracked trajectory data. We review the well known object tracking techniques i.e. the Mean-Shift and the CAMSHIFT algorithm and discuss their properties. Then we introduce the computational technique known as particle filtering (PF) and explain how we have applied it to the tasks of pedestrian tracking. We compare the PF approach against the Mean-Shift and the CAMSHIFT algorithms in terms of tracking robustness and the required computational demand. We conclude, that on the tested video sequences, the PF tracker outperforms the Mean-Shift and by a small margin the CAMSHIFT algorithm. The PF tracker requires more computational power, however, its tracking performance can be flexibly adjusted to the application requirements.**

## I. Introduction

HUMAN BEINGS, whether in a standstill or in motion, have an extraordinary visual capability of detecting objects and tracking them in the environment. This powerful property of the human visual system allows people, e.g. to manoeuvre in crowded pavements without bumping into other pedestrians. Implementation of object tracking functionality in computer vision systems is a challenging task and has attracted researchers' interest for decades. This is because there are numerous applications in which object tracking is important. They range from civilian applications (human-computer interaction, robotics, surveillance systems, crowd sourcing systems) to military applications (guided missile systems) [1], [2], [3]. Tracking of objects in video sequences is a particularly difficult task due to the following reasons:

- varying illumination of the monitored scene,
- loss of depth information in mono-camera image acquisition systems,
- varying size and shape of the tracked object (due to changes in orientation and distance to the camera),
- occlusions of the tracked objects,
- motion of the tracking camera (i.e. both the tracked object and the background move in reference to the camera).

The object tracking task can be subdivided into the three major steps:

1) Object detection in a scene and determining its location (i.e. applying methods for segmenting out the object of interest from the background).
2) Identifying object position changes in consecutive image frames, termed object tracking.
3) Analysis of the object tracking data (e.g. determination of the motion trajectory, path prediction, etc.).

The latter processing step strongly depends on the application. Our research goal is to develop a system that would serve as a vision based travel aid for the visually impaired. Although, a number of GPS-based navigation systems or remote guidance systems were designed especially for the visually impaired, they are expensive and offer poor position accuracy in urban environments [4]. Also, electronic travel aids in which ultrasound or laser sensors are embedded into white canes and also more advanced sensory substitution solutions (e.g. using auditory display techniques) have not found wider acceptance among the visually impaired users [5]. Our approach to aiding the visual impaired in mobility and travel is to track the position of a blind pedestrian in a city environment on the basis of video sequences. Then the positioning data will be integrated with the digital map data to work out the navigation instructions for the blind user. Such a solution does not require any additional hardware to be carried by a blind pedestrian except for a mobile phone that would serve as a communication device between the system and the user.

In this communication we report on our preliminary studies aimed at developing and testing robustness of the vision based object tracking methods with special focus on systems used for pedestrians' tracking. In Section II we review the related work and highlight the widely used Mean-Shift and CAMSHIFT object tracking algorithms. In Section III we introduce the particle filtering algorithm and explain how we apply this powerful computing technique to person tracking. In section IV we describe the experimental tests of the algorithms and evaluate their performance on example object tracking tasks.

## II. Review of related work

### A. Object detection methods

Every tracking method requires an object detection technique to distinguish the tracked object from the background and/or other objects. The result of tracking strongly relies on the applied object detection method. These methods can be subdivided into four categories [1]:

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 643636 "Sound of Vision."

978-83-60810-66-8/$25.00©2015, IEEE

- **Point detectors** are used to detect characteristic points in a processed image (i.e. corners, edges). Among the most commonly used methods are: the *Moravec's operator*, the *Harris point detector* and the *Scale invariant Feature Transform* (SIFT). A comparative survey [6] provides more information on the topic.
- **Segmentation** leads to partitioning an image into characteristic regions (i.e. perceptually similar regions that can be used for tracking). The *Mean-shift* clustering [7], [8], the *Graph-cuts* and *Active contours* are within the most relevant image segmentation techniques.
- **Background modelling** is based on an assumption that moving objects appear on a very largely stable background [9], therefore the foreground can be obtained by "subtracting" the estimated background image from the current frame.
- **Supervised classifiers** that are based on a large collection of labeled samples composed of feature vectors. Samples are used as training data for the supervised learning method to generate a function that maps inputs (e.g. object features) to the desired outputs (e.g. class labels unambiguously binding the object with given features). These methods include neural networks, *adaptive boosting* (e.g. *Viola-Jones object detection* framework used to detect pedestrians [10]), *decision trees* and *Support Vector Machines* (SVM) in many variations [3].

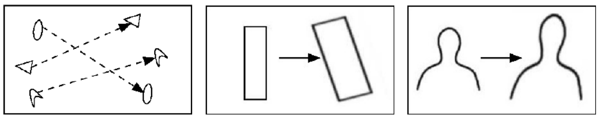
Among the image features commonly used in object tracking are [1], [11]:

- **Color** is used as a feature in histogram-based methods, where the object is represented by its appearance. Color spaces such as *L*a*b* and *HSV* (Hue, Saturation, Value) are more preferred in image processing, due to more perceptual uniformity.
- **Edges** are strong changes in the intensity or color in an image generated by object boundaries. Edges are less sensitive to illumination changes than color features.
- **Optical flow** defines the apparent motion of an object by a dense field of displacement vectors across consecutive image frames.
- **Texture** of an object described by a number of properties (such as lightness, density, regularity, linearity, directionality, smoothness, etc. [12]) is an excellent feature to track.

### B. Approaches to object tracking

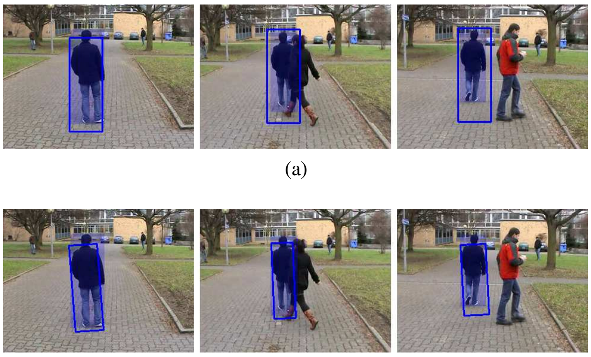
Object tracking techniques can be subdivided into the three major categories:

- **Point Tracking** relies on the positions and motion of points representing the target object in consecutive frames. Therefore, object tracking can be defined as the problem of finding points' correspondences (Fig. 1a).
- **Kernel Tracking** relies on the shape or appearance of the target object (referred to as *kernel*), which is represented by a geometric primitive (e.g. a rectangular patch or an ellipse), see Fig. 1b. The most popular representative of this category is the *Mean-Shift* [7], [8], [13] and its modification, the *CAMSHIFT* (*Continuously Adaptive Mean-Shift*) algorithm [14] (see Fig. 2).
- **Silhouette Tracking** relies on the information encoded inside the region of a tracked object which usually, due to the complexity of its shape, cannot be described well using simple geometric primitives (Fig. 1c).



(a) Multi-point correspondence (b) Parametric transformation of a rectangular patch (c) Contour evolution

Fig. 1. Different tracking approaches (source: [1])



(a)

(b)

Fig. 2. In the Mean-Shift algorithm (a) the search window size is fixed throughout the tracking session, whereas in the CAMSHIFT (b) the search window continuously adapts itself in size and orientation to fit the target object

The idea of applying particle filtering to object tracking was independently proposed by several research groups and is described in [15], [16] and [17] among others. Although the computational and probabilistic origins of the particle filter usually remains (more or less) the same, different approaches use different features to define the target model. Typically, edge-based image features are used [18], [15], [19], but color-based image features are also an option. The latter is even more robust against the out-of-plane rotations (e.g., when a person turns around), scale and rotation invariant [20], but more sensitive to the illumination changes (as indicated in Section II-A). For this reason, the color-based particle filter has become the object of our study.

Reference [2] presents recent trends in object detection, while [3] evaluates the state-of-the-art tracking algorithms.

## III. Application of Particle Filtering to person tracking

### A. Particle filter basis

Particle filtering, also called the *Sequential Monte Carlo* (SMC) [21] method, is a simulation-based technique which stems from the Monte Carlo method. The latter is a simple, yet effective, way of finding an optimal solution for multidimensional problems by randomly generating a large number of possible system states. This enables to observe the overall system behaviour and select the best solution.

In the particle filter approach, the distribution of the $\mathbf{s}_t$ is approximated by a set of so called particles. Every particle is represented by the vector:

$$\mathbf{c}_t^{(n)} = \left[\mathbf{s}_t^{(n)} \ \pi_t^{(n)}\right]^T \tag{1}$$

where the superscript $(n)$ denotes a particle number ranging from 1 to $N$ being the size of the particle set and $t$ denotes a time instant. Hence, the vector $\mathbf{s}_t^{(n)}$ represents the system state, that is the variables of interest, $\pi_t^{(n)}$ is a particle weight, i.e. a value that reflects how accurately a given particle approximates the system state.

The posterior distribution of $\mathbf{s}_t$ is approximated by the probability mass function:

$$p(\mathbf{s}_t|\mathbf{y}_{0:t}) \approx \sum_{n=1}^{N} \delta(\mathbf{s} - \mathbf{s}_t^{(n)}) \cdot \pi_t^{(n)} \tag{2}$$

where $\delta(\cdot)$ is the Dirac delta function.

The algorithm can be summarized in its basic form as follows [22]:

1) **Initialization**. At the algorithm's outset, all particles' states $\mathbf{s}_0^{(n)}$ are randomly initialized according to a given distribution. The weights $\pi_0^{(n)}$ are assigned equal values of $\frac{1}{N}$.
2) **Prediction.** New particle states are predicted on the base of the transition equation:

$$\mathbf{s}_t^{(n)} = \mathbf{f}\big(\mathbf{s}_{t-1}^{(n)}, \mathbf{u}_{t-1}, \mathbf{w}_{t-1}^{(n)}\big) \tag{3}$$

where $\mathbf{u}_{t-1}$ is a driving vector and $\mathbf{w}_{t-1}^{(n)}$ is the noise vector introduced to the state due to the measurement error of $\mathbf{u}_{t-1}$. Each particle is perturbed with an individually generated vector $\mathbf{w}_{t-1}^{(n)}$; $\mathbf{f}(\cdot)$ is the transition function that calculates a new state on the base of the previous one and the driving signals.
3) **Measurement update**. Each measurement $\mathbf{z}_t$ updates the weights of the particles by the equation:

$$\pi_t^{(n)} = \pi_{t-1}^{(n)} \cdot p\big(\mathbf{z}_t|\mathbf{s}_t^{(n)}\big) \tag{4}$$

where $p\big(\mathbf{z}_t|\mathbf{s}_t^{(n)}\big)$ is a conditional probability density of measuring $\mathbf{z}_t$ given the particle state $\mathbf{s}_t^{(n)}$. Particles that diverge in the long run from measurements will have small weights $\pi_t^{(n)}$.
4) **Weights' normalization.** For the sake of the next steps, the weights need to be normalized so that they sum up to 1:

$$\pi_t^{(n)} := \frac{\pi_t^{(n)}}{\sum\limits_{i=1}^{N} \pi_t^{(n)}}. \tag{5}$$

5) **State estimation**. The system state is the weighted average of all particles' states:

$$\bar{\mathbf{s}}_t = \sum_{i=1}^{N} \mathbf{s}_t^{(n)} \cdot \pi_t^{(n)}. \tag{6}$$

6) **Resampling**. After a number of algorithm iterations, all but a few particles have negligible weights and therefore do not participate in the simulation effectively. This situation is detected by calculating the so-called *degeneration indicator*, expressed by:

$$d_t = \frac{1}{N \sum\limits_{i=1}^{N} \left(\pi_t^{(n)}\right)^2}. \tag{7}$$

As the weights start to differ, the $d_t$ indicator decreases. If $d_t$ falls below a given threshold, then a process called resampling is evoked and a new set of particles is created. Resampling causes that the probability density function of $\mathbf{s}_t$ is refined in the next iterations of the algorithm. Thus, a better estimate can be found.
7) **Go to point 2**

### B. *Implementation of particle filter for person tracking*

The implementation of the particle filter tracker described herein is based on the idea behind the *ConDensation Algorithm* [15], once implemented in the OpenCV library, yet presently deprecated due to some imperfections in the resampling part.

Target object is represented by a rectangular patch (Fig. 4a). Such representation is suitable for representing simple non-rigid targets and works well in terms of kernel-based tracking methods [1]. Target model is defined by a (8×8×4 bins) HSV histogram of the target object's representation and an initial size of the representation rectangle (during the tracking process it is scaled in relation to its initial size). Reduced number of levels for the V-component of the histogram makes the method less sensitive to the changes in lighting conditions. The use of HSV histogram may, however, make the method more sensitive to noise [1]. Current version of the algorithm does not update the target model automatically, but a new target model can be built on-the-fly.

Due to the fact that the principle of operation of the particle filter was described in the preceding section, in the subsequent paragraphs we would like to limit ourselves just to the adjustments which make use of particle filter in object tracking.

*a) Initialization:* A set of $N$-samples is generated. Each sample represents the quantities of the target object's representation and is defined by a state vector and an initial weight:

$$\begin{aligned} \mathbf{s} &= [x, y, \frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}, k]^T \\ \mathbf{s}_0^{(n)} &= [x, y, 0, 0, 1]^T \text{ for } n = 1, ..., N \\ \pi_0^{(n)} &= N^{-1} \text{ for } n = 1, ..., N \end{aligned} \tag{8}$$

where $(x, y)$ are the coordinates of the center of the representation rectangle, $\left(\frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}\right)$ represents velocity vector (motion model), $k$ is the scale of the representation rectangle in relation to its initial size and $N$ is the number of samples.

*b) Evolution of the samples:* In every new frame a state of each sample is generated by a second order linear difference equation based on prior observations:

$$\mathbf{s}_t^{(n)} = \mathbf{A}\ \mathbf{s}_{t-1}^{(n)} + \mathbf{w}_{t-1}^{(n)} \tag{9}$$

where $\mathbf{w}_{t-1}^{(n)} \sim N(\mathbf{0}, \mathbf{Q}_t)$ is a vector of multivariate normally distributed random variates (stochastic component) and $\mathbf{A}$ is a *transition matrix* (deterministic component) defined for state vector (8):

$$A = \begin{bmatrix} 1 & 0 & \mathrm{d}t & 0 & 0 \\ 0 & 1 & 0 & \mathrm{d}t & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

*c) Weighting the samples:* Each sample represents a hypothetical location of the tracked object. To weight the sample set, a histogram of each hypothetical target representation is compared to the target model. To quantify similarity between the two histograms, the *Hellinger distance* is used:

$$\begin{aligned} d_t^{(n)} &= \sqrt{1 - \frac{1}{M\sqrt{\bar{H}_1 \bar{H}_2}} \sum_{u=1}^{M} H_1(u) \cdot H_2(u)} \\ \bar{H}_k &= \frac{1}{M} \sum_{u=1}^{M} H_k(u), \end{aligned} \tag{11}$$

where $H_1$ and $H_2$ are the histograms to be compared and $M$ is a total number of histogram bins. The result is a score in range $[0, 1]$, where the first value indicates a perfect match and the latter a complete mismatch. The distance is used in the measurement equation to update the weights of the samples, according to:

$$\pi_t^{(n)} = \pi_{t-1}^{(n)} \cdot p(\mathbf{z}_t | \mathbf{s}_t) \tag{12}$$

$$\pi_t^{(n)} = \pi_{t-1}^{(n)} \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{\left(d_t^{(n)}\right)^2}{2\sigma^2} \right), \tag{13}$$

where $\sigma$ is the standard deviation. The higher the weight, the higher chance that the particle will be drawn during the next iteration (particles with the lowest weights are to be replaced).

TABLE I
DETAILS OF THE TEST VIDEO SEQUENCE

| Parameter | Value |
|---|---|
| Format | 320 × 240 at 25 frames/s |
| Length | 1017 frames (∼ 41 sec.) |
| Target object | person |
| Keywords | moving cam, moving target, non-rigid target, rotation, similar distractors, full occlusion, outdoor |
| Link to dataset | [23] /datasets/seqI.zip |

The state of the tracked object at a time-step $t$ can be therefore estimated as the mean state [15]:

$$\bar{\mathbf{s}}_t = \mathrm{E}\left[\mathbf{s}_t^{(n)}\right] = \sum_{n=1}^{N} \pi_t^{(n)} \cdot \mathbf{s}_t^{(n)}. \tag{14}$$

*d) Re-initialization:* The process of propagation of samples is being continued until the mean state moves off the image or the probability of the mean state drops below a certain threshold. It usually means that the target object has been lost due to a mismatch between the predicted and actual motion. In such scenario, particles are redistributed in accordance with the continuous uniform distribution to re-acquire the target:

$$\mathbf{s}_t^{(n)} = \begin{bmatrix} x \\ y \\ \mathrm{d}x/\mathrm{d}t \\ \mathrm{d}y/\mathrm{d}t \\ k \end{bmatrix} \sim \begin{bmatrix} U(0, W_{image}) \\ U(0, H_{image}) \\ U(-0.05, 0.05) \\ U(-0.05, 0.05) \\ U(1.0, 2.0) \end{bmatrix} \tag{15}$$

where $(W_{image}, H_{image})$ is the size of the image.

## IV. Results of Experimental Tests

### A. Test results and performance measures

The implemented PF tracker has been tested and evaluated on a few short video sequences from the BoBoT benchmark on tracking dataset [23]. Here we present results obtained for one of the particularly difficult sequence which shows an individual walking along a pavement and is being followed by a camera (Fig. 4a). Ever and again a different person appears and passes by, shadowing the tracked pedestrian (so-called *occlusion*, cf. Fig. 4b), or is dressed in the same manner (so-called *similar distractor*, cf. Fig. 4c), trying to confuse the algorithm. Table I provides more details on the dataset.

The basic evaluation of the tracking algorithm relies on the overlap rate calculated every frame of the video sequence and defined as $\frac{\mathrm{area}(R_T \cap R_G)}{\mathrm{area}(R_T \cup R_G)}$, where $R_T$ is the rectangular region of the tracking result and $R_G$ denotes the ground truth [3]. If the overlap is larger than or equal to $1/3$, it is considered a hit (as shown in Fig. 5), otherwise a miss of target is denoted. Tests with a various number of particles show that $N = 100$ particles seems to be a good trade-off between accuracy and processing time of the algorithm (cf. Fig. 3). An interested reader is referred to the video sequence published at [25], which shows how the result of the tracking algorithm depends on the number of particles.

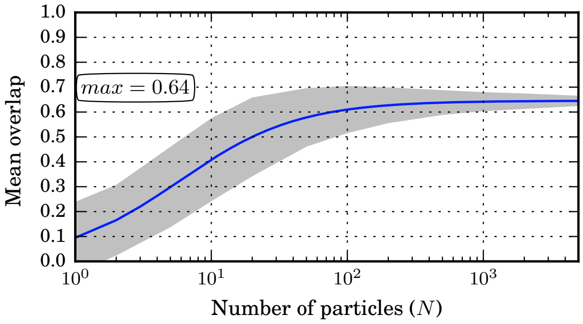
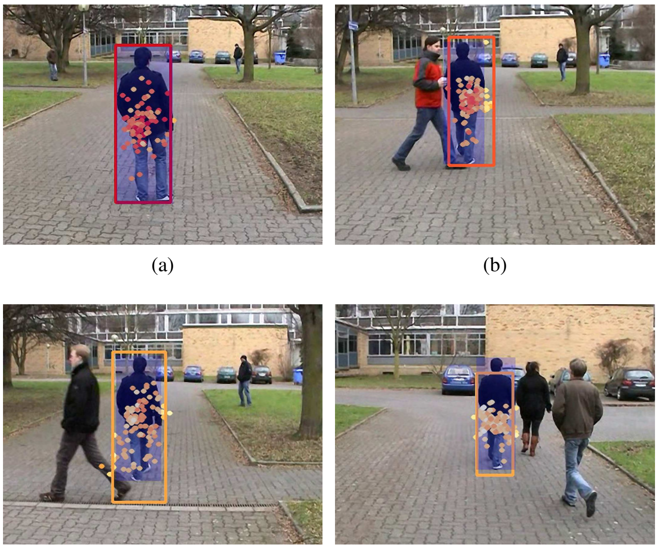


Fig. 3. The mean of ground truth and the tracking result overlap for $N = 100$ particles is equal to 0.62, which seems to be a trade-off between accuracy and processing time. Gray region is the standard deviation



(a) (b)

(c) (d)

Fig. 4. PF tracker in action — tracked person is being followed by "particles" (each particle represents a hypothetical location of the tracked object). The tracking result is an outline around the target. Blue semi-transparent rectangle is the ground truth (source: [24])

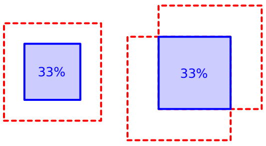
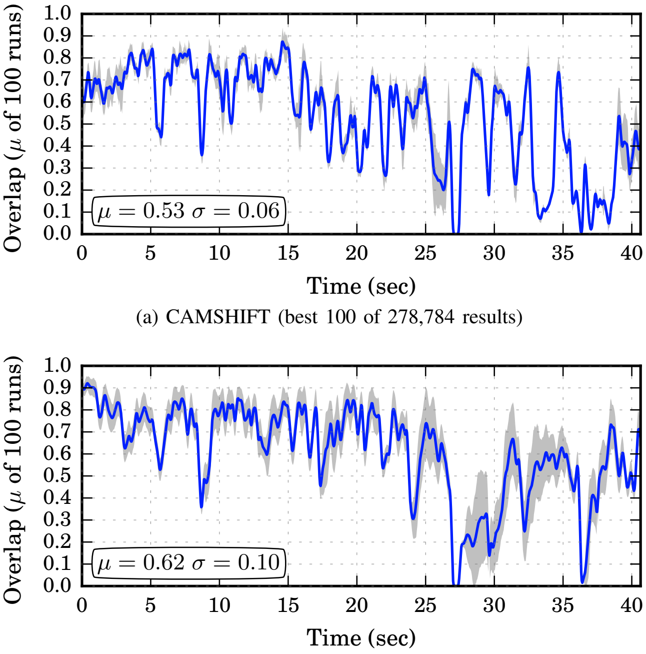


Fig. 5. Two examples of the ground truth (blue rectangle in the middle) and the tracking result (red dashed rectangle) overlap of 1/3



(a) CAMSHIFT (best 100 of 278,784 results)

(b) Implemented PF tracker (for $N = 100$ particles)

Fig. 6. The ground truth and tracking result overlap per frame for the test video sequence. Grey region is the standard deviation

Fig. 6b shows the overlap for each frame of the test video sequence (mean of 100 iterations) using $N = 100$ particles. At the beginning the overlap reaches up to 92% and oscillates around 80% for most of the time. Each drop below $1/3$ (a miss) is caused by a full occlusion, when the algorithm has lost the target and needed to be re-initialized. Slightly worse results in the last quarter of the video sequence are caused by the out-of-plane rotation of the tracked individual. After the person turned right, his right side is visible on the video instead of his back, on which the target model used to be built (as it was mentioned before, current version of the algorithm does not update the target model automatically, yet an adaptive approach is within the further work). Notwithstanding, the results are satisfactory — the PF tracker copes well with similar distractors and occlusions. A non-rigid target and motion of both the camera and the target also pose no problem.

### B. PF based tracker vs. the MeanShift and the CAMSHIFT

As indicated in Section II-B, color-based image features are robust against the out-of-plane rotations, scale and rotation changes. Additionally, the use of histogram matching results in relatively low computational cost. For this reason, color-based segmentation and tracking methods such as *Mean-Shift* and its extension *CAMSHIFT* (*Continuously Adaptive Mean-Shift*, cf. Fig. 2 upon the differences between these two methods), both within the OpenCV library, gained great popularity over the years.

CAMSHIFT is a parameter-free[1] tracking technique. It re-

[1] In the OpenCV library user can, however, define the maximum number of Mean-Shift iterations to converge

TABLE II
COMPARISON OF THE CAMSHIFT AND THE IMPLEMENTED PF TRACKER

| | **CAMSHIFT** | **PF tracker** |
|---|---|---|
| Mean of ground truth and tracking result overlap | $0.53 \pm 0.06$ | $0.62 \pm 0.10$ |
| Mean of hit ratio[a] | $0.79 \pm 0.02$ | $0.87 \pm 0.04$ |
| Mean number of Mean-Shift iterations per frame | $1.90 \pm 0.52$ | n.a. |
| Mean execution time per frame[b] | $0.73 \pm 0.03$ ms | $5.55 \pm 0.00$ ms |

[a] Hit ratio is the percentage of overlaps greater or equal to $1/3$.

[b] It should be noted that both the PF tracker and the evaluation script for the CAMSHIFT were implemented in Python, which tends to be slow.

quires a probability image of the target object (i.e. the back-projection of the object histogram) and an initial search window. Our test CAMSHIFT-based pedestrian tracker was based on the sample from the OpenCV library source code. It utilizes the HSV histogram of the target object masked with a binary image resulting from the thresholding of the initial image of the target object with different boundaries of colors. The purpose of masking was to extract as much of the target object from the background as possible to avoid false positives. A total number of over 270,000 different masks were used and only first one hundred results in terms of mean overlap were taken into account. The results are presented on Fig. 6a and tabulated in Table II.

As one can note, in several cases the CAMSHIFT-based pedestrian tracker performs marginally better than the implemented PF tracker, although the mean overlap and hit ratio of the former one are clearly lower. The CAMSHIFT-based tracker is also firmly faster. Please bear in mind though, that the result of the CAMSHIFT tracker strongly relies on the probability image of the target object. In this context it was selected carefully so as to correspond strictly to the video sequence used.

Finally, we encourage the reader to watch our short comparison video sequence available at [24].

## V. CONCLUSIONS

In this study we have undertaken the problem of object tracking in video sequences with a special focus on pedestrian tracking. Our objective was to compare the performance of the Mean-Shift and Camshift trackers to the powerful computation technique known as particle filtering. We have explained theoretical background of the PF and shown how we have adapted it for the purpose of pedestrian tracking. From the tests of the compared trackers on example video sequences (shot from a camera in motion) we conclude that the PF can outperform the the Mean-Shift and Camshift trackers. This is, however, at the cost of higher computational demand (cf. Table II summarizing quantitative comparison of the performance of the compared trackers). We argue, however, that the PF can be employed to track multi-modal distributions, i.e. that are not confined to Gaussian distributions. Finally, the strong advantage of the PF is that its computing implementations can be mapped to parallel processing architectures with a flexibly chosen number of particles. Our intention is to employ the developed tracking technique in a multi-camera video system aimed at aiding the visually impaired in mobility and navigation.

## ACKNOWLEDGMENT

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 643636 "Sound of Vision."

## REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, dec 2006. doi: http://dx.doi.org/10.1145/1177352.1177355

[2] J. Kulchandani and K. Dangarwala, "Moving object detection: Review of recent research trends," in *Pervasive Computing (ICPC), 2015 International Conference on*, Jan 2015. doi: 10.1109/PERVASIVE.2015.7087138 pp. 1–5.

[3] Y. Wu, J. Lim, and M. Yang, "Online object tracking: A benchmark," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, June 2013. doi: 10.1109/CVPR.2013.312 pp. 2411–2418. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2013.312

[4] P. Baranski and P. Strumillo, "Enhancing positioning accuracy in urban terrain by fusing data from a GPS receiver, inertial sensors, stereo-camera and digital maps for pedestrian navigation," *Sensors*, vol. 12, no. 6, pp. 6764–6801, 2012. doi: http://dx.doi.org/10.3390/s120606764

[5] M. Bujacz, P. Skulimowski, and P. Strumillo, "Naviton-a prototype mobility aid for auditory presentation of three-dimensional scenes to the visually impaired," *Journal of the Audio Engineering Society*, vol. 60, no. 9, pp. 696–708, 2012.

[6] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1615–1630, Oct 2005. doi: http://dx.doi.org/10.1109/cvpr.2003.1211478

[7] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," *Computer Vision and Pattern Recognition*, vol. 2, no. 1, pp. 142–149, 2000. doi: http://dx.doi.org/10.1109/cvpr.2000.854761

[8] D. Comaniciu and V. Ramesh, "Mean shift and optimal prediction for efficient object tracking," in *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 3, 2000. doi: http://dx.doi.org/10.1109/icip.2000.899297. ISSN 1522-4880 pp. 70–73.

[9] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference, 2002. ISBN 0130851981

[10] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, vol. 2, Oct 2003. doi: http://dx.doi.org/10.1109/iccv.2003.1238422 pp. 734–741.

[11] B. Deori and D. M. Thounaojam, "A survey on moving object tracking in video," *International Journal on Information Theory*, vol. 3, no. 3, July 2014. doi: http://dx.doi.org/10.5121/ijit.2014.3304

[12] A. Materka and M. Strzelecki, "Texture analysis methods–a review," Technical University of Lodz, Institute of Electronics, Brussels, Tech. Rep., 1998, cOST B11 report.

[13] A. Yilmaz, K. Shafique, N. Lobo, X. Li, T. Olson, and M. A. Shah, "Target-tracking in flir imagery using mean-shift and global motion compensation," in *Workshop on Computer Vision Beyond the Visible Spectrum, Kauai*, 2001, pp. 54–58.

[14] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision*, 1998. doi: http://dx.doi.org/10.1109/acv.1998.732882

[15] I. Michael and B. Andrew, "Condensation — conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998.

[16] N. Gordon and D. Salmond, "Bayesian state estimation for tracking and guidance using the bootstrap filter," *Journal of Guidance, Control and Dynamics*, vol. 18, no. 6, pp. 1434–1443, 1995. doi: http://dx.doi.org/10.2514/6.1993-3701

[17] G. Kitagawa, “Monte carlo filter and smoother for non-gaussian nonlinear state space models,” in *Journal of Computational and Graphical Statistics*, vol. 5, no. 1, 1996. doi: http://dx.doi.org/10.2307/1390893 pp. 1–25.

[18] T. Heap and D. Hogg, “Wormholes in shape space: Tracking through discontinuous changes in shape,” in *International Conference on Computer Vision*, 1998. doi: http://dx.doi.org/10.1109/iccv.1998.710741 pp. 344–349.

[19] J. MacCormick and A. Blake, “A probabilistic exclusion principle for tracking multiple objects,” vol. 1, pp. 572–587, 1999. doi: http://dx.doi.org/10.1007/978-1-4471-0679-1_6

[20] K. Nummiaro, E. Koller-Meier, and L. Van Gool, “A color-based particle filter,” in *First International Workshop on Generative-Model-Based Vision*, A. Pece, Ed., vol. 2002/01. Datalogistik Institut, Kobenhavns Universitet, 2002, pp. 53–60.

[21] J. S. Liu and R. Chen, “Sequential monte carlo methods for dynamic systems,” *Journal of the American Statistical Association*, vol. 93, pp. 1032–1044, 1998. doi: http://dx.doi.org/10.2307/2669847

[22] S. Ceranka and M. Niedzwiecki, “Application of particle filtering in navigation system for the blind,” in *Proceedings of the IEEE 17th International Symposium on Signal Processing and its Applications*, 2003. doi: http://dx.doi.org/10.1109/isspa.2003.1224922 pp. 495–498.

[23] D. A. Klein, “Bobot - bonn benchmark on tracking,” 2010. [Online]. Available: http://www.iai.uni-bonn.de/~kleind/tracking/

[24] M. Owczarek, “Pedestrian tracking in video sequences: Camshift-based vs. particle filter-based approach,” 2015, visited on 2015-05-05. [Online]. Available: https://vimeo.com/mateuszowczarek/pf2014b

[25] ——, “Color-based particle filter pedestrian tracker: different number of particles,” 2015, visited on 2015-05-05. [Online]. Available: https://vimeo.com/mateuszowczarek/pf2014a

DOI: 10.15439/2015F166
ACSIS, Vol. 5

# Enhanced movement tracking with Kinect supported by high-precision sensors

Tomasz Pałys
Cybernetics Faculty at the Military University of Technology
ul. S. Kaliskiego 2,
00-908 Warsaw, Poland
Email: tomasz.palys@wat.edu.pl

Witold Żorski
Cybernetics Faculty at the Military University of Technology
ul. S. Kaliskiego 2,
00-908 Warsaw, Poland
Email: wzorski@wat.edu.pl

*Abstract*—**The paper presents a proposal of sensor fusion combining data derived from a Kinect device and high-precision sensors. The main idea was to enhance the tracking of a human arm in order to obtain precise coordinates. The Kinect plays the role of a calibration device and sensors' data are used in kinematics equations for enhanced tracking of the arm. This way the resulting information has less uncertainty and can be directly transmitted to another system, e.g. robotics one. The system has been implemented using kinematics-based approach with transformation operators as well as quaternions. In terms of accuracy the main result is that the Kinect performs very similar with sensors.**

## I. Introduction

THE idea of creating a device that allows to recognize human body shape in details without having to hold any remote controller on it inspired Microsoft to create the Kinect device (2010). It was a very novel and extraordinary approach to the challenge. Soon, the Kinect attracted researchers in many fields, especially from robotics and computer vision, as well as some communities that have released various SDK for Kinect allowing to use it as a kind of measurement tool. Since the lunch of the Kinect for Windows a lot of papers have been devoted to the scientific application of this device.

Some papers may serve as tutorials, e.g. for calibration of the Kinect imaging sensor [12], or help while building a new Kinect-based system [9], or source of references [4] for Kinect-based computer vision researchers, covering topics which include: preprocessing, object tracking and recognition, human activity analysis, hand gesture analysis, and indoor 3-D mapping, and many others.

In the beginning, researchers presented novel approaches to human detection using depth information taken by the Kinect [21], methods of obstacles detection where the Kinect was used as a capturing device [14], methods to quickly and accurately predict 3D positions of body joints from a depth image [17].

Later, considered problems became more complex or sophisticated. The Kinect was used to evaluate noise, accuracy, resolution, and latency of tracking skeletons; and measure the range in which the person being tracked must be in order to achieve these values [7]. Another tested application of the Kinect was a virtual bilateral Man-Machine interaction through the Kinect sensors, allowing the user to control and monitor Windows programs by gestures without a need of peripherals use [1].

There are even attempts to use the Kinect in the field of medicine or clinical rehabilitation as a monitoring device [18]. Another interesting use of the Kinect are methods for calibration using streaming information from depth cameras [19]. This is still a challenge in the case of hands generic precision and agility due to complexity of fingers joints.

As we consider the Kinect, the area of robotics is very popular in some purposes, for instance, enhancing the navigation of mobile robots, object detection and recognition, scene reconstruction, 3D inspection and others. The good example is an automated vehicles inspection by scanning and dynamically exploring regions of interest over an automotive vehicle body under visual guidance [11]. In such applications the very important aspect is quality of sensors [10] responsible for 3D data acquisition [13].

At present, there are many available solutions dedicated to context-aware robotics system [8], [16], e.g. for orientation determination of pedestrian motions by using an inertial measurement unit module and a complementary separate-bias Kalman filter [22].

In this paper we would like to propose a system based on the Kinect assisted by a set of high-precision sensors, that would be responsible for enhance movement tracking of a human arm with the accuracy suitable in the field of robotics. The idea of such a task is presented in Fig. 1, where we can see a man wearing sensors – one on the waistline and three on the natural links of his right arm. The arm’s configuration is mapped by the system. The Fig. 2. gives us a general view about the system and its components.

978-83-60810-66-8/$25.00©2015, IEEE

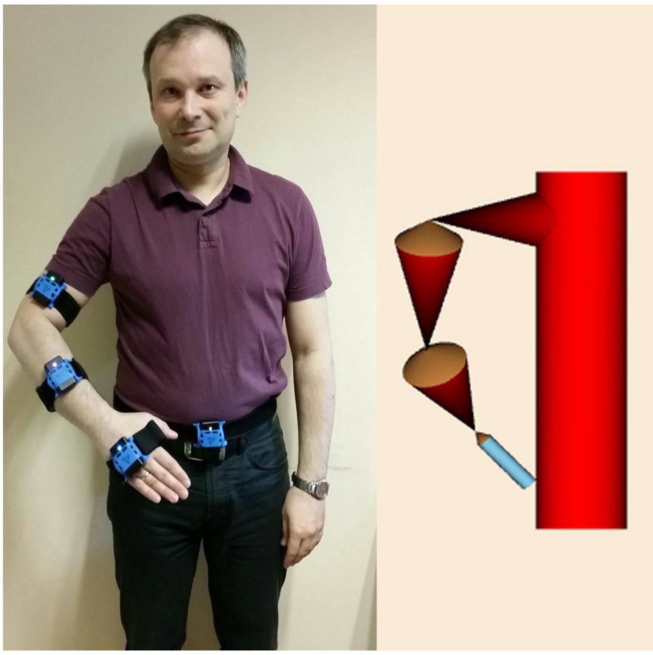


Fig. 1. The idea of mapping a human arm

## II. Presentation of the System

The used computer vision system consists of an x86 computer equipped with the Kinect for Windows v2 device [23], which is supported by the YEI 3-Space sensors [24]. Microsoft Visual Studio 2012 and .NET Framework 4.5 are on the basis of applied software environment. Fig. 2 shows a visual conception of the used computer vision system.

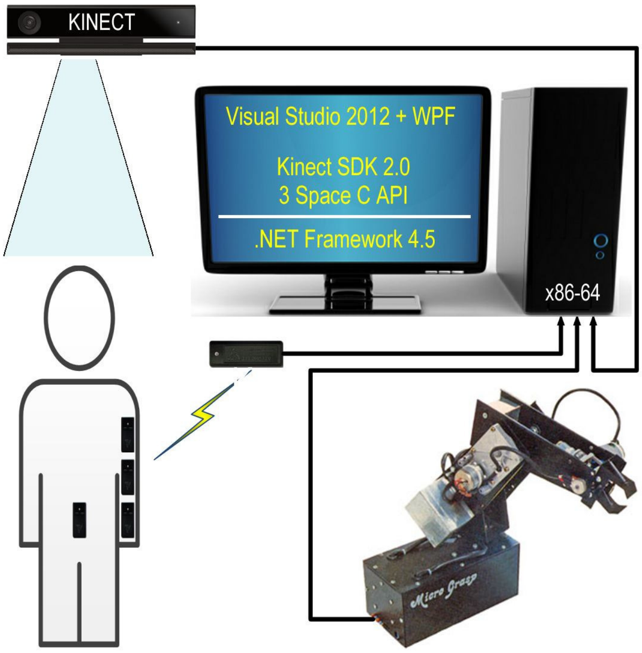


Fig. 2. The system conception

The presented system is based on the second-generation Kinect for Windows (well described in [3]) with an improved motion tracking functionality over its predecessor, a wider field of view, a high definition camera, and the ability to track up to six bodies at once. The Kinect can capture audio, color, and depth environment features, and process the depth data to generate skeleton data.

The YEI 3-Space Sensor (Fig. 3) is a miniature, high-precision, high-reliability, Attitude and Heading Reference System (AHRS) / Inertial Measurement Unit (IMU) with USB 2.0 communication interfaces in a single unit. The AHRS/ IMU uses a gyroscope, an accelerometer, and compass sensors in conjunction with advanced processing and on-board quaternion-based algorithms to determine orientation relative to an absolute reference or relative to a designated reference orientation in real-time. The gradient descent calibration process and high update rates increase accuracy and greatly reduce and compensate errors. The YEI 3-Space Sensor USB unit features are accessible via a well-documented open communication protocol that allows access to all sensor data and configuration parameters. Versatile commands allow access to raw sensor data, normalized sensor data, and filtered absolute and relative orientation outputs in formats incl. quaternions and Euler angles (pitch/roll/yaw). We a going to use both, Euler angles and quaternions, in order to compare results and also to verify our implementation this way.

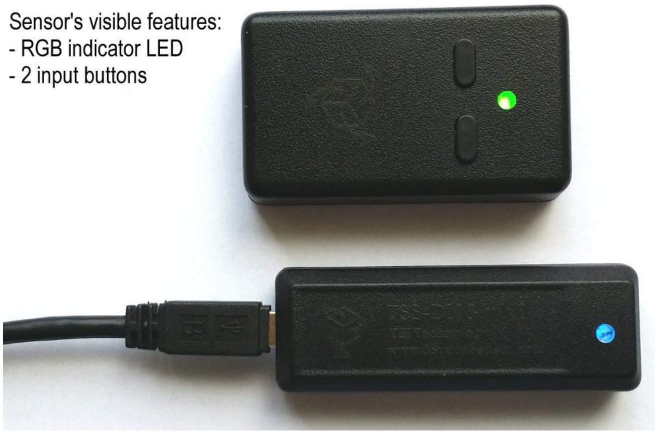


Fig. 3. The YEI 3-Space Sensor and the communication dongle

## III. Software Environment

Considering Visual Studio 2012 as the primal software environment we can distinguish three essential layers of the system: “application layer” on the top, “management layer” in the middle, and “data layer” at the bottom. The structure of the used software is presented in Fig. 4.

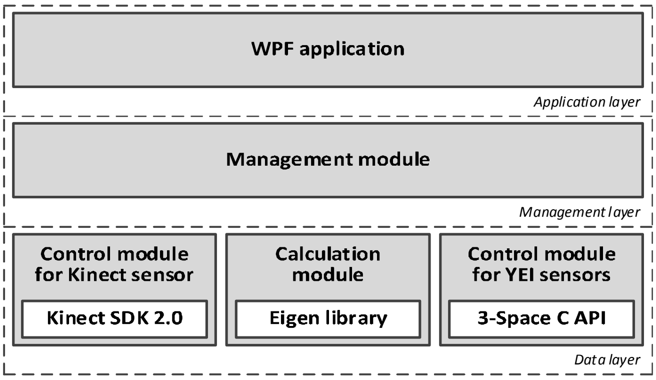


Fig. 4. The structure of the software environment

The application layer stands for the created WPF application [15]. It is the main element responsible for every activity aspect of the system, particularly for the GUI.

The management layer includes management control system tools placed in the management module library.

The set of libraries on the bottom layer includes: a control module for the Kinect sensor (it uses Kinect SDK 2.0), a calculation module (it uses the Eigen library ver. 3.2.2), and a control module for YEI sensors (it uses 3-Space C API ver. 2.0.6).

### A. Microsoft Visual Studio as a development tool

The use of Visual Studio 2012 as a development tool was involved by the possibility to fully utilize the Kinect for Windows SDK and assets of the Windows Presentation Foundation that provides a consistent programming model for building applications with a graphical subsystem. An additional aspect was the integration possibility of Microsoft .NET Common Language Runtime (managed code) with the 3-Space C API library and the Eigen library (unmanaged code).

### B. The Software Development Kit for Kinect

The Kinect for Windows SDK enables to create applications under Visual Studio 2012 (.NET Framwork 4.5 is required) that support advanced gesture and voice recognition using Kinect sensor technology on computers running modern Windows. To develop speech-enabled Kinect applications, the Microsoft Speech Platform SDK must be installed.

### C. The Eigen library

The Eigen is a high-level C++ open source library of template headers for linear algebra, matrix and vector operations, numerical solvers; this facilitates to achieve high performance of required calculations. In particular, the Eigen has also a possibility to solve linear systems using several types of decompositions, and includes several very useful classes and functions for image processing.

### D. The 3-Space API

The use of the 3-Space API was inevitable due to YEI sensors. This API is available for C/C++ and Python languages and can be used to write independent software for YEI sensors. Because of the chosen Visual Studio environment the 3-Space C API has been adopted.

## IV. Kinect's Role

As mentioned earlier and suggested in Fig. 1, the assumption is that the system is able to perform tracking of a human arm with the use of Kinect and sensors. To be exact, the Kinect is use mostly at the stage of calibration and later for verification purposes. In short, to perform the calibration of the system a person wearing four sensors have to take a posture shown in Fig. 5. (the arm must be aligned to ensure proper calibration), and next press a button on sensor $S_1$. At this moment sensors are reset, and the system reads coordinates of points $P_1$ to $P_4$ from the Kinect. Starting from this point, the Kinect plays the role of the secondary sensor, suitable for verification purposes.

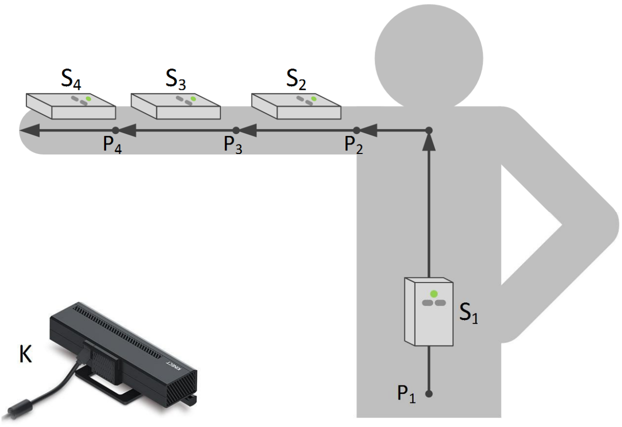


Fig. 5. The calibration of the system

## V. Kinematics Solutions

The use of sensors affixed to a human stature involves a need for an adequate mathematical instrument appropriate to perform required transformations on received data. The obvious one derives from robotics, where the kinematics plays the role of a fundamental description tool [6].

In order to describe the location of each link (a person's body part) relative to its neighbors a frame (a coordinate system) has been affixed to each link; see Fig. 6.

### A. Transformation operators

In the field of robot kinematics transformation operators allow simple and clear matrix representation of rotations and translations [2]. Let us start from coordinates of point $P_1$ (see Fig. 5) received from the Kinect during calibration procedure:

$$ {}^{K}P_1 = [p_1^x, p_1^y, p_1^z, 1]^T . \qquad (1)$$

Next, we can use other coordinates (points P$_2$, P$_3$, P$_4$) received from the Kinect to calculate crucial distances between sensors (see Fig. 6):

$$\begin{cases} x_{21} = p_2^x - p_1^x \\ y_{21} = p_2^y - p_1^y \end{cases}, \quad x_{32} = p_3^x - p_2^x, \quad x_{43} = p_4^x - p_3^x. \qquad (2)$$

Now, we can create transformation operators that are adequate to the situation presented in Fig. 6. The assumption is, that we are able to receive Euler angles $(\alpha,\beta,\gamma)$ from sensors. The next thing is that we use differences of corresponding angles, e.g. $\alpha_3 = \alpha_{S3} - \alpha_{S2}$.

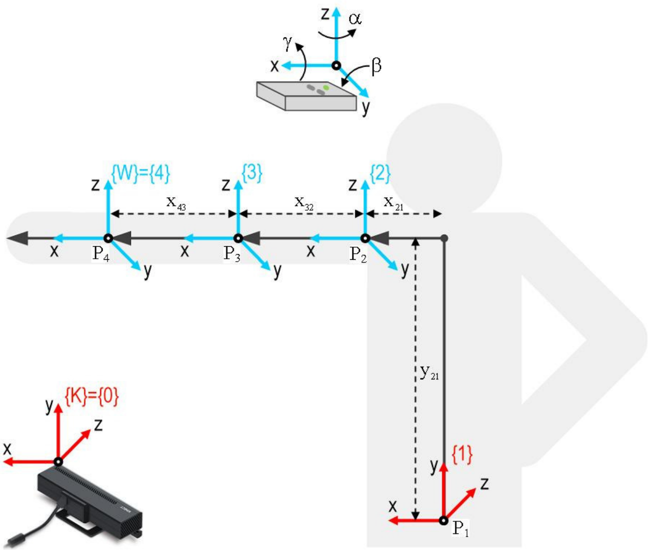


Fig. 6. A model of the human skeleton as a kinematic chain (see Fig. 5)

Taking into account our assumptions as to angles the set of transformation operators is as follows:

$${}^0_1T = {}^K_1T = \begin{bmatrix} 1 & 0 & 0 & p_1^x \\ 0 & 1 & 0 & p_1^y \\ 0 & 0 & 1 & p_1^z \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (3)$$

$${}^1_2T = \begin{bmatrix} 1 & 0 & 0 & x_{21} \\ 0 & \cos(-\pi/2) & -\sin(-\pi/2) & y_{21} \\ 0 & \sin(-\pi/2) & \cos(-\pi/2) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

$$\cdot \begin{bmatrix} c\alpha_2 c\beta_2 & c\alpha_2 s\beta_2 s\gamma_2 - s\alpha_2 c\gamma_2 & c\alpha_2 s\beta_2 c\gamma_2 + s\alpha_2 s\gamma_2 & 0 \\ s\alpha_2 c\beta_2 & s\alpha_2 s\beta_2 s\gamma_2 + c\alpha_2 c\gamma_2 & s\alpha_2 s\beta_2 c\gamma_2 - c\alpha_2 s\gamma_2 & 0 \\ -s\beta_2 & c\beta_2 s\gamma_2 & c\beta_2 c\gamma_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (4)$$

$${}^2_3T = \begin{bmatrix} c\alpha_3 c\beta_3 & c\alpha_3 s\beta_3 s\gamma_3 - s\alpha_3 c\gamma_3 & c\alpha_3 s\beta_3 c\gamma_3 + s\alpha_3 s\gamma_3 & x_{32} \\ s\alpha_3 c\beta_3 & s\alpha_3 s\beta_3 s\gamma_3 + c\alpha_3 c\gamma_3 & s\alpha_3 s\beta_3 c\gamma_3 - c\alpha_3 s\gamma_3 & 0 \\ -s\beta_3 & c\beta_3 s\gamma_3 & c\beta_3 c\gamma_3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (5)$$

$${}^3_4T = \begin{bmatrix} c\alpha_4 c\beta_4 & c\alpha_4 s\beta_4 s\gamma_4 - s\alpha_4 c\gamma_4 & c\alpha_4 s\beta_4 c\gamma_4 + s\alpha_4 s\gamma_4 & x_{43} \\ s\alpha_4 c\beta_4 & s\alpha_4 s\beta_4 s\gamma_4 + c\alpha_4 c\gamma_4 & s\alpha_4 s\beta_4 c\gamma_4 - c\alpha_4 s\gamma_4 & 0 \\ -s\beta_4 & c\beta_4 s\gamma_4 & c\beta_4 c\gamma_4 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (6)$$

Thus, the link transformations can be multiplied together to find the single transformation that relates frame {W} to frame {K}:

$${}^K_W T = {}^0_4 T = {}^0_1 T \, {}^1_2 T \, {}^2_3 T \, {}^3_4 T \qquad (7)$$

Finally, we can calculate coordinates of points from P$_2$ to P$_4$ with respect to frame {K} in the following way:

$$\begin{cases} {}^K P_2 = {}^0_1 T \, {}^1_2 T \cdot [0,0,0,1]^T \\ {}^K P_3 = {}^0_1 T \, {}^1_2 T \, {}^2_3 T \cdot [0,0,0,1]^T \\ {}^K P_4 = {}^K_W T \cdot [0,0,0,1]^T \end{cases} \qquad (8)$$

In the case of a robotics system we will prefer coordinates of point P$_4$ (i.e. for the wrist) with respect to frame {1} that can be calculated in the following way:

$${}^1 P_4 = {}^1_2 T \, {}^2_3 T \, {}^3_4 T \cdot [0,0,0,1]^T \qquad (9)$$

This is the case when the arm is fully independent of the Kinect as long as the calibration is done.

### *B. The quaternions*

The quaternions [5], a number system that extends the complex numbers, are members of a noncommutative division algebra. Dual-quaternions may formulate and solve a problem more concisely, more rapidly and in fewer steps, with result more plainly to others, and practice with fewer lines of code effortlessly debugged. There is no loss of efficiency – dual-quaternions can be just as efficient (if not more efficient) than using matrix methods. Additionally, there are some very important reasons for using dual-quaternions: singularity-free, un-ambiguous, shortest path interpolation, etc. This mathematical tool becomes very popular in the computer world, especially in computer graphics – even the used Kinect and sensors are quaternion-based devices. In a general case dual-number quaternions [20] allow compact representation of both rotations and translations. This mathematical tool was integrated into the current system with little coding effort.

In short, a dual-quaternion consists of two quaternions: $q_r$, $q_d$ (eight elements). These two quaternions are called the real part and the dual part:

$$q = q_r + q_d \varepsilon \,, \qquad (10)$$

where $\varepsilon$ is an additional dual number.

The dual-quaternion can represent only rotation by angle $\varphi$ if the dual part is set to zero:

$$q_R = [\cos(\frac{\varphi}{2}) - (\hat{v}_x i + \hat{v}_y j + \hat{v}_z k)\sin(\frac{\varphi}{2})][0,0,0,0]\,, \qquad (11)$$

where $\hat{v}$ is a unit axis.

Analogously, to represent only translation by vector $[t_x, t_y, t_z]$, the real part is set to an identity and the dual part represents the translation:

$$q_T = [1,0,0,0][0,\frac{t_x}{2},\frac{t_y}{2},\frac{t_z}{2}]\,. \qquad (12)$$

Now, combining the rotational and translational transforms into a single unit quaternion to represent a rotation followed by a translation we get:

$$q = q_T \times q_R \,. \qquad (13)$$

The following equation defines how we can transform point $p$ into point $p'$, using the received unit dual-quaternion:

$$p' = q \cdot p \cdot q^* \,, \qquad (14)$$

where $q$ and $q^*$ represent a dual-quaternion transform and its conjugate.

In the considered system we can directly use quaternions received from the Kinect and sensors. However, if we have only Euler angles $(\alpha, \beta, \gamma)$, there is an easy way to receive the quaternion responsible for the rotation (see Eq. 11):

$$\begin{cases} q_x = \cos(\gamma/2) - i\sin(\gamma/2) \\ q_y = \cos(\beta/2) - j\sin(\beta/2) \\ q_z = \cos(\alpha/2) - k\sin(\alpha/2) \end{cases}$$

$$q_R = [q_z \cdot q_y \cdot q_x][0,0,0,0]\,. \qquad (15)$$

## VI. Results

Results we obtained seem to be quite satisfying. Both tools, transformation operators and quaternions, were independently implemented and gave exactly the same values, what may be considered as a confirmation of a proper solution (implementation). Additionally, the coordinates received from the Kinect are concurrent with those calculated by the system – this has been checked in many cases; a sample is presented in Fig. 7–9, respectively for coordinates X, Y, and Z. The considered figures present wrist coordinates (the values are in metres) received during a person activity for an example period of 8 seconds.

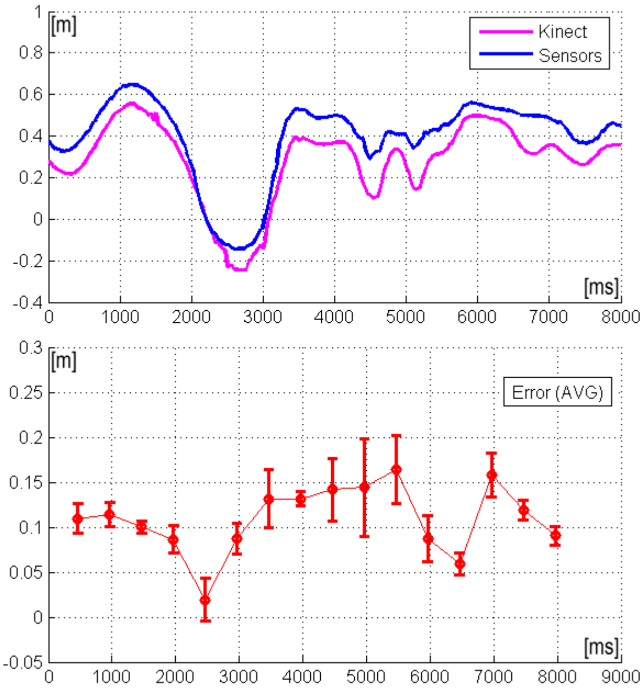


Fig. 7. Wrist coordinate X – values and the average error

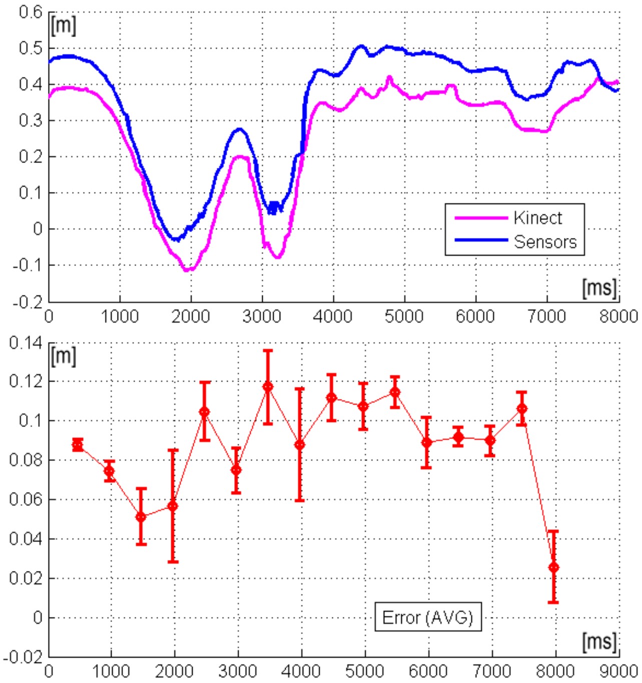


Fig. 8. Wrist coordinate Y – values and the average error

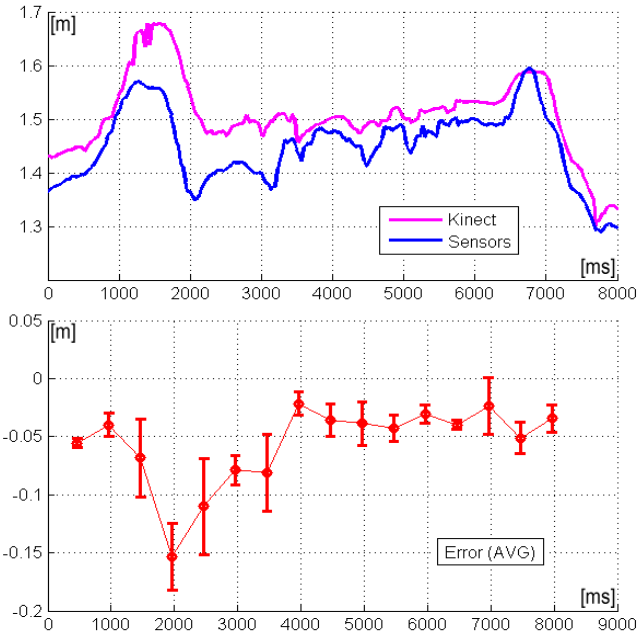


Fig. 9. Wrist coordinate Z – values and the average error

## VII. Conclusion

Authors are aware that the idea of the Kinect device supported by additional sensors is a strong negation of the Kinect nature. Nevertheless, we need to accent that the Kinect served mainly as a fine calibration device and for some verification purposes. Additionally, the Kinect fails for some "configurations" of a human posture, environment objects are serious obstacles, and the space range of the device is very narrowed. This set of adversities has been reduced by the use of the set of four high-precision sensors.

Some gains of the work are worth to be mentioned as advantages of the system: a high level of accuracy limited mainly by the process of calibration, immunity to terrain obstacles, and wide operating range.

We plan to improve the system accuracy, further simplify and revamp the process of calibration, and remove some technical limitations. We want to use only quaternions, even in the case of data received from the Kinect and sensors as they are native for these devices. We would like to perform some robotics experiments in order to face unknown.

## References

[1] Andaluz V., Gallardo C., Santana J., Villacres J., *Bilateral Virtual Control Human-Machine with Kinect Sensor*, Andean Region International Conference, 2012, pp. 101-104. http://dx.doi.org/10.1109/Andescon.2012.32

[2] Craig J. J., "*Introduction to Robotics: mechanics and control*", 2nd ed., Addison-Wesley Publishing Company, 1989. http://dx.doi.org/10.1016/0005-1098(87)90105-1

[3] Fankhauser P., et al., *Kinect v2 for Mobile Robot Navigation: Evaluation and Modeling*, IEEE International Conference on Advanced Robotics, 2015 (the paper is accepted).

[4] Jungong H., Ling S, Dong Xu, Shotton J., *Enhanced Computer Vision with Microsoft Kinect Sensor: A Review*, IEEE Transactions on Cybernetics, Vol.43(5), 2013, pp. 1318-1334. http://dx.doi.org/10.1109/TCYB.2013.2265378

[5] Kenwright B., *A Beginners Guide to Dual-Quaternions: What They Are, How They Work, and How to Use Them for 3D*, WSCG 2012 Communication Proceedings, pp.1-13.

[6] Kim Jung-Ha, Kumar V. R., *Kinematics of robot manipulators via line transformations*, Journal of Robotic Systems, Vol.7(4), 1990, pp. 649–674. http://dx.doi.org/10.1002/rob.4620070408

[7] Livingston M., Sebastian J., Zhuming Ai, Decker J., *Performance measurements for the Microsoft Kinect skeleton*, Virtual Reality Short Papers and Posters, 2012, pp. 119-120. http://dx.doi.org/10.1109/VR.2012.6180911

[8] Luo R.C., Bo-Han Shih, Tsung-Wei Lin, *Real time human motion imitation of anthropomorphic dual arm robot based on Cartesian impedance control*, IEEE International Symposium on Robotic and Sensors Environments, 2013, pp. 25-30. http://dx.doi.org/10.1109/ROSE.2013.6698413

[9] Ming A., Enomoto K., Shinozaki M., Sato R., Shimojo M., *Development of an entertainment robot system using Kinect*, 8th Europe-Asia Congress on Mechatronics, 2014, pp. 127-132. http://dx.doi.org/10.1109/MECATRONICS.2014.7018621

[10] Murawski K., *Measurement of Membrane Displacement with a Motionless Camera Equipped with a Fixed Focus Lens*, Metrology and Measurement Systems, Vol.22(1), 2015, pp. 69-78. http://dx.doi.org/10.1515/mms-2015-0011

[11] Nakhaeinia D., Fareh R., Payeur P., Laganiere R., *Trajectory planning for surface following with a manipulator under RGB-D visual guidance*, IEEE International Symposium on SSRR 2013, pp. 1-6. http://dx.doi.org/10.1109/SSRR.2013.6719365

[12] Pagliari D., Menna F., Roncella R., Remondino F., Pinto L., *Kinect Fusion improvement using depth camera calibration*, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-5, 2014, pp. 479-485. http://dx.doi.org/10.5194/isprsarchives-XL-5-479-2014

[13] Pinto A., Costa P., Moreira A., Rocha L., *Evaluation of Depth Sensors for Robotic Applications*, IEEE International Conference on Autonomous Robot Systems and Competitions, 2015, pp. 139-143. http://dx.doi.org/10.1109/ICARSC.2015.24

[14] Rakprayoon P., *Kinect-based obstacle detection for manipulator*, IEEE/SICE International Symposium on System Integration (SII), 2011, pp. 68-73. http://dx.doi.org/ 10.1109/SII.2011.6147421

[15] Ren Yu, Gubing Lu, Feng Lu, *A Method of Rotation Transformation for 3D Object by Changing Camera Attributes in WPF*, International Conference on Information Engineering and Computer, 2010, pp. 1-4. http://dx.doi.org/10.1109/ICIECS.2010.5678267

[16] Rosado J., Silva F., Santos V., Zhenli Lu, *Reproduction of human arm movements using Kinect-based motion capture data*, IEEE International Conference on Robotics and Biomimetics, 2013, pp. 885-890. http://dx.doi.org/10.1109/ROBIO.2013.6739574

[17] Shotton J., Fitzgibbon A., Cook M., Sharp T., *Real-time human pose recognition in parts from single depth images*, IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1297-1304. http://dx.doi.org/10.1109/CVPR.2011.5995316

[18] Tao G., Archambault P. S., Levin M. F., *Evaluation of Kinect skeletal tracking in a virtual reality rehabilitation system for upper limb hemiparesis*, International Conference on Virtual Rehabilitation, 2013, pp. 164-165. http://dx.doi.org/10.1109/ICVR.2013.6662084

[19] Vicente A., Faisal A., *Calibration of kinematic body sensor networks: Kinect-based gauging of data gloves "in the wild"*, IEEE International Conference on Body Sensor Networks, 2013, pp. 1-6. http://dx.doi.org/10.1109/BSN.2013.6575526

[20] Walker M. W., Shao L., Volz R. A., *Estimating 3-D location parameters using dual number quaternions*, CVGIP: Image Understanding Vol.54(3), 1991, pp. 358-367. http://dx.doi.org/10.1016/1049-9660(91)90036-O

[21] Xia Lu, Chen Chia-Chih, Aggarwal J. K., *Human detection using depth information by Kinect*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011, pp. 15-22. http://dx.doi.org/10.1109/CVPRW.2011.5981811

[22] Zhang R., Reindl L., *Pedestrian motion based inertial sensor fusion by a modified complementary separate-bias Kalman filter*, Sensors Applications Symposium, 2011, pp. 209-213. http://dx.doi.org/10.1109/SAS.2011.5739766

[23] https://www.microsoft.com/en-us/kinectforwindows/

[24] http://www.yeitechnology.com/yei-3-space-sensor

DOI: 10.15439/2015F240
ACSIS, Vol. 5

# Analysis of the sound attack in context of computer evaluation of the singing voice quality

Edward Półrolniczak
West Pomeranian University of Technology
Faculty of Computer Science and Information Technology
ul. Żołnierska 52, 71-210 Szczecin, Poland
Email: epolrolniczak@wi.zut.edu.pl

Michał Kramarczyk
West Pomeranian University of Technology
Faculty of Computer Science and Information Technology
ul. Żołnierska 52, 71-210 Szczecin, Poland
Email: mkramarczyk@wi.zut.edu.pl

***Abstract*—Sound attack is meant as the initiation of the tone. Good attack involves the simultaneous start of exhalation along with the emission of sound. This situation is known as a soft attack. Much more common is a hard type of attack, which is normally used while speaking. It involves tightening of the vocal cords before the expiration. Unfortunately, this state is tiring for the ligaments and can cause damage to the voice. The idea was to propose a method to analyse this problem, which in the future would become a part of set of methods for computer analysis of the singing voice. This article presents a method of the extraction of attack parameters. The estimation of these parameters is carried out with the recorded audio samples. In the study sound sample of 'a' registered several times for each pitch will be used.**

***Index Terms*—singing voice, sound attack, ADSR, quality of singing**

## I. Introduction

The motivation for taking up the research on the attack in the singing is the need of assessment of singing quality. The problem is to find some criteria of evaluation of the sound attack while singing. The goal is to find the criteria of automatic evaluation and involve computer methods to evaluate the attack in singing using these criteria. If properly defined, they allow for self-correction of selected voice parameters. It may be useful for training lessons of voice production. It can be useful to help singers make a progress. It can be very important for the choirs constantly working on the voice. Some criteria of evaluation, if properly defined, may allow for self-correction of selected voice parameters. The considered methods should evaluate these features in a similar way as human experts. To achieve that it is necessary to propose a computer method to analyse this feature and its parameters.

It should be noted here that it is difficult to find scientific publications which present the approach to computer evaluation of singing carried out in similar way as described in this article. The approach presented here reflects expert's evaluation in the computer aided assessment.

The singing voice is produced by the vocal instrument consisting of three basic components: the respiratory apparatus, the oscillating vocal folds and the vocal tract. Breathing has decisive influence on all activities related to the voice emission. Vocal folds open and close during breathing, but during speaking and singing, they start to oscillate with the fundamental frequency of sound which comes out of the mouth. At the interface between these two phases (open and close) of the sound generation the musical phenomenon known as an attack on the sound occurs.

The ADSR envelope is a practical model that describes a single sound (note). It can be used for sound analysis [1], [2] and synthesis [3]. It is the essential description of the sound waveform in the MIDI standard [4]. In terms of sound synthesis attack is interpreted as a part of ADSR envelope. The attack is used for modification of a first phase of amplitude envelope [5] of generated sound in which sound gains the highest amplitude. In most basic models it's followed by decay section, during which amplitude is decreased. After that there is a stage of a sustain which is characterized by a stable amplitude. The ADSR envelope is finished then by a stage of release. During this part amplitude is completely decreasing . The ADSR sections are illustrated in fig. 1.

The word "attack" is probably too strong description of the release of air through glottal folds. Anyway it is the term commonly used in the literature. Some other, related terms used in literature are: initiation, onset [6] or transient. To be precise it is needed to define some terms related to each other: onset, attack, transient. The reason for making these distinctions clear is that different applications have different needs. The similarities and differences between these key concepts should be considered. Referring to the definitions found at [7] the attack is the time interval during which the amplitude envelope of the sound increases.

The transients are short intervals during which the signal evolves quickly (increases and decreases quickly) in some nontrivial, unpredictable way. It is often connected to the situation where the excitation of the sound is applied and then damped (leaving only a slow decay). So the transient would be a part of the amplitude envelope starting from the onset, including the whole attack and the most important part of decay (the strongest one). The onset [3], [8] is a single instant chosen to mark a transient. In most cases it is connected with the start of transient (also start of attack) or with the earliest time when the transient can be detected.

The sound attack in the singing is related to the breathing. During vocal development care should be taken to teach the attack on the breath [9]. Proper breath preceding the phonation

978-83-60810-66-8/$25.00©2015, IEEE

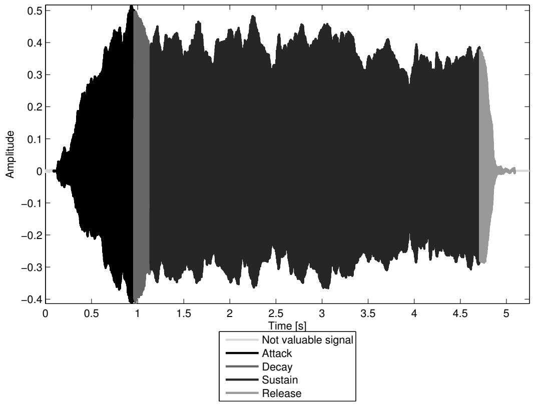


Figure 1. Example of the typical split of a sung phrase divided into sections ADSR (Attack-Decay-Sustain-Release)

phase will result in a good beginning of each phrase. It is especially important in phrases which begin with a vowel sound. Attack is more precise in the case of professional singers. Choral singers are less precise in that stage of voice production. It has to be noticed that practice can clear up that problem. Choir members are usually developing their voice in groups. Thus the vocal abilities are similar in a group of choir singers. It should be recalled here that the presented investigation concerns choir voices.

There are three known kinds of attack on sound depending on moment of tightening vocal cords in relation to the beginning of exhalation and on how strong the tightening of vocal cords is in the early stages of sound production:

- soft - when the vocal cords tightening moment coincides with the beginning of exhalation. This is the most favourable attack.
- hard - tightening of the vocal cords begins before the expiration. Unfortunately, this state is tiring for the ligaments and can cause damage to the voice.
- exhaling (aspirated) - short exhaust ahead of tightening of the vocal cords. Vocal cords do not close completely and the remaining gap influences the sound.

In this paper attack part of a signal is defined as first part of the sung sound. It ends by either stabilisation of amplitude (sustain) or total decrease of amplitude (decay).

In the article [10] an attempt to identify fast attacks has been presented. Fast attack transients are named there simply as attacks. The authors have defined the attacks as zones of short duration (a few ms) and fast variation of the signal short time spectrum with an abrupt increase in energy particularly noticeable in high frequencies since energy is usually concentrated in the low frequencies. The attack detection and modelling method developed there was based on the following requirements: the method should not use additive analysis results, in order to be usable for other purposes (segmentation, instrument recognition, etc.), it should succeed in every type of sound (particularly polyphonic sounds) with good time accuracy, it should be simple to use: the analysis parameters should, as much as possible, be adjusted automatically, it should be tested on a data base of sounds including polyphonic mixtures of percussive and non-percussive sounds.

## II. Research Conditions

The database consisting of representative samples presenting the abilities of choir singers is important point of that study. The database used here is the extension of the database created under the research project of West Pomeranian University of Technology: "Computerized methods of supporting the process of training choir voices" [11]. The recorded singers are singing in the choir of the same university. The samples in the database are divided into categories reflecting different aspects of the singer's practices. The content of the database allows to extract the selected parameters of singing voice. The exercises were selected from a set usually used during the voice production training. It is possible to investigate for example intonation [12], vibrato feature, tremolo, sonority, noise [13] and other features. It is possible to perform some more general investigations over the database as, for example, singing voice quality assessment [14], [15]. For that study the exercise containing the vowel "a" sung at one pitch for a few seconds was chosen. In fact, the most interesting part here is situated at the beginning of voiced part of the sample.

The recordings used in this article were taken in the specially arranged room, with proper conditions for the recording session. All of the sound material was recorded with a 24 bit resolution, with the sampling frequency of 48 kHz. Higher recording parameters give some wider possibilities of editing of the recordings. The sessions were recorded as a whole for each singer. The division into elementary parts, called sequences or phrases (piano, singer), was done with use of a program with a tool for automatic segmentation. The process was performed under supervision of an expert to provide the possible best prepared samples.

Eight males, representing baritone voice, were subjects for the experiment. They were chosen as representative voices from the group of singers. Each person was in good vocal and health condition. The samples were recorded twice to have the possibility to compare the results. The sound pitches selected for the study were chosen to be comfortable to sing for all of the singers. The pitches covered the range from pitch number 10 in octave 2 (sound A) to pitch number 5 in octave 4 (sound e2). There were 2 sets of the samples for each singer. The sets consisted of 5 samples, so together there were 80 samples available to analyse (8 singers, 2 sets, 5 samples for each set).

## III. Research Idea

The idea of the research was to develop a descriptors of attack on sound. Those descriptors should map experts' evaluation of analysed signal. For this reason, the evaluations of attack on sound were carried out parallelly by three experts. Evaluation marks given by the experts will be presented in the following part of this publication. Finally both results, obtained by computer methods and given by human experts, will be compared to decide whether the similar conclusion can be drawn.

### A. Experts' Assessment

Three experts were asked to assess the sound attack in the samples of singing. They were instructed about the definition of the attack on sound defined in this study. They had a possibility to listen to the voice examples before they started the evaluation. It was done to teach the experts the rules of the assessment and to make sure the samples would be evaluated in the same way by all of the experts. The experts had to assess the attack feature using 1-3 scale, where 1 meant so called soft attack, 2 meant medium attack and 3 meant hard attack. The experts were also evaluating the quality of the attack on the sound. To assess the quality they used a rating scale 1-5 where 1 was bad quality and 5 meant very good quality. It should be reminded here that every expert had to evaluate 2 sets of samples for each singer. Every expert was assessing the set of the samples according to the procedure:

- at first they were able to listen to 6 different, randomly selected, samples,
- 1st set of samples was played as first and then it was evaluated by the expert,
- 2nd set of samples was played in the random order.

The experts did not see the ratings of each other. During the assessment process three experts have assessed 8 male singers. There were 2 sets for each singer. Each set consisted of 5 samples. Concluding, samples were assessed 240 times. Taking into account there were 2 marks for each sample (the type of the attack and the quality of the attack) it gave 480 marks.

### B. Computer Aided Evaluation

During the preliminary studies, it has been established that the attack on the sound will be analysed according to the definition described in the first section of the article. At the beginning of the study it was assumed that the factors will be applied to three areas of the voice signal: time, amplitude and pitch [2]. Parameters of time have been given in seconds. The generalized Hilbert envelope was used to describe the characteristics of the amplitude. To illustrate each envelope the following parameters, which are also described in fig. 2, were selected:

- angle of attack on volume - directional factor of linear function of approximation for the amplitude envelope (fig. 2b),
- mean square error between the linear approximation and the actual envelope (fig. 2b),
- ratio of the energy of attack in relation to the energy of the entire signal represented by equation 1,
- values of derived envelope calculated by eq. 2 - the total sum of values and the number of values above and below zero (fig. 2c).

According to the theory, if both the first and the second parameter has a low value, we should be dealing with perfectly soft attack. Other parameters were selected as complementary ones, in order to test the possibility of their use.

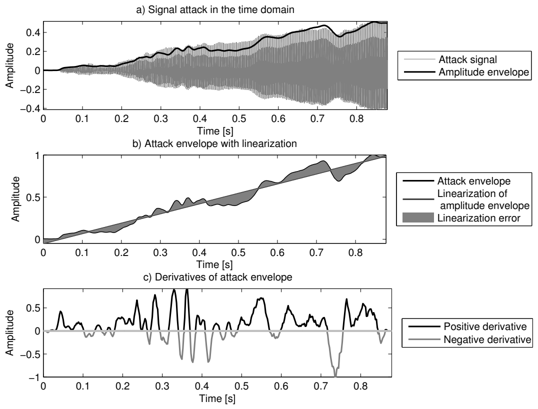


Figure 2. Presentation of some attack parameters based on amplitude values

$$E = \frac{\sum(Attack^2)}{\sum(Signal^2)},$$

where $E$-is energy parameter,

$Attack$-fragment of signal during the attack phase,

$Signal$-whole signal.

(1)

$$f'(x) = \begin{cases} 0 & : x = 1 \\ f(x) - f(x-1) & : x > 1 \end{cases}, x \in \mathbb{N}_+,$$

where $f'(x)$-is derivative in point x,

$f(x)$-is base signal

$x$-is index of element in vector.

(2)

The third set of analysed parameters concerned the pitch. This set meant to show how the singer sets the pitch during the attack on the sound. The parameters are displayed in fig. 3, and include following sets of data:

- angle of attack on pitch - directional factor of linear function of approximation for the pitch track (fig. 3b),
- mean square error between the linear approximation and the actual pitch track (fig. 3b),
- maximum pitch value registered with the upper pitch limit given by the value of expected tone plus 200 Hz,
- values of derived pitch track calculated with eq. 2 - the total sum of values and the number of values above and below zero (fig. 3c).

If the first parameter is less than zero that means the singer is attacking the sound from a higher frequency and trying to reduce the frequency to match the sound with the referenced tone. In the opposite case the singer increases the frequency in order to match the tone. Other parameters were selected as complementary ones, to test the possibility of their use.

## IV. The Results

As it was mentioned before, the investigation consisted of two parts. In the first part experts were asked to assess

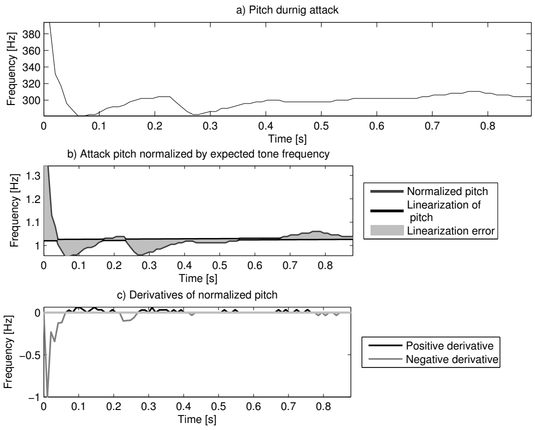


Figure 3. Presentation of some attack parameters based on pitch values

the samples from the database. In the second part a set of calculations over the samples was performed. After that, correlations between the results were searched.

### A. Computer Aided Evaluation Results

As it was planned before, the special calculations over the chosen samples were performed to obtain information about the attack on the sound. One of the most natural parameters of the attack calculated on the sound sample was the angle of attack calculated over the volume envelope. The parameter is interpreted here as a directional factor of a linear function which is an approximation of the amplitude of the envelope (fig. 2b). The values obtained for 5 samples for each singer are presented together in fig. 4. It can be seen in the graph that the angle of the attack is higher in case of the boundary samples. It is because these pitches are too high or too low to be enough comfortable for singing. It can be observed that in the tested group there are singers having problems at every sample and the singers having some discomfort only at boundaries. It is apparent from the graph that the best are singers s21 and s27, as the angle is generally lower for those singers.

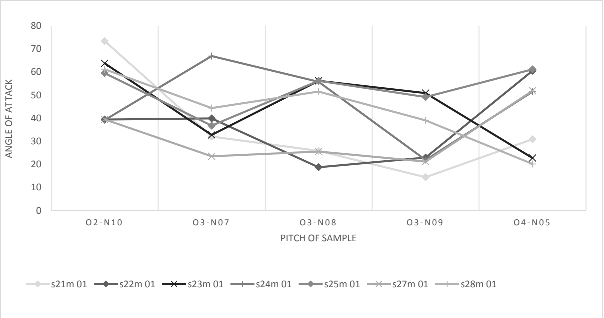


Figure 4. Angle of amplitude attack parameter (version 1 of set of samples)

The results obtained for mean square error between the linear approximation of the sound attack and the actual envelope (fig. 2b) are presented in fig. 5. Analysing that graph, and comparing to the previous one it should be noted that the parameter for singers s21 and s27 differentiate the singers. The singer s21 has higher values of these parameters than s27 and for this reason it can be stated that s27 is better, more stable, than s21 (although, considering the second attempt, s21 became more stable). Combining those two parameters together it is possible to obtain a decision rule better describing the abilities of the singers.

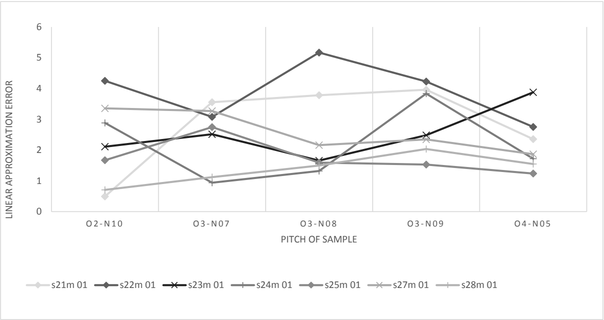


Figure 5. Mean square error parameter for the actual attack envelope (version 1 of set of samples)

Another example of the calculated parameter is ratio of the energy of attack in relation to the energy of the entire signal (represented by equation 1). The results visible in fig. 6 again highlight the singers s21 and s27. Those singers have greater energy values for the attack than the others. Especially if the energy values for each of the samples are summed up the differences between those singers and the others become really clear. The other results will not be described here in details, as the ones presented above sufficiently illustrate the general idea.

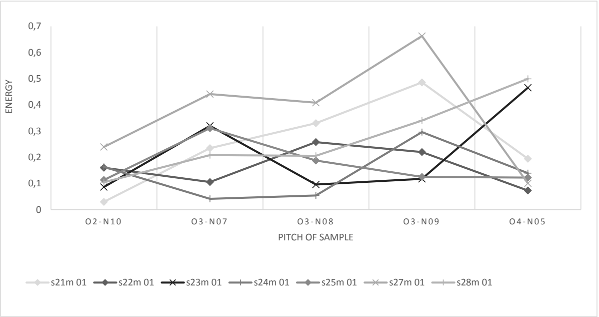


Figure 6. Energy parameter calculated over attack envelopes (version 1 of set of samples)

### B. Experts Results

Experts were asked to assess the feature of sound attack of the sung samples. They were instructed about the rules adopted in this study. They had a possibility to listen to the voice examples before evaluation of the process. The procedure was set to guarantee the assessment of the samples in the same way. The experts evaluated, among others, whether the sound attack in the recorded sample is soft, medium or hard. Those marks have been counted for each singer. That gave an entry information about the approach of the singer to the sound initiation. To illustrate that the figures 7 and 8 characterising the results of a good singer are presented. He was, generally, evaluated in the same way by all the experts. Some problems were observed on both ends of the analysed range. Additionally, from figure 8 one can see that the second trial (every singer was recorded twice) was better than the first one. Each expert has generally evaluated the initiation of the sound as soft, in this case. It should be noticed that all the samples lying in the middle range of the pitch were evaluated as soft. This is a result consistent with expectations of the experts.

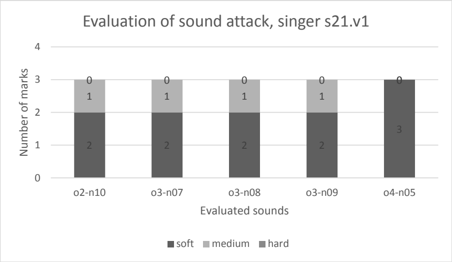


Figure 7. Evaluation of the singer s21 (sample version 1)

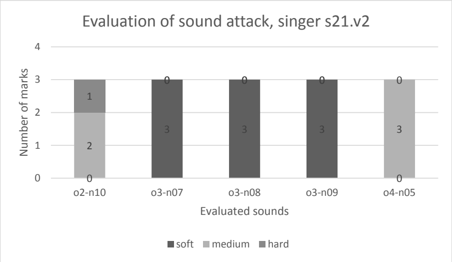


Figure 8. Evaluation of the singer s21 (sample version 2)

Another situation is presented in figure 9. The singer's results presented in that figure are not stable. This may be connected with some evaluation problems or, on the other hand, with low abilities of the singer.

All evaluation results were collected in one table and compared with the results obtained in an automatic way.

### C. Correlation of The Results

The results, both given by the experts and obtained from computer calculations were compared to find an answer to the question - whether the computer methods proposed here are able to evaluate singing samples in the similar way as human experts do. To this end the relation between the vector containing experts' decisions and the vector containing values of each previously calculated parameters for each singer was estimated. The example provided here shows comparison of the expert marks given to each singer with the calculated angle parameters.

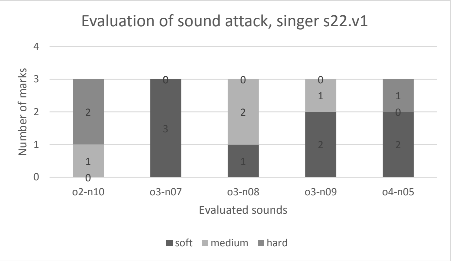


Figure 9. Evaluation of the singer s22 (sample version 1)

The upper part of the table I ("Summary of weighted expert's evaluation marks") summarize the marks given by the experts. The marks given by three experts were summed up and weighed to obtain general mark for each singer. For each mark (obtained in the previous step) it has been assigned an adjective describing the type of the attack according to the rule: mark 1-3 meant hard attack, 4-6 meant medium, 7-9 meant soft attack. The numerical marks were also averaged to give the average adjective mark. Similar operations were carried out on the second part of the table consisting of the values calculated using previously described procedure. The bottom part of the table contains the comparison of the marks, both, calculated and given by the experts. The row "Consistent values" shows how many values are consistent in both sets (expert's marks and calculated values). Taking into account all 5 samples (for each singer) it gives the mean value of consistency at the level of about 60 percent. Taking into account only the three middle samples the average consistency level between experts and computer evaluation also gives the value of about 60 percent. It leads to the conclusion that to achieve better results other parameters should be also involved.

The approach presented here is based on the expert's way of assessing the singing. The parameters proposed and calculated here are directly connected to the singing signal and can be easily observed in the signal. In the present article an example application of one of the parameters has been described. Therefore, further improvement of efficiency of computer assessment using the remaining, previously mentioned, parameters seems to be possible.

## V. Conclusion

The article focuses in general on the problem of the sound attack in singing. Sound attack is understood here as an initiation of the sound. This characteristic is very subjective in perception, but it can be used as a criterion of singing quality evaluation. The attack on sound can be divided into several types. One of the types is a soft attack. This is identified with the beginning of the sound in a soft, fine, smooth manner. Much more common is a hard type of attack, which is normally used while speaking. Unfortunately, this state is tiring for a singer and can cause damage of the voice. That is why the hard attack is not advised while singing for a long time. The goal was to involve computer methods to evaluate

Table I
THE ANGLE OF THE ATTACK

Summary of weighed experts evaluation marks - 1-3: hard, 4-6: medium, 7-9: soft

| s22m 01 | mark | s23m 01 | mark | s24m 01 | mark | s25m 01 | mark | s27m 01 | mark | s28m 01 | mark | s21m 01 | mark |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | medium | 5 | medium | 6 | medium | 4 | medium | 5 | medium | 5 | medium | 8 | soft |
| 9 | soft | 9 | soft | 7 | soft | 3 | hard | 9 | soft | 5 | medium | 8 | soft |
| 7 | soft | 7 | soft | 8 | soft | 3 | hard | 9 | soft | 3 | hard | 8 | soft |
| 8 | soft | 7 | soft | 9 | soft | 4 | medium | 8 | soft | 3 | hard | 8 | soft |
| 7 | soft | 7 | soft | 8 | soft | 6 | soft | 8 | soft | 6 | medium | 9 | soft |
| Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark |
| 7 | soft | 7 | soft | 7,6 | soft | 4 | medium | 7,8 | soft | 4,4 | medium | 8,2 | soft |

Calculated values of the angle parameter

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | medium | 4 | medium | 6 | medium | 4 | medium | 6 | medium | 4 | medium | 3 | hard |
| 6 | medium | 7 | soft | 3 | hard | 6 | medium | 8 | soft | 6 | medium | 7 | soft |
| 8 | soft | 4 | medium | 4 | hard | 4 | medium | 7 | soft | 5 | medium | 7 | soft |
| 8 | soft | 5 | medium | 8 | soft | 5 | medium | 8 | soft | 6 | medium | 9 | soft |
| 4 | medium | 8 | soft | 5 | medium | 4 | medium | 5 | medium | 8 | soft | 7 | soft |
| Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark | Mean value | Mean mark |
| 6,4 | soft | 5,6 | medium | 5,2 | medium | 4,6 | medium | 6,8 | soft | 5,8 | medium | 6,6 | medium |

Consistent values (%)

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 60 | | 60 | | 40 | | 40 | | 80 | | 40 | | 80 |

Consistent values excluding boundary pitches (%)

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 66 | | 33 | | 33 | | 33 | | 100 | | 33 | | 100 |

the attack in a singing to help the singer make progress. The methods should evaluate that feature in a similar way as human experts. To achieve that it was necessary to propose a method to analyse this problem. The results of the above automatic computer evaluation were compared to the marks given by the human experts. For the purpose of automatic evaluation it was assumed that the attack on the sound can be represented by objective parameters of signal describing subjective human impression. There was a need to answer the question whether among the calculated parameters were such that reflect experts' evaluation and thus can be useful to construct computer aided assessment system. The evaluation of the sound attack would be a part of that system. The investigation consisted of two parallel parts. In one of them the experts were assessing recordings of singing. In the other part a set of signal parameters was calculated for each singer in the context of sound attack. Among others the following parameters were calculated: the angle of attack calculated over the volume envelope, mean square error between the linear approximation of the sound attack and the actual envelope, the energy of attack in relation to the energy of the entire signal. During the study it has been found that among estimated features angle parameter, mean square error parameter and energy parameter can construct feature vector applied in a computer method of a sound attack quality assessment. The results concerning sound attack presented here may be useful for constructing a singing quality assessment system.

A large number of the results obtained for this study requires further, deeper analysis and may lead to subsequent applications.

## ACKNOWLEDGMENT

The authors would like to thank the members of The Jan Szyrocki Memorial Choir of West Pomeranian University of Technology for their cooperation in building database of singing voices. The authors would also like to thank the experts who spent a lot of time listening to and evaluating the singing samples.

## REFERENCES

[1] Y. Meron and K. Hirose, "Separation of singing and piano sounds." in *ICSLP*, 1998.

[2] M. Muller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 6, pp. 1088–1110, 2011.

[3] A. Holzapfel, Y. Stylianou, A. C. Gedik, and B. Bozkurt, "Three dimensions of pitched instrument onset detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1517–1527, 2010.

[4] L. Mazurowski, "Computer models for algorithmic music composition," in *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*. IEEE, 2012, pp. 733–737.

[5] K. Jensen, "Envelope model of isolated musical sounds," in *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, 1999.

[6] J. Davids and S. LaTour, *Vocal Technique: A Guide for Conductors, Teachers, and Singers*. Waveland Press, 2012.

[7] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 1035–1047, 2005.

[8] C.-C. Toh, B. Zhang, and Y. Wang, "Multiple-feature fusion based onset detection for solo singing voice." in *ISMIR*, 2008, pp. 515–520.

[9] R. M. Alderson, *Complete handbook of voice training*. Parker Publishing Company, 1979.

[10] X. Rodet and F. Jaillet, "Detection and modeling of fast attack transients," in *Proceedings of the International Computer Music Conference*, 2001, pp. 30–33.

[11] M. Łazoryszczak and E. Półrolniczak, "Audio database for the as sessment of singing voice quality of choir members," *Elektronika: konstrukcje, technologie, zastosowania*, vol. 54, no. 3, pp. 92–96, 2013.

[12] E. Półrolniczak and M. Łazoryszczak, "Quality assessment of intonation of choir singers using f0 and trend lines for singing sequence," *Metody Informatyki Stosowanej*, pp. 259–268, 2011.

[13] E. PÓŁROLNICZAK and M. KRAMARCZYK, "Computer analysis of the noise component in the singing voice for assessing the quality of singing," *Przegląd Elektrotechniczny*, vol. 91, pp. 79–83, 2015.

[14] E. Polrolniczak and M. Kramarczyk, "Formant analysis in assessment of the quality of choral singers," in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2013*, Sept 2013, pp. 200–204.

[15] P. Zwan and B. Kostek, "System for automatic singing voice recognition," *Journal of the Audio Engineering Society*, vol. 56, no. 9, pp. 710–723, 2008.

DOI: 10.15439/2015F10
ACSIS, Vol. 5

# Robotic Arm Detection in Space with Image Recognition Made in Linux with the Hough Circles Method

Roland Szabó, Aurel Gontean
Politehnica University Timişoara,
Faculty of Electronics and
Telecommunications, Applied
Electronics Department, V. Pârvan
Av., no. 2, 300223, Timişoara,
România
Email: roland.szabo@etc.upt.ro,
aurel.gontean@upt.ro

***Abstract*—This paper presents a method to recognize a robotic arm in space using the Hough circles method. The robotic arm has colored bottle stoppers glued at the joints, which are recognized with color filtering. After this step, the biggest colored spot is detected and marked using the Hough circle method. The joints are numbered, this way each joint's position is known. The joints can be united with lines and this way a skeleton of the robotic arm can be drawn which can be loaded in a PC to create a control application.**

## I. Introduction

THIS paper presents a robotic arm detection and control using the Hough circles algorithm.

The control of a robotic arm is very useful in the industry, because in today's consumer world only robotic arms can do all the tasks precisely and with high speed [1]. The robotic arms need to be automated as much as possible, because this way, they can do even more tasks with less human intervention, which will lead to higher productivity [2].

Mostly in the industry, robotic arms are blind; they don't have vision system [3]. Vision system adds artificial intelligence and more autonomy to a robotic arm [4]. Mostly in the industry robotic arms are programmed previously, they know exactly all the movements which need to be executed [14]. This way, the robotic arms are not very flexible for today's dynamic world. If the robotic arm has a vision system added, than it can change its tasks during execution, this way there is no need to stop a production line [15]. A vision system can also make auto calibration for the robotic arm, this way the robotic arm can make small adjustments during execution, this way there is no need to stop the production line and this will save money and time [16].

In this experiment a robotic arm detection and control method was made. The robotic arm has glued colored bottle stoppers, the image has applied color filters and the joints are detected with the Hough circles method [8].

The joints are numbered, so exact position of them it's known. The joints can be united with straight lines, this way the skeleton of the robotic arm can be created and this way the movement of the robotic arm can be loaded in a PC [9]. With this information a control algorithm can be created just based on the images of the robotic arm taken with the connected cameras [10].

## II. Problem Formulation

The laboratory from the university had two Lynxmotion AL5 type robotic arms, the AL5A and the AL5B. These robots had no cameras or some other control system connected to them. The idea was to create a smart control system with image recognition, where the position in space of the robotic arm can be detected and to have a possibility to control the robotic arm using this information [11]. The intention was to implement the control system in C/C++ with the OpenCV library and run it on the Linux operating system. The used cameras were the Logitech C270 cameras which are compatible with a wide range of operating systems, including Ubuntu Linux 12.04 LTS which was used in this experiment. The Linux operating system is also a very good choice, because the system can be ported, with only a small amount of effort on other embedded devices which can run the Linux operating system.

## III. Problem Implementation

### *A. Used Equipment*

The experiment was made with the Lynxmotion AL5B robotic arm (Fig. 1). This robotic arm is used mainly for teaching purposes, but it can be also used for tasks which reflect industrial behavior of a robotic arm, because it copies exactly the characteristics of an industrial robotic arm.

The used robotic arm is controlled with the SSC-32 servo control board (Fig. 2). The communication with the servo control board can be done from a master control system, like a PC, via the RS-232 serial interface. The SSC-32 servo control board can generate PWM signals up to 32 motors; the used robotic arm has only 6 motors, so the servo control board is more than enough. The movement angle of each motor from the robotic arm is proportional with the PWM signal's duty cycle [12].

978-83-60810-66-8/$25.00©2015, IEEE

The motors from the robotic arm are special Hitec servomotors with gears for increasing the torque; they have also a special circuit board which moves the motors proportionally to the PWM signal generated by the SSC-32 servo control board [13].

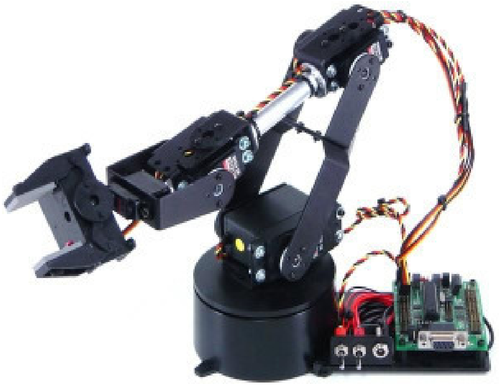


Fig. 1 Lynmotion AL5B robotic arm

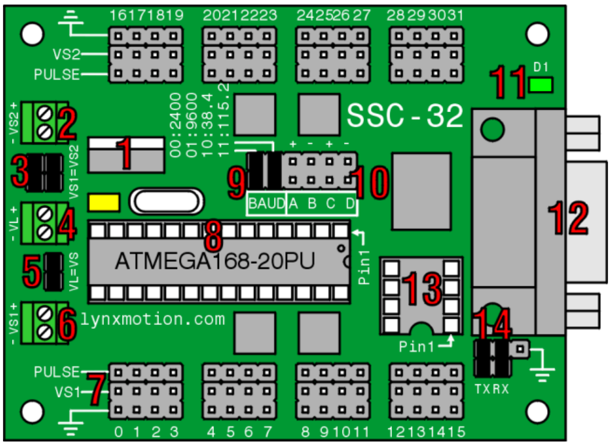
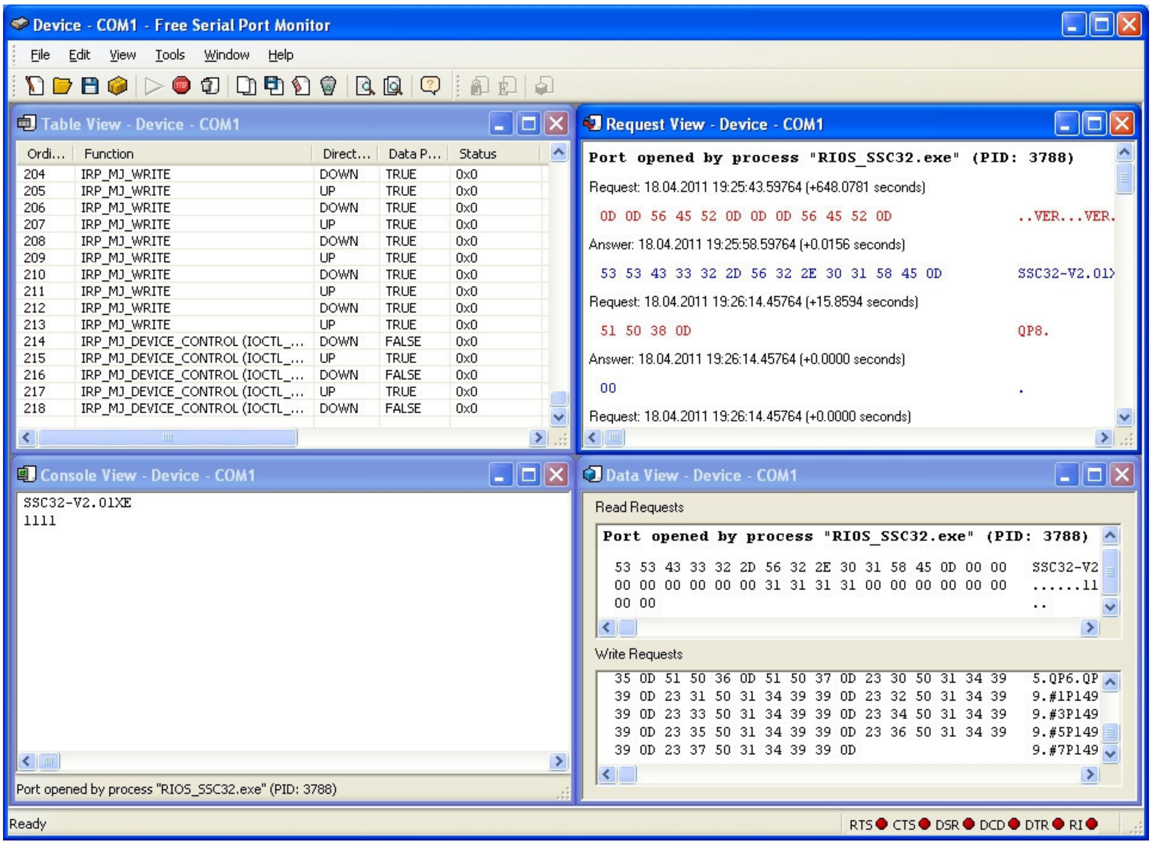


Fig. 2 SSC-32 servo control board

### *B. Theoretical Background*

For controlling the robotic arm, SCPI (Standard Commands for Programmable Instruments) commands are sent through the serial interface from the PC to the SSC-32 servo control board. These commands were captured using the HHD Free Serial Port monitor program (Fig. 3).


Device - COM1 - Free Serial Port Monitor
File
Edit
View
Tools
Window
Help
Table View - Device - COM1
Request View - Device - COM1
Console View - Device - COM1
Data View - Device - COM1
Read Requests
Write Requests
Ready
RTS ● CTS ● DSR ● DCD ● DTR ● RI ●

| Ordi... | Function | Direct... | Data P... | Status |
|---|---|---|---|---|
| 204 | IRP_MJ_WRITE | DOWN | TRUE | 0x0 |
| 205 | IRP_MJ_WRITE | UP | TRUE | 0x0 |
| 206 | IRP_MJ_WRITE | DOWN | TRUE | 0x0 |
| 207 | IRP_MJ_WRITE | UP | TRUE | 0x0 |
| 208 | IRP_MJ_WRITE | DOWN | TRUE | 0x0 |
| 209 | IRP_MJ_WRITE | UP | TRUE | 0x0 |
| 210 | IRP_MJ_WRITE | DOWN | TRUE | 0x0 |
| 211 | IRP_MJ_WRITE | UP | TRUE | 0x0 |
| 212 | IRP_MJ_WRITE | DOWN | TRUE | 0x0 |
| 213 | IRP_MJ_WRITE | UP | TRUE | 0x0 |
| 214 | IRP_MJ_DEVICE_CONTROL (IOCTL_... | DOWN | FALSE | 0x0 |
| 215 | IRP_MJ_DEVICE_CONTROL (IOCTL_... | UP | TRUE | 0x0 |
| 216 | IRP_MJ_DEVICE_CONTROL (IOCTL_... | DOWN | FALSE | 0x0 |
| 217 | IRP_MJ_DEVICE_CONTROL (IOCTL_... | UP | TRUE | 0x0 |
| 218 | IRP_MJ_DEVICE_CONTROL (IOCTL_... | DOWN | FALSE | 0x0 |

```
Port opened by process "RIOS_SSC32.exe" (PID: 3788)
Request: 18.04.2011 19:25:43.59764 (+648.0781 seconds)
 0D 0D 56 45 52 0D 0D 0D 56 45 52 0D          ..VER...VER.
Answer: 18.04.2011 19:25:58.59764 (+0.0156 seconds)
 53 53 43 33 32 2D 56 32 2E 30 31 58 45 0D    SSC32-V2.01>
Request: 18.04.2011 19:26:14.45764 (+15.8594 seconds)
 51 50 38 0D                                  QP8.
Answer: 18.04.2011 19:26:14.45764 (+0.0000 seconds)
 00                                           .
Request: 18.04.2011 19:26:14.45764 (+0.0000 seconds)
```

```
SSC32-V2.01XE
1111
```

```
Port opened by process "RIOS_SSC32.exe" (PID: 3788)
```

```
Port opened by process "RIOS_SSC32.exe" (PID: 3788)
 53 53 43 33 32 2D 56 32 2E 30 31 58 45 0D 00 00   SSC32-V2
 00 00 00 00 00 00 31 31 31 31 00 00 00 00 00 00   ......11
 00 00                                             ..
```

```
 35 0D 51 50 36 0D 51 50 37 0D 23 30 50 31 34 39   5.QP6.QP
 39 0D 23 31 50 31 34 39 39 0D 23 32 50 31 34 39   9.#1P149
 39 0D 23 33 50 31 34 39 39 0D 23 34 50 31 34 39   9.#3P149
 39 0D 23 35 50 31 34 39 39 0D 23 36 50 31 34 39   9.#5P149
 39 0D 23 37 50 31 34 39 39 0D                     9.#7P149
```

Fig. 3 HHD Free Serial Port Monitor

Next it will be presented the SCPI commands for controlling the robotic arm.

The first command is the version reading command, which is used for establishing connection over serial interface between the master control computer and the SSC-32 servo control board.

The commands after are the commands which initialize the motors. Actually these commands test all the digital ports from the SSC-32 servo control board. There are tested all 32 digital ports, even those which have no motors connected to them. This is a self test before controlling the robotic arm; this way is known if the system runs normally.

The “ALL SERVOS 1500” command places all servos in the position 1500. The position 1500 is the middle position of each motor. The motors are mechanically blocked to be able to make a maximum rotation of 180°, so the middle position is at 90°. The robotic arm is assembled mechanically in such a way, when all motors are in middle position, it can be said that the robotic arm woke up. The robotic arm it’s in a position which resembles a Greek capital gamma letter (“Γ”). The “ALL SERVOS 1500” command places the robotic arm in the home position, a position which is well known by the system, because all the motors are at the position 1500.

The last commands are used to control each motor of the robotic arm. The first number after the “#” it’s the motor number. The second number after the “P” letter it’s the motor position. If it’s 1500, than the motor it’s in the middle position. The last number which is after the “S” letter is the execution time, which is usually set to 1000 ms. If this number is 0, than this means that the motor is set to the maximum possible speed, so this way the execution of the movement will be made at the maximum speed which can be executed by the motor.

```
  // SSC-32 VERSION
  \r\rVER\r

  // INITIALIZE MOTORS
  QPL0\rQP0\r
  QP1\r
  QP2\r
  //...
  QP31\r

  // ALL SERVOS 1500 (MIDDLE POSITION)
  #0P1500S0\r#1P1500S0\r#2P1500S0\r#3P1500S0\
r#4P1500S0\r#5P1500S0\r
```

```
  // GRIPPER
  #4P1500S1000\r

  // WRIST ROTATE
  #5P1500S1000\r

  // WRIST
  #3P1500S1000\r

  // ELBOW
  #2P1500S1000\r

  // SHOULDER
  #1P1500S1000\r

  // BASE
  #0P1500S1000\r
```

If the minimum robotic value (500) and maximum robotic value (2500) is known, than equation (1) can be obtained.

$$\alpha = \frac{\Delta\omega}{180° - 0°} = \frac{2500 - 500}{180° - 0°} = \\ = 11,(1)\ robotic\ values \qquad (1)$$

Equation (1) implies equation (2).

$$1°\sim 11,(1)\ robotic\ values \qquad (2)$$

In Fig. 4 there is the block diagram of the experiment. As it can be seen, we have a PC, the Logitech C270 webcam and the Lynxmotion AL5B robotic arm. The system is made for 2D control of the robotic arm, but for 3D control, the camera and the algorithm needs to be duplicated. The software part is a little bit more interesting. First it’ s used a HSV (Hue, Saturation Value) filter. After this, the Hough circle detection it’s made.

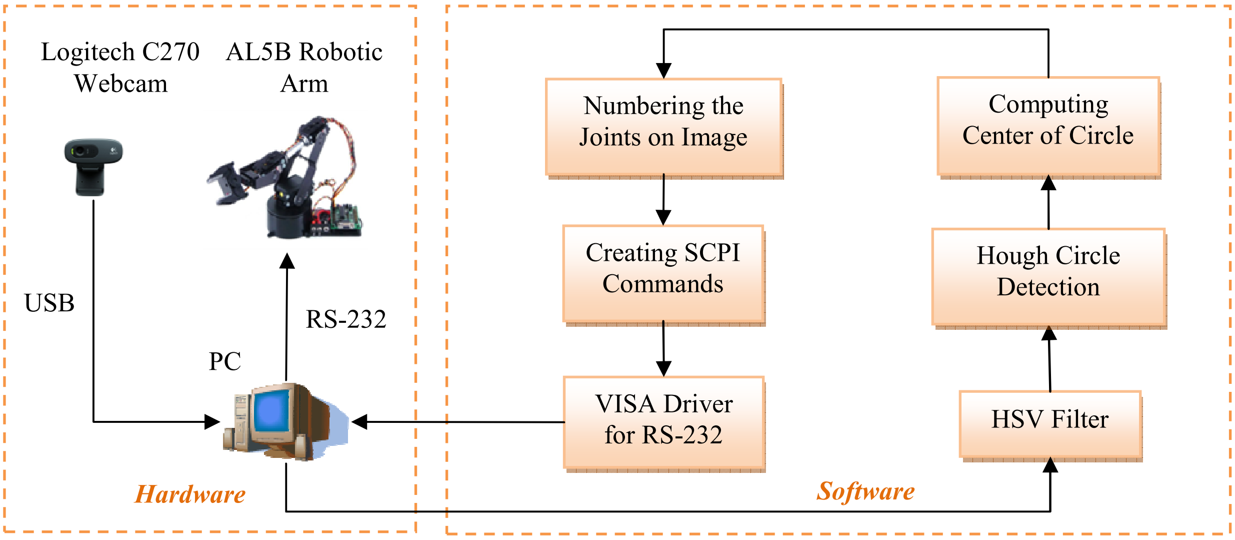


Fig. 4 Block diagram of the experiment

The Hough circle algorithm can detect round spots in an image. Round spots are present on the image after color filtering. The biggest round spot will be chosen, which will be a colored bottle stopper on a joint of the robotic arm. The Hough circle algorithm draws a circle around this detected colored spot.

After this step, the center of the circle is detected, which is actually the coordinate of a joint of the robotic arm and also is the place where the number of the joint will be placed. After this step, the joints can be united and this way the position of the robotic arm in space is known. After knowing the position of the robotic arm, SCPI commands are created and sent using the VISA RS-232 serial driver from National Instruments.

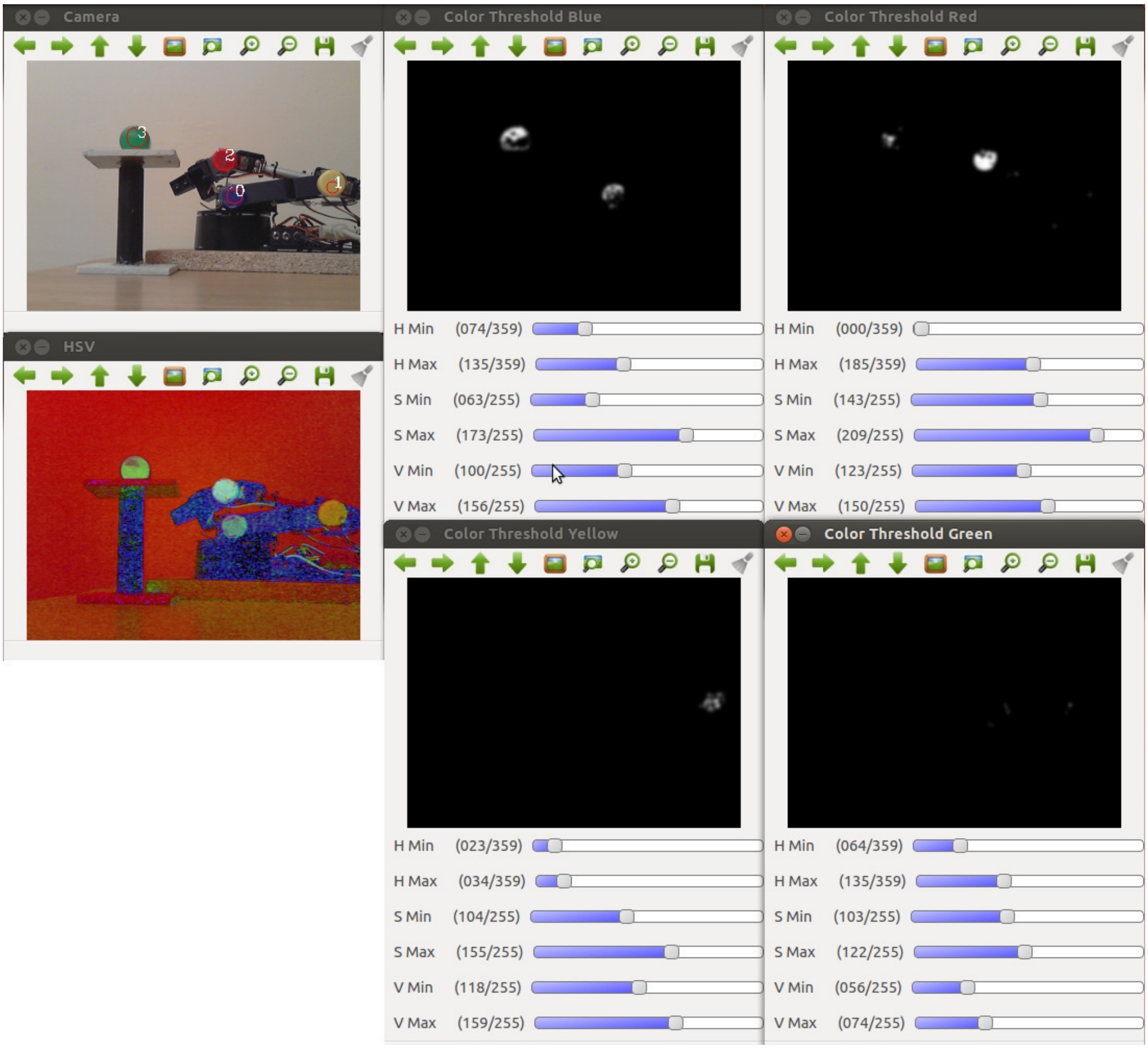
### C. Software Implementation

The software was implemented in Ubuntu Linux 12.04 LTS operating system using the C/C++ programming language with the OpenCV library and the Qt GUI. As it can be seen on the first image, the joints are recognized with the circles around the colored bottle stoppers glued to them. The joints are numbered.

The second image is the HSV filter of the original image.

The next four images are the color filters for blue, yellow, red and green with the sliders for the HSV values.

As we can see the joints were recognized, the only task is to unite the joints with straight lines, which is quite easy because the coordinates are known. They are exactly at the center of the circles. The numbers are drawn at the centers of the circles. The only task is to write the equation of a segment which goes though two points in space. After the step when the joints are united, the skeleton of the robotic arm is created. This data can be loaded in a PC and the position of the robotic arm will be known. Based on this info SCPI commands can be sent to the robotic arm to place its gripper in the desired position.



Color Threshold Blue

H Min (074/359)
H Max (135/359)
S Min (063/255)
S Max (173/255)
V Min (100/255)
V Max (156/255)

Color Threshold Red

H Min (000/359)
H Max (185/359)
S Min (143/255)
S Max (209/255)
V Min (123/255)
V Max (150/255)

Color Threshold Yellow

H Min (023/359)
H Max (034/359)
S Min (104/255)
S Max (155/255)
V Min (118/255)
V Max (159/255)

Color Threshold Green

H Min (064/359)
H Max (135/359)
S Min (103/255)
S Max (122/255)
V Min (056/255)
V Max (074/255)

Fig. 5 Robotic arm control application under Ubuntu Linux 12.04 LTS in C/C++ programming language with OpenCV library using the Hough circles method for joint position detection

## IV. Conclusion

It was presented a system which can recognize the joints of a robotic arm using the Hough circles algorithm.

The robotic arm recognition is useful in the industry because it can ease the maintenance of the production line. There is no need to program previously or to calibrate the robotic arm, these operations are done on-the-go during execution.

The Hough circles algorithm works flawlessly for joints detection. There was also made a test with the Hough circle algorithm to follow a bouncing ping-pong ball. The test was done with no errors. After this test it could be told that the Hough circles algorithm can follow any movement of a robotic arm, it can even follow a bouncing ping-pong ball.

The Hough circles algorithm is implemented in OpenCV, so it can be used with ease.

Further enhancements would be to port the robotic arm control application on an embedded system.

The first step would be to port it on a Raspberry PI with Raspbian Linux "Wheezy" v7.0. After this it can be ported on FPGA boards which can run Ubuntu Liunx 12.04 LTS like the ZYBO or the ZedBoard development boards from Digilent.

On all the three systems the code can be ported directly as it is written in C/C++ or in Python. All the boards have the possibility to install OpenCV, so all the algorithms can be implemented with just minor changes.

The final task is to actually create the ASIC with Linux running on it and the code written in C/C++ with the OpenCV library. This would recognize the robotic arm's joints using the Hough circles algorithm. The task to create an ASIC is not the most complicated one, because there are tools from Mentor Graphics which can convert the VHDL codes in Verilog and after it can create the layout of the chip. This ASIC will behave similar to a PC or to a development board, but it will be a compact embedded IC which will be portable and will have the possibility to be placed in on PCBs with reduced size.

This idea to do everything portable and embedded came due to the tendency of the industry and consumer electronics predictions. This fact is described also in the German factories in the "Industrie 4.0" standard which explains that in the future the PC will not exist and it's no more needed. The task is to have many small smart devices like smart phones, tablets, smart watches and embedded devices which can do the same task as a PC, but they are much more flexible and much more compact. These devices are all connected in a big wireless network, all the devices communicate with each other in a smart manner and everything is more automated and works better, faster with less human intervention. This big wireless network became possible with the introduction of IP v6 which gave more IP addresses to the continuously growing devices connected to the big internet and to the intranets.

## References

[1] W. G. Hao, Y. Y. Leck, L. C. Hun, "6-DOF PC-Based Robotic Arm (PC-ROBOARM) with efficient trajectory planning and speed control," *4th International Conference On Mechatronics*, Kuala Lumpur, 2011, pp. 1–7, http://dx.doi.org/10.1109/ICOM.2011.5937171.

[2] W. Yang, J. H. Bae, Y. Oh, N. Y. Chong, B. J. You, S. R. Oh, "CPG based self-adapting multi-DOF robotic arm control," *International Conference on Intelligent Robots and Systems*, Taipei, 2010, pp. 4236–4243, http://dx.doi.org/10.1109/IROS.2010.5651377.

[3] E. Oyama, T. Maeda, J. Q. Gan, E. M. Rosales, K. F. MacDorman, S. Tachi, A. Agah, "Inverse kinematics learning for robotic arms with fewer degrees of freedom by modular neural network systems," *International Conference on Intelligent Robots and Systems*, 2005, pp. 1791–1798, http://dx.doi.org/10.1109/IROS.2005.1545084.

[4] N. Ahuja, U. S. Banerjee, V. A. Darbhe, T. N. Mapara, A. D. Matkar, R.K. Nirmal, S. Balagopalan, "Computer controlled robotic arm," *16th IEEE Symposium on Computer-Based Medical Systems*, New York, 2003, pp. 361–366, http://dx.doi.org/10.1109/CBMS.2003.1212815.

[5] M. H. Liyanage, N. Krouglicof, R. Gosine, "Design and control of a high performance SCARA type robotic arm with rotary hydraulic actuators," *Canadian Conference on Electrical and Computer Engineering*, St. John's, CA, 2009, pp. 827–832, http://dx.doi.org/10.1109/CCECE.2009.5090244.

[6] M. Mariappan, T. Ganesan, M. Iftikhar, V. Ramu, B. Khoo, "A design methodology of a flexible robotic arm vision system for OTOROB," *International Conference on Mechanical and Electrical Technology*, Singapore, 2010, pp. 161–164, http://dx.doi.org/10.1109/ICMET.2010.5598341.

[7] H. Guo-Shing, C. Xi-Sheng, C. Chung-Liang, "Development of dual robotic arm system based on binocular vision," *International Automatic Control Conference*, Nantou, 2013, pp. 97–102, http://dx.doi.org/10.1109/CACS.2013.6734114

[8] R. Szabó, A. Gontean, "Controlling a Robotic Arm in the 3D Space with Stereo Vision," *21th Telecommunications Forum*, Belgrade, 2013, pp. 916–919, http://dx.doi.org/10.1109/TELFOR.2013.6716380.

[9] R. Szabó, A. Gontean, "Robotic arm control in 3D space using stereo distance calculation," *International Conference on Development and Application Systems*, Suceava, 2014, pp. 50–56, http://dx.doi.org/10.1109/DAAS.2014.6842426.

[10] R. Szabó, A. Gontean, "Remotely Commanding the Lynxmotion AL5 Type Robotic Arms," *21th Telecommunications Forum*, Belgrade, 2013, pp. 889–892, http://dx.doi.org/10.1109/TELFOR.2013.6716373.

[11] R. Szabó, A. Gontean, "Creating a Programming Language for the AL5 Type Robotic Arms," *36th International Conference on Telecommunications and Signal Processing*, Rome, 2013, pp. 62–65. [Online]. Available: http://dx.doi.org/10.1109/TSP.2013.6613892.

[12] R. Szabó, A. Gontean, "Full 3D Robotic Arm Control with Stereo Cameras Made in LabVIEW," *Federated Conference on Computer Science and Information Systems*, Kraków, 2013, pp. 37–42.

[13] R. Szabo, A. Gontean, "Robotic Arm Control with Stereo Vision Made in LabWindows/CVI," *37th International Conference on Telecommunications and Signal Processing*, Berlin, 2014, pp. 635–639.

[14] M. Seelinger, E. Gonzalez-Galvan, M. Robinson, S. Skaar, "Towards a robotic plasma spraying operation using vision," *IEEE Robotics & Automation Magazine*, vol. 5, issue 4, 1998, pp. 33–38, 49, http://dx.doi.org/10.1109/100.740463.

[15] R. Kelly, R. Carelli, O. Nasisi, B. Kuchen, F. Reyes, "Stable visual servoing of camera-in-hand robotic systems," *IEEE/ASME Transactions on Mechatronics*, vol. 5, issue 1, 2000, pp. 39–48, http://dx.doi.org/10.1109/3516.828588.

[16] V. Lippiello, F. Ruggiero, B. Siciliano, L. Villani, "Visual Grasp Planning for Unknown Objects Using a Multifingered Robotic Hand", *IEEE/ASME Transactions on Mechatronics*, vol. 18, issue 3, 2013, pp. 1050–1059, http://dx.doi.org/10.1109/TMECH.2012.2195500.

[17] M. Kazemi, K. K. Gupta, M. Mehrandezh, "Randomized Kinodynamic Planning for Robust Visual Servoing", *IEEE Transactions on Robotics*, vol. 29, issue 5, 2013, pp. 1197–1211, http://dx.doi.org/10.1109/TRO.2013.2264865.

[18] R. T. Fomena, O. Tahri, F. Chaumette, "Distance-Based and Orientation-Based Visual Servoing From Three Points", *IEEE*

*Transactions on Robotics*, vol. 27, issue 2, 2011, pp. 256–267, http://dx.doi.org/10.1109/TRO.2011.2104431.

[19] N. C. Orger, T. B. Karyot, "A symmetrical robotic arm design approach with stereo-vision ability for CubeSats," *6th International Conference on Recent Advances in Space Technologies*, Istanbul, 2013, pp. 961–965, http://dx.doi.org/10.1109/RAST.2013.6581353.

[20] F. Medina, B. Nono, H. Banda, A. Rosales, "Classification of Solid Objects with Defined Shapes Using Stereoscopic Vision and a Robotic Arm," *Andean Region International Conference*, Cuenca, 2012, pp. 226, http://dx.doi.org/10.1109/Andescon.2012.71.

[21] M. Puheim, M. Bundzel, L. Madarasz, "Forward control of robotic arm using the information from stereo-vision tracking system," *14th International Symposium on Computational Intelligence and Informatics*, Budapest, 2013, pp. 57–62, http://dx.doi.org/10.1109/CINTI.2013.6705259.

[22] T. P. Cabre, M. T. Cairol, D. F. Calafell, M. T. Ribes, J. P. Roca, "Project-Based Learning Example: Controlling an Educational Robotic Arm With Computer Vision," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 8, issue 3, 2013, pp. 135–142, http://dx.doi.org/10.1109/RITA.2013.2273114.

[23] G. S. Gupta, S. C. Mukhopadhyay, M. Finnie, "WiFi-based control of a robotic arm with remote vision," *Instrumentation and Measurement Technology Conference*, Singapore, 2009, pp. 557–562, http://dx.doi.org/10.1109/IMTC.2009.5168512.

[24] L. Haoting, W. Wei, G. Feng, L. Zhaoyang, S. Yuan, L. Zhenlin, "Development of Space Photographic Robotic Arm based on binocular vision servo," *Sixth International Conference on Advanced Computational Intelligence*, Hangzhou, 2013, pp. 345–349, http://dx.doi.org/10.1109/ICACI.2013.6748528.

[25] C. Wen-Chung, C. Chih-Wei, "Automatic Mobile Robotic Manipulation with Active Eye-to-Hand Binocular Vision," *33rd Annual Conference of the IEEE Industrial Electronics Society*, Taipei, 2007, pp. 2944–2949, http://dx.doi.org/10.1109/IECON.2007.4460000.

[26] P. C. Nunnally, J. M. Weiss, "An inexpensive robot arm for computer vision applications," *Energy and Information Technologies in the Southeast*, Columbia, vol. 1, 10989, pp. 1–6, http://dx.doi.org/10.1109/SECON.1989.132303.

[27] T. Kizaki, A. Namiki, "Two ball juggling with high-speed hand-arm and high-speed vision system," *IEEE International Conference on Robotics and Automation*, Saint Paul, MN, 2012, pp. 1372–1377, http://dx.doi.org/10.1109/ICRA.2012.6225090.

DOI: 10.15439/2015F131
ACSIS, Vol. 5

# Assessment of Software System Presentation Layers Based on an ECORAM Reference Architecture Model

Michał Turek, Jan Werewka, Kamil Sztandera, Grzegorz Rogus
AGH University of Science and Technology
Al. Mickiewicza 30
Kraków, Poland
Email: {mitu, werewka, sztandera, rogus}@agh.edu.pl

***Abstract*– Software systems are constructed using many different solutions. In practice it is important to assess different software systems. The most effective method of comparing software systems is to use a predefined reference architecture model as the basis for comparison. In this paper an ECORAM presentation layer reference architecture model is proposed. ECORAM is an extended common reference architecture model which contains a crucial number of elements used in systems investigation. The reference architecture is described using ArchiMate notation. For each element of the reference architecture a search for a mapping to the selected solution architecture is performed. When mapping fails it is assumed that a given feature of the reference architecture doesn't exist in the target system. Comparisons can be based on different aspects and the quality of metrics for predefined aspects may be defined. In this paper a comparison examples of different visualization solutions of business intelligence and game systems are presented.**

## I. Introduction

SOFTWARE systems are characterized by different architecture solutions. The solution architecture should be based on a reference architecture which is a template for a family of software systems or specific domains of interest.

The reference architecture [1] models the abstract architectural elements in the domain of interest, independent of the technologies, protocols, and products that are used to implement a specific solution for the domain, and represents the essence of the architecture of a collection of systems [2]. The purpose of the reference architecture is to provide guidance on developing architectures for new versions of a system or extended families of systems and products.

The reference model [1] describes the important concepts and relationships in the domain, focuses on what distinguishes the elements of the domain, and the reference architecture elaborates further on the model to show a more complete picture.

One of the main perspectives on Reference Architecture Models is that they are knowledge repositories which facilitate knowledge transfer and communication. However, they can also serve as frameworks, lexicons of terms and naming conventions, as well as structural relationships. One of the main challenges is to make the inherently abstract Reference Architecture Model concrete and understandable by providing sufficient specific information and guidelines.

When building software solution architecture in a given domain it is very important to have reference architecture. When obtaining software reference architecture it is important to have domain concepts and architecture reference models.

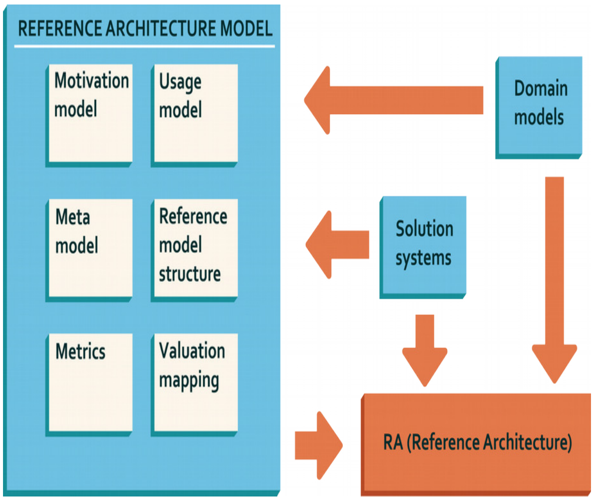


Fig 1. A reference architecture model in a reference architecture building context.

Domain models and solution systems may be used when building reference architecture models (fig. 1). The reference architecture model consists of the following elements:

- Motivation to build the reference architecture
- A usage model based on services (how the reference model will be used)
- A meta model defining the relationship in the set of related models.
- A reference model structure, consisting of elements and their descriptions.
- Metrics for defining the values of the reference model features.
- Valuation mapping. The elements of the reference architecture model can have different values for different stakeholders.

The reference architecture is developed based on a reference model, domain models and solution systems.

978-83-60810-66-8/$25.00©2015, IEEE

## II. Related works

The related works concerning the development of reference architecture models and the visualization capabilities of business intelligence systems are as follows.

Publication [3] describes a so-called RAModel (Reference Architecture Model) for reference architectures which can be included in reference architectures and presents all possible elements, broken down by type and relationship. The RAModel guides and improves productivity when new reference architectures are built and can offer effective support in the use and evolution of existing architectures. The RAModel also provides information on the elements and associated relationships which could be contained in reference architectures, independently from the application domains or purposes of such architectures. In the paper [4] a ProSARA process is presented for building reference architectures, focusing on how to design, represent, and evaluate such architectures. Publication [5] provides a framework for classifying software reference architectures. This framework is based on three dimensions: goals (why something is defined), context (where it will be used, who defines it, when it is defined), and design (what it describes, how specific and detailed this description is, how is it represented).Two purposes for using reference architecture are distinguished to standardize specific architectures (aimed at the interoperability of systems/components) and to enable the design of specific architectures (aimed at providing guidelines for the design of systems in the form of blueprints, patterns, etc.).

The question arises concerning to what extent reference architecture can be used for general application. In [6], CORA (COmmon Reference Architecture) is described for general use, and examples of its utilization are given, but the work proposes more of a reference model than a reference architecture.

In Business Intelligence Systems Presentation Layer solutions start to play an important role. In paper [7], which concerns big data analytics, two views of data presentation are described:

- Data visualization. The goal of data visualization is to communicate information clearly and effectively. In BI systems data volume is increasing, so it is important to use new visualization solutions.

- Analytics of data originating from different sources. The goal of web analytics is to automatically retrieve, extract, and evaluate information for knowledge discovery from web services and documents. Web content involves several data types such as text, symbolic data, metadata, hyperlinks and multimedia data including image, audio, and video. Multimedia analytics has to deal with multimedia data. Social media content contains text, multimedia, locations and comments.

Paper [8] indicates that using more interactive visual methods in BI would be beneficial for users. The paper contains confirmation of the following statements: Most visual methods currently applied in Business Intelligence are static or employ only very limited forms of interactivity; Increasing the interactivity of visual methods is desired by users. According to users, interactivity helps them gain information and knowledge in business data analysis.

In article [9] the visualization capability of BI systems in terms of developing effective visualization for addressing business problems is crucial to the success of BI. To amplify decision makers' perception and cognition of business data, BI systems often deploy a variety of visualization techniques. In the paper a Context Adaptive Visualization (CAV) framework is proposed which can guide the design and implementation of the visualization systems to be integrated into or used together with BI systems. The system requirements for flexible visualization development are based on visual solutions like creation, modification enhancement, integration and transformation.

## III. ECORAM presentation layer

Presentation plays a crucial role in IT systems. Historically, the presentation layer was separated from the application logic in order to enable different presentations without redesigning the whole application [10]. The presentation layer, especially the Graphical User Interface (GUI), has become a very important part of software for both users and developers [11]. There is a set of specific applications that relies mostly on User Interface (UI) and Human–computer interaction (HCI), with games and game controllers being among the most important representatives of this kind of application. From a design perspective, the presentation layer represents a specific set of problems that lead to a number of different solutions. The general idea of presentation layer design is to keep the presentation-related code separate from the domain-related code. The domain code should be completely unaware of the presentation logic in order to create reusable objects and components that are easy to test and maintain. The purpose of the presentation layer is to manipulate the presentation medium, give the user the look and feel experience (e.g. manipulate UI controls, present multimedia content), provide output to and receive input from the user.

The ECORAM (ECO Reference Architecture Model) is constructed from architecture building blocks (ABBs) that are comprised of information about the styles and architectural templates or features used [12]. The ECORAM is based on a layered model described in ArchiMate [13], so ABBs are grouped and presented in layers to emphasize their common properties. ArchiMate is an architecture description language (ADL), which is very suitable for communicating solutions among business and technology staff. The established approach assumes the developing of a new presentation model designed for product managers and system architects. This model should enhance understanding of user viewpoint dependencies and technical needs related to the entire service delivery in corporate architectures.

The main reason for developing the ECORAM was to express all system endpoint characteristics. Firstly, there is a need to classify the device held by the user. The first branch of the ECORAM is extremely simplified, adjusted just to describing the general device type. Then, on a deeper level, a feature set must be considered. The user endpoint may be full-featured, reduced or extremely basic.

Once the presentation platform is ready, another matter should be considered, namely the general type of information processed. This information may be current or historic, concurrent or object-based, fully verified or possibly damaged, complete or only partial, and so on.

The final branch should be a broad classification of the interaction with the user. Here, all kinds of HCI, content exchange or presentation services will be considered.

As long as the system has a user interface, it can be classified using the ECORAM. Finally, after a long period of discussions and fixes, a basic model branch list was established. All model branches have been included in a compact set.

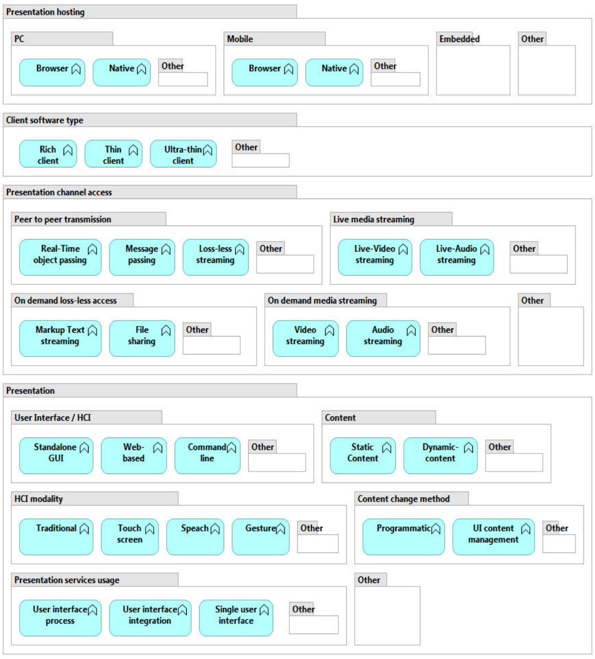


Fig 2. ECORAM presentation layer components in full set.

### *A. Presentation Hosting*

The first selected group of entities concerns presentation hosting. It represents three types of physical devices used to display information:

- PC – Any kind of personal computer (desktop, netbook, notebook, etc.): Browser – an application running in a Web browser on a PC; Native Client – a native application running on a specific OS.

- Mobile – Any kind of mobile device working on a specific OS (mobile phone, smart phone, tablet, etc.): Browser – an application running in a Web browser on a mobile device; Native Client – an application running on a mobile device.

- Embedded – a system with specific hardware dedicated to performing a computing function within a mechanical or electrical system (Real Time Systems, controllers, MP3 players, traffic lights, etc.). For embedded systems native, headless and other solutions can be used.

### *B. Client Software Type*

This division specifies the set of features and capabilities available in the solution and the degree of its dependence on the operating system. Most importantly, client types can easily be divided into three kinds:

- Rich Client – a typical standalone application deployed on a user's machine (Workstations, i.e. desktop UI, etc.) or a RIA (Rich Internet Application) running in a Web Browser, usually with a graphical user interface displaying data using a range of controls. Suitable for disconnected and occasionally connected scenarios.

- Thin Client – an application that depends heavily on a server in order to provide functionality. All the computing is performed on the server side, which distinguishes this type of software from rich clients.

- Ultra-thin Client – Server Based Computing (Independent Computing Architecture), delivering only presentation data and/or sending user actions.

### *C. Presentation Channel Access*

The next section specifies an abstract information flow between the user and the system. It should answer the question: "How do we deliver the presentation content?". Information can either be stored or retransmitted to a final user, or passed without any storage involvement. It can also be just a plain object, or possibly an infinite stream. Each user can also order just a part of a steam previously stored. Consequently, we propose the following list of presentation channel components:

- Peer to peer transmission - a transmission established for data generated in a real time. ECORAM assumes three kinds of such communication. 1) Real-Time object passing, occurring when an object is passed from the sender to the receiver in real time. These objects are based on patterns or classes and their size is limited. Objects can be cloned and sent to an unlimited number of receivers. (SOAP, REST etc.). 2) Message passing, messages can be read later (so storage is required). Messages may be cloned and delivered to a number of receivers without being deleted from storage (SMS, EMAIL, PAGER etc.). 3) Loss-less streaming, when a stream of data (taken in real time from a sensor or some such) is sent. It is encoded using dedicated protocol. The data size in this case is undefined.

- Live media streaming – a media stream does not require the use of any rigorous data quality assurance mechanisms. A stream may be defective or contain misrepresented data. In the case of this kind of stream, delay reduction is of the greatest importance: Live-Video streaming – Video content streaming; Live-Audio streaming – Audio content streaming.

- On demand lossless access – a stream with storage (on demand) – a limited size stream stored in data storage (similar to messages). Encoded using dedicated protocols: Markup Text streaming – a structured text described with mark-ups exchanged on demand between the sender and the receiver (RSS, TELETEXT, etc.); File Sharing – Files exchanged on demand between the sender and the receiver.

- On Demand Media Streaming – a media stream with data storage – on receipt, no type of rigorous data quality assurance mechanisms is required. Video Streaming – Video content streaming; Audio streaming – Audio content streaming.

### *D. Presentation*

A presentation mode, applied in a system interface, must deal with a user's (or human's) influence on that interface, and subsequently on the system itself. The following list divides interface presentation possibilities considering the type of user interaction:

- User Interface/HCI – The user interface is a part of a software program that allows the user to interact with computers and run their tasks. Human–computer interaction (HCI) describes the interaction between users and computers: Standalone GUI – a UI separated as a standalone application; Web-based – a User Interface based on a web application and characterized by request response navigation (interaction with the user); Command Line – Console-based applications offering an alternative text-only user experience, and typically run within command shells such as a Command window or a Power Shell. Most suitable for administrative or development tasks.

- HCI modality – the HCI modality is a general class of a sense through which the human can receive output from or give input to the computer: Traditional – Keyboard, mouse or touchpad; Touch Screen –Touch (multi-touch) or gestures on a Touch Screen; Speech – AUI (Audio User Interface) equipped with speech recognition solutions; Gesture – An interaction based on head and body movements

- Content – Describes whether the presented content will be changed by the user (dynamic content) or not (static content): Static content – content doesn't change, e.g. a static web-page; Dynamic Content – Presented content generated dynamically.

- Content change method – Describes how the presented content will be changed by the user: Programmatic – content refreshed or generated automatically, independently from the user's actions; UI Content management – Refreshed or generated depending on a user's actions, including Content Management Systems (CMS). A CMS is a computer program that allows content to be published, edited, modified and maintained from a central interface, and provides procedures to manage the workflow in a collaborative environment

- Presentation Services Usage – Describes the usage of different presentation services and user interactions between them: User Interface process – Synchronizes and orchestrates user interactions between presentation services for one actor and a one-time execution span [6]; User interface integration – Integrates different user interfaces and existing presentation services by offering various presentation integration services (i.e. WSRP, WS rendering); The User Interface Integration itself behaves like a User Interface, providing a presentation service by integrating a 'real' presentation service [6]; Single User Interface – One user interface – no need for integration

As we can see, the ECORAM is optimized to just four main sections, enabling the quick classification of a system. On the other hand, internal section granularity will support the expressing of all vital differences between two or more presentation layer structures, even those used in completely different kinds of IT systems.

## IV. ECORAM presentation layer application in Business Intelligence

ECORAM may be used for analyzing and developing human interactive computer systems (HICS). It is assumed that in BI systems interaction is performed in a timely manner. This means the reaction time is in accordance with human reaction. The application of an ECORAM for BI systems can be explained briefly as follows:

- The motivation is to use the reference model for a broad spectrum of applications. In particular, the model should be useful for BI systems.

- The model can be used in constructing BI systems , and also for investigating multimedia systems as potential data sources.

- The metamodel can be used as a default template for an application layer framework in an IT system of any kind. It shows meta-relations and meta-entities (here sections) that can be used or affected.

- Reference model structure, which is a composition of reference model elements. The description of model elements is also included here. The proposed structure can be applied for BI systems.

- Metrics defining the importance of the model features for BI systems.

- Valuation mapping – feature values important for BI System stakeholders.

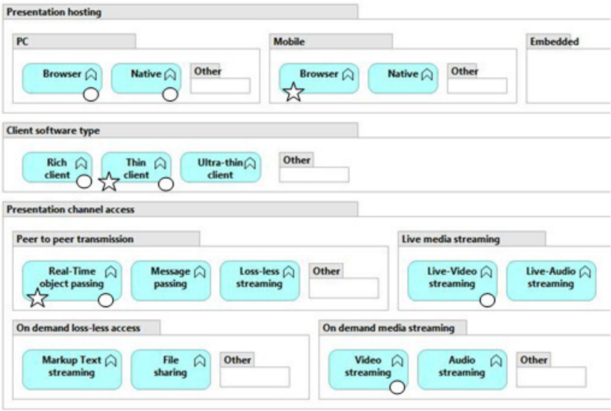


Fig 3. Sample mapping of asset usage: Taxi service management BI system (star symbols) and Integrated DVR Monitoring BI system (circle symbols).

Assessing software solution architecture is not an easy task. For the presentation layer some quality metrics can be distinguished, e.g. usability, user experience, and playfulness. ISO Standards 9126-3 [14] describe usability as product quality. According to ISO 9241-210 [15], "user experience represents a person's perceptions and responses resulting from the use or the anticipated use of a product, system or service". In terms of user experience, playfulness can come from a design that engages people’s attention or involves them in an activity for play, amusement, or creative enjoyment.

Fig. 3 shows the features used by two different BI systems and displays only a part of the ECORAM diagram. A full ECORAM diagram can be used to compare even totally different types of system, including different system domains Each component in a solution’s architecture is checked for features (assets) which also exist in the ECORAM. Various icon symbols are used to distinguish particular solution architectures (see Fig. 3). The diagram is used for a comparison of different systems. In a taxi business intelligence system [16] the customer uses a mobile device to connect to online taxi booking services and track the route of the approaching taxi. Additionally, by collecting information for many BI systems we can obtain statistical data which will give insight into how systems in a given domain are built. In [17] a Laboratory Management System (LIMS) delivering Real-Time Manufacturing Intelligence monitors a wide range of variables of both the manufacturing and the laboratory databases

## IV. ECORAM PRESENTATION LAYER VIDEO GAMES APPLICATION

ECORAM may be used for game systems comparison. Finally, three different popular game systems were considered: Angry Birds, the World of Tanks, and Kinect Sports.

Angry Birds is a puzzle video game developed by the Finnish computer game development company Rovio Entertainment. Inspired primarily by a sketch of stylized wingless birds, the game was first released for iOS in December 2009. Since then, over 12 million copies of the game have been purchased from the iOS App Store, which has prompted the company to design versions for other touchscreen-based smartphones, most notably those using the Android, Symbian, Windows Phone and BlackBerry 10 operating systems. It has since expanded to video game consoles and for PCs.

The World of Tanks is a multiplayer online game developed by a Belarusian company called Wargaming.net, featuring mid-20th century fighting vehicles. The focus is on player vs. player game play with each player controlling a tank or an armored vehicle. The World of Tanks debuted as an eSports game at the World Cyber Games 2012. On 10 June 2013, it was announced that the World of Tanks was coming to the Xbox 360 in summer 2013 as the World of Tanks Xbox 360 Edition. Xbox 360 players will use servers separate from those playing the Windows version and players of each version will have separate accounts.

Kinect Sports is a sports video game developed by Rare and published by Microsoft Game Studios for Xbox 360. The game is a collection of six sports simulations and eight mini-games, designed to demonstrate the motion-sensing capabilities of Kinect. The six sports included are: Bowling, Boxing, Track & Field, Table Tennis, Beach Volleyball and Football (Soccer in North America). Standing in front of a Kinect sensor, players compete by mimicking actions performed in real-life sports, such as throwing a javelin or kicking a football. The game received generally positive reviews from critics and sold over three million units as of April 2011.

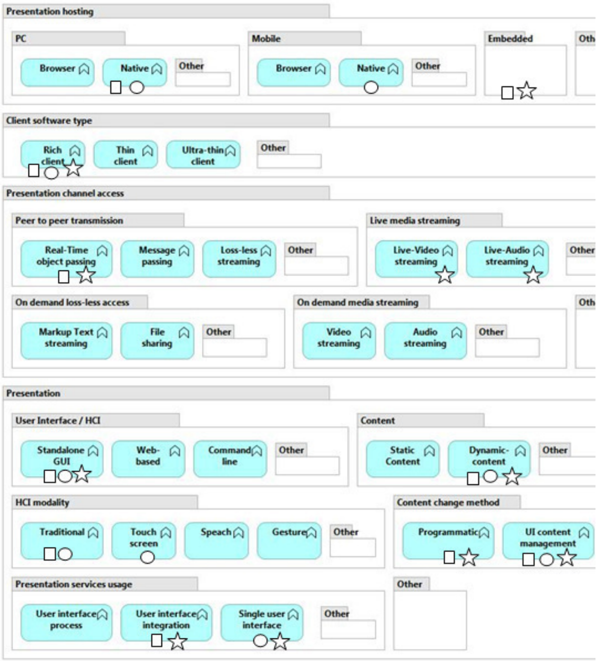


Fig 4. Sample system comparison: Kinect Sports (star symbols) and Angry Birds (circle symbols), World of Tanks (rectangle symbols).

The features used in the games can be evaluated from different viewpoints. The example considered here is playfulness. Playfulness motivates one to use the product and learn new features and technologies of the device. Paper [18] provides user interface characteristics influencing playfulness: creative enjoyment, challenge, curiosity, ability to customize the user interface, fun-in-doing, exploration, feedback, fantasy, metaphor and social interaction. A different perspective for understanding social game design is given by SoPlay heuristics [19]: accessibility – easy to approach, to understand and play; interrupt ability –spontaneous and irregular play sessions; continuity –a game world attracting the player to come back; discovery –new experiences and surprises; virility – growth in the player’s social network; narrativity –elicits curiosity; expression –self-discovery and customization; sharing – collaborating with friends, sociability among friends, social competition with others.

We have selected the following key aspects for a playful user interface: Creative enjoyment, challenge, feedback; Accessibility – making the game easy to approach, understand and play; Ability to customize the user interface; Social interaction.

$$PF(selected\ aspects) = \Sigma_{i=1}^{CE\ aspects}(\alpha_i CE_i) + \Sigma_{i=1}^{C\ aspects}(\beta_i C_i) + \Sigma_{i=1}^{SI\ aspects}(\gamma_i SI_i) \qquad (1)$$

For the selected games, a weighted quality factor can be determined to assess the play ability of the presenta-

TABLE 1.
KEY ASPECTS OF A PLAYFUL (PF) USER INTERFACE

| Key Aspect | Positive influence aspects (make use of) |
|---|---|
| Creative enjoyment (CE) | Touch screen, Speech, Gesture |
| Customization (C) | Dynamic content, UI content management |
| Social interaction (SI) | Live Video Streaming, Live Audio Streaming, User interface integration |

tion layer solution. Assuming that all coefficients αi, βi, γi are equal to 1, the playfulness of each game stemming from the presentation layer (fig. 4) can be determined in conjunction with Table 2. The results are as follows: PF(Angry Birds aspects)=3, PF(World of Tanks aspects)=4, PF(Kinect Sports aspects)=6. The simplified calculation demonstrate that the Kinect Sports game is characterized by the highest playfulness. The calculation of quality factors demonstrate only an approach to comparing different systems based on a proposed model.

## V. CONCLUSIONS

The process of ECORAM model development may be a part of knowledge management in a software development company [20]. The cooperative knowledge discovery model [21] including communication, coordination and cooperative decision making could be applied for developing and assessment of software solutions.

Reference architectures should accelerate work, reduce operating expenses and improve the quality of software system development, mainly due to reuse. Before reference architecture is designed it is important to develop an architecture reference model. To prove the flexibility of the model, a multiple system comparative analysis has to be performed, including system domains for auctions, games and BI. Each system chosen usually consists of many different presentation layer components that can be compared and placed in an ECORAM diagram. The diagram is capable of expressing all the differences with a sufficient level of detail.

Another possibility is to investigate the popularity of model elements in terms of usage by different user groups. For different kinds of users, we can now identify the presentation layer components used. The constant monitoring of a user group's activities and the system service counters will generate important statistical data for investment planning for future software line development. The solution will also make common feature exploration processes easier and will help system developers to design and refactor system architecture.

Finally, an ECORAM diagram can also be useful as a presentation tool during any kind of system requirements-establishment discussion with customers. The ECORAM diagrams are also used by computer science students when presenting their solution architectures, with the aim of quickly identifying the class of a developed software system.

It is possible to provide a development time-line in ECORAM diagrams (by creating a diagram set with timestamps) - and express a product line development regarding particular presentation layer components gradually involved. Building time-lined ECORAM diagrams for a set of similar products can easily express trends in product developments of that kind – helping to make a proper investment decisions in own product.

## REFERENCES

[1] Reference Architecture Foundation for Service Oriented Architecture Version 1.0, 2011, p 120,http://docs.oasis-open.org/soa-rm/soa-ra/v1.0/soa-ra.pdf

[2] Muller, G.: A Reference Architecture Primer. Buskerud University College, 2013, Gaudí documents, 2013, pp.21, www.gaudisite.nl

[3] Nakagawa E. Y, Oquendo F., Becker M.: RAModel: A Reference Model for Reference Architectures, 2012 Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture

[4] Nakagawa E. Y., Guessi M., Maldonado J. C., Feitosa D., Oquendo F.: Consolidating a Process for the Design, Representation, and Evaluation of Reference Architectures. Conference on Software Architecture (WICSA), 2014 IEEE/IFIP 2014, pp. 43 – 152

[5] Angelov S., Grefen P., Greefhorst D.: A Classification of Software Reference Architectures: Analyzing Their Success and Effectiveness, 2009 IEEE/IFIP WICSA/ECSA

[6] Elzinga T., van der Vlies J., Smiers L., The CORA Model, A Practical guide on using a COmmon Reference Architecture to design and deliver integrated IT solutions successfully, Sdu Customer Service, 2009

[7] Han Hu , Yonggang Wen, Tat-Seng Chua, And Xuelong Li: Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, IEEE Access, Vol. 2, 2014, pp. 652- 687

[8] Aigner W.: Current Work Practice and Users' Perspectives on Visualization and Interactivity in Business Intelligence, 17th International Conference on Information Visualization, 2013 IEEE, pp. 299-306.

[9] Bai X., White D., Sundaram D.: Context Adaptive Visualization for Effective Business Intelligence, Proceedings of ICCT2013, pp. 786-790

[10] G140: Separating information and structure from presentation to enable different presentations, http://www.w3.org/TR/WCAG-TECHS/G140.html

[11] Fowler M.: Development of Further Patterns of Enterprise Application Architecture, http://martinfowler.com/eaaDev

[12] TOGAF® Version 9.1, Open Group Standard, The Open Group, 2009-2011 pp. 692

[13] The Open Group: ArchiMate 2.0 Specification (2009-2012), p. 183

[14] ISO/IEC TR 9126-3:2003 Software engineering - Product quality - Part 3: Internal metrics, International Org. for Standardization (ISO)

[15] ISO FDIS 9241-210:2009. Ergonomics of human system interaction - Part 210: Human-centered design for interactive systems, International Org. for Standardization (ISO)

[16] Ch. Wang, H. Chen: From Data to Knowledge to Action: A Taxi Business Intelligence System, 15th International Conference on Information Fusion (FUSION),2012, p. 1632 – 1628

[17] J. Cooley, J. Petrusich, Delivering optimal real-time manufacturing intelligence, Proceedings of PICMET '13: Technology Management for Emerging Technologies,2013, p. 1658 – 1668.

[18] Ekaterina Kuts: Playful User Interfaces: Literature Review and Model for Analysis, Breaking New Ground: Innovation in Games, Play, Practice and Theory. Proceedings of DiGRA 2009

[19] Janne Paavilainen: Designing Social Network Games with SoPlay Heuristics, MindTrek 2010 Conference Workshop, Tampere, Finland.

[20] K. Jamróz, D. Pitulej, J. Werewka: Adapting Enterprise Architecture at a Software Devel-opment Company and the Resultant Benefits, in P. Avgeriou and U. Zdun (Eds.): ECSA 2014, LNCS 8627, pp. 170–185 (2014)

[21] X. Dai, N. Matta, G. Ducellier: CKD: a Cooperative Knowledge Discovery Model for Design Project. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, 1363–1369

# 5th Workshop on Advances in Programming Languages

PROGRAMMING languages are programmers' most basic tools. With appropriate programming languages one can drastically reduce the cost of building new applications as well as maintaining existing ones. In the last decades there have been many advances in programming languages technology in traditional programming paradigms such as functional, logic, and object-oriented programming, as well as the development of new paradigms such as aspect-oriented programming. The main driving force was and will be to better express programmers' ideas. Therefore, research in programming languages is an endless activity and the core of computer science. New language features, new programming paradigms, and better compile-time and run-time mechanisms can be foreseen in the future.

The aims of this event is to provide a forum for exchange of ideas and experience in topics concerned with programming languages and systems. Original papers and implementation reports are invited in all areas of programming languages.

## Topics

Major topics of interest include but are not limited to the following:

- 
- Automata theory and applications
- Compiling techniques
- Domain-specific languages
- Formal semantics and syntax
- Generative and generic programming
- Grammarware and grammar based systems
- Knowledge engineering languages, integration of knowledge engineering and software engineering
- Languages and tools for trustworthy computing
- Language theory and applications
- Language concepts, design and implementation
- Markup languages (XML)
- Metamodeling and modeling languages
- Model-driven engineering languages and systems
- Practical experiences with programming languages
- Program analysis, optimization and verification
- Program generation and transformation
- Programming paradigms (aspect-oriented, functional, logic, object-oriented, etc.)
- Programming tools and environments
- Proof theory for programs
- Specification languages
- Type systems
- Virtual machines and just-in-time compilation
- Visual programming languages

## Steering Committee

**Janousek, Jan,** Czech Technical University, Czech Republic

**Luković, Ivan,** University of Novi Sad, Serbia

**Mernik, Marjan,** University of Maribor, Slovenia

**Slivnik, Boštjan,** University of Ljubljana, Slovenia

## Event Chair

**Porubän, Jaroslav,** Technical University of Kosice, Slovakia

## Program Committee

**Barisic, Ankica,** Universidade Nova de Lisboa, Portugal

**Horvath, Zoltan,** Eotvos Lorand University, Hungary

**Janousek, Jan,** Czech Technical University, Czech Republic

**João Varanda Pereira, Maria,** Instituto Politecnico de Braganca, Portugal

**Kardaş, Geylani,** Ege University International Computer Institute, Turkey

**Kollár, Ján,** Technical University of Kosice, Slovakia

**Kosar, Tomaž,** University of Maribor, Slovenia

**Liu, Shih-Hsi Alex,** California State University, United States

**Luković, Ivan,** University of Novi Sad, Serbia

**Mandreoli, Federica,** University of Modena, Italy

**Martínez López, Pablo E. "Fidel",** Universidad Nacional de Quilmes, Argentina

**Mernik, Marjan,** University of Maribor, Slovenia

**Milasinovic, Boris,** University of Zagreb Faculty of Electrical Engineering and Computing, Croatia

**Moessenboeck, Hanspeter,** Johannes Kepler Universitat Linz, Austria

**Papaspyrou, Nikolaos,** National Technical University of Athens, Greece

**Rangel Henriques, Pedro,** Universidade do Minho, Portugal

**Sierra Rodríguez, José Luis,** Universidad Complutense de Madrid, Spain

**Slivnik, Boštjan,** University of Ljubljana, Slovenia

**Splawski, Zdzislaw,** Wroclaw University of Technology, Poland

**van der Meer, Arjan,** Eindhoven University of Technology

**Watson, Bruce,** Stellenbosch University, South Africa

DOI: 10.15439/2015F274
ACSIS, Vol. 5

# Usability of a Domain-Specific Language for a Gesture-Driven IDE

Michaela Bačíková, Martin Maričák, Matej Vančík
Technical university of Košice
Letná 9, 042-00 Košice, Slovakia
Email: michaela.bacikova@tuke.sk, {martin.maricak, matej.vancik}@student.tuke.sk

***Abstract*—User interfaces (UIs) are advancing in every direction. The usage of touch screen devices and adaptation their UIs lives its boom. However integrated development environments (IDEs) that are used to develop the same UIs are oversleeping the time. They are directed to developing usable software, but forgot to be usable by themselves. Our goal is to design a new way of user interaction for common IDEs with the help of touch. The target group are hybrid devices formed by a physical keyboard and either an integrated, or separate, touch screen display. In this paper we describe a set of general purpose and domain-specific gestures which represents a language for working with a touch-driven IDE and provide a method their design. We performed two studies with developers from industry and university and developed a prototype of a gesture-driven IDE to evaluate the usability of the presented approach.**

## I. Introduction

The accessibility of alternative devices such as wearable gadgets, smartphones, tablets, tablet PCs, wall-sized displays, touch-enabled displays, kinect and others, has given a rise to different design guidelines for application user interfaces (UIs) which, though still young and imperfect, become more and more dominant. According to the world statistics a large part of interaction has moved to touch devices and the vast majority of mobile devices now supports touch control. Touch interfaces are natural and easy to get used to and they also affect the way people try to control other devices such as common PCs or notebooks. Therefore, *hybrid devices*[1] have emerged and operation systems try to adapt their UIs and functionality to touch.

Although touch control support has moved a large leap forward, it still has its disadvantages. Touch gestures are not entirely uniform on all platforms and in many cases, a touch display alone is not sufficient to fully handle all desired functionalities. A typical example is *writing on a virtual keyboard*. Vriting on a virtual keyboard is slow, imprecise and the keyboard occupies a large part of the screen [1]. Still, the situation is much better than before and the rise of platform guidelines for touch devices with different screen sizes helps the situation.

Despite the disadvantages, the situation is much better than in case of integrated development environments (IDEs). They are made to improve developer productivity as much as possible [2] but it is as if developers were primarily focusing on making UIs and interaction better just for their users while forgetting about themselves. The support of touch for the most common IDEs literally overslept the time. Not only the current IDEs are not ready for advanced work on touch displays, they are not even prepared for everyday touch use stereotypes and are largely limited to interaction via traditional hardware devices (e.g., the ubiquitous keyboard and mouse) [3]. It is not possible to use swiping to scroll, pinch to zoom or tapping the same way as in every other touch-enabled UI. From the IDEs we have analyzed, the only IDE ready touch is Eclipse, statistically the most used IDE for Java language. However, it still offers just fundamental features. As for the second in line, IntelliJ IDEA, despite marked innovative and usable, does not even support the most fundamental ones. Instead of browsing or scrolling through code, the swipe gesture selects text, infringing the common interaction stereotypes and adjustment of panel sizes or zooming by using pinch gesture is not possible at all.

Programming requires the speed of interaction and creativity, touch support is unambiguously a step forward because it brings a different form of user experience and according to Greene [4], supports creative thinking. Our goal is not to force the work exclusively via touch gestures but to create a symbiosis of keyboard, touch and, alternatively, mouse. Begel [5] experimentally verified that programmers are so used to physical keyboard that it would be disadvantageous or at most impossible for them to explicitly use virtual keyboard, which, in addition, has a low precision and does not enable writing by ten fingers.

Supporting common general gestures in IDE may not be enough to speed up the developer work. A specific solution is necessary to not just support the basic gestures, but enable advanced interaction adjusted for a specific IDE platform and programming language. In this paper, we show a method of designing a set of gestures for such advanced interaction as a way of designing the lexical part of a visual language. The focus is mainly on usability [6], [7], [8], domain usability and specificity [9], simplicity of use, learnability and compliance with standard design patterns and guidelines.

The main contributions of this paper are:

- Definition of more complex general-purpose gestures in the context of IDEs.

[1] In this paper, the term *hybrid devices* will be used to refer to devices with a keyboard and a touch display such as tablet PCs or combinations of desktop PCs/notebooks with an external touch display.

978-83-60810-66-8/$25.00©2015, IEEE

- A method of collecting, identifying and designing a set of domain-specific gestures via multiple user surveys and interviews including industry programmers and gesture recorder.
- Verbal definition of a domain-specific language of gesture patterns to support highly specialized work with code and consequently a design and realization of drawn gesture input and recognition in a form of a gesture-driven IDE for Android. When compared to a physical keyboard, we hypothesize that touch gestures are *more learnable and memorable than key shortcuts*.
- Evaluation of the proposed approach by means of usability evaluation.

Although our prototype is designed for Android devices, we do not focus on touch devices explicitly (although they may be a possible target), our target group are hybrid devices.

## II. A Touch-Enabled IDE

Here, we will describe a design of fundamental gesture interaction for touch devices related to IDEs. Elementary gestures are based on usability guidelines [10], [11], analysis of existing desktop and touch-based IDEs and discussions with industrial developers. Primarily, we followed the Google Material design guidelines [12] that also define the most common gesture patterns (Fig. 1).

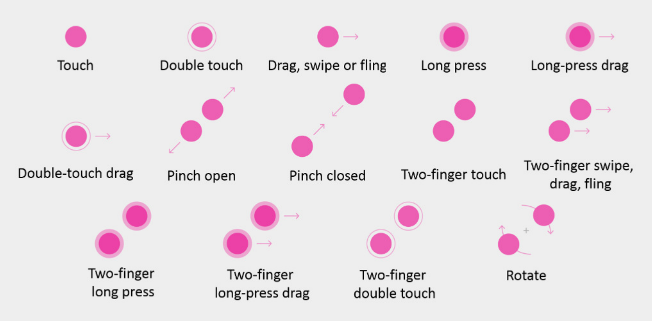


Fig. 1. The mechanics of touch gestures according to the Google Material Guidelines

We identified the following two fundamental categories of gestures related to their design:

1) *General-purpose gestures (GPGs)* - used for standard interaction. They can be used in different context, however their meaning is always the same or at least very similar. The most common GPGs are touch, double-touch, swipe or pinch. This category can be considered platform-independent although according to Wroblewski [13], not all gestures are applied on all platforms consistently to the same functionalities.
2) *Drawn gestures* - a specific type of gestures that have no standard or specification indicating their form, purpose or semantics. For example, drawing a circle can be used for restoration, rotation or selecting multiple items in a list. The design of semantics of drawn gestures is often intuitive and closely linked to the context of their use.

Breaking the stereotypes of *GPGs* can cause serious usability issues, mainly in the case of advanced or experienced touch device users. Therefore it is good to abide the standard interaction in case of this category. In case of an IDE, it is advantageous to design additional gestures similar to general-purpose ones by using multiple fingers to support more action shortcuts. When designing the set of GPGs, we primarily followed the guidelines of the Android platform and we were also inspired by the iOS platform guidelines for multi-touch gestures [14].

Based on the simple gestures displayed in Fig. 1, we further designed the following additional multi-touch gestures for common IDEs:

1) Double-touch to *select words*.
2) Tripple-touch to *select lines*.
3) Two-finger horizontal swipe for *undo/redo* inspired by the usage on Apple trackpads [14].

## III. Domain-Specific Language for a Gesture-Driven IDE

IDEs are used in the domain of programming and to be able to support at least the standard interaction with an IDE, we need to design *domain-specific gestures*.

The set of gestures supported by an application can be perceived as a *sign language*, which the user has to learn to be able to communicate effectively. A correct conceptual model helps users to better understand actions performed upon a UI, thus speeds up the learning process.

From this point of view, the list of domain-specific gestures (both general-purpose and drawn) described further in this paper can be perceived as a *domain-specific language* of the touch-enabled IDE:

1) Concrete syntax is defined by the *gestures* themselves: touch or movement, finger count, touch count and movement directions.
2) Abstract syntax is defined by the *actions* that the gestures *cause*.
3) Semantics is defined by their *type and implementation*.

Because of the limited space and for we design a simple visual language (the set of gestures is kept small), we will not formally describe the semantics of the gesture DSL in this paper. We expect that common language constructs, such as loops, conditions, comments or class and method declarations are known to common programming language designers. Therefore we expect the semantics to be clear from the purpose of the gesture. For each gesture, the semantics is a generated programming language construct. We use templates to generate the constructs similar to the ones used in common IDEs using the Velocity template engine. Concrete and abstract syntaxes are described verbally.

### A. General-Purpose Gestures

We designed the following GPGs for touch-enabled IDEs:

1) Swiping:
   a) Two-finger swipe for *"stuck shift"* - gradual shift from method to method by swiping two fingers.

   b) Three-finger swipe to the *end/beginning of the file*.
2) Code folding:
   a) Three-finger pinch for *method folding*.
   b) Four-finger pinch for *"semantic zoom"*.
   c) Five-finger pinch to *fold all class members*.

"*Semantic zooming*"[2] is a way of grouping content into specific predefined categories. In case of an IDE it would be a combination of related methods or blocks of code such as getters, setters, void methods, public methods, imports, nested classes, etc., into a single group. After performing the gesture, the user is promted to select a semantic group. After selection, everything else not belonging to the selected group is folded. This enables to focus only on the important parts of the code.

In case of pinch gestures, the action of pinching-in standardly folds class members and the opposite direction of pinching-out causes unfolding action.

### *B. Drawn Gestures*

As we stated in section II, drawn gestures are not subject to any specifications or conventions. Therefore their design needs to be approached with caution. They are specific for the target domain of use, therefore *domain experts should be included* in the design and development process. A domain analysis was performed to determine the way how programmers interpret different language constructs visually. We interviewed multiple developers from industry and their claims were supplemented by a survey. In the next subsections we will describe the results of the survey that was performed in two iterations.

*1) First Survey - Collection of Patterns:* In order to elicit the correct conceptual model most effectively, a support application for Android touch devices called *"Gesture Recorder"* (Fig. 2) was created. The whole survey process was performed in the application, using a Google Nexus 7 tablet, which asked users to interpret given programming language constructions by directly inducing gestures on touch screen. We used a standard Android tablet for the survey.

We wanted to cover a wide spectrum of programmer experience, therefore we included 68 participants, from which 8 considered themselves advanced programmers, 38 intermediate and 22 were beginner programmers. During the survey, each participant performed the survey independently to prevent influence and was assisted by one member of our research team in order to be able to correctly work with the survey application.

The survey application prompted the participants to interpret the following constructs by a drawn gesture: class declaration, comment, variable declaration, loop statements (both for, while and do-while), condition statement, method declaration and deleting a line.

For each prompt, an example of a language construct was displayed in the application to help the user to better connect his/her mental model with a new gesture. After all language constructs, participants were given a possibility to add their own idea for a new gesture and language construct. Then they were to fill a questionnaire about their programming experience. The results for each participant were sent to a distant server, manually checked and analyzed.

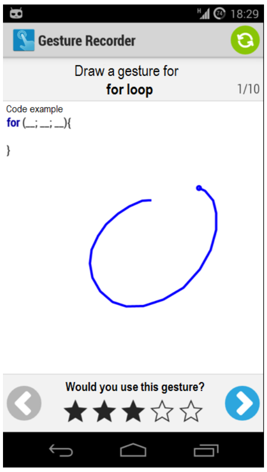


Fig. 2. The GestureRecorder survey application

*a) Evaluation Method:* Similarity of gesture shapes was taken into account via their abstractions into patterns. Differences in size, shape and format were abstracted into a single unified pattern while the direction of drawing was preserved. Differences were neglected only if the conceptual model of the gesture interpretation remained unchanged. The patterns with low repetition rate were omitted. Fig. 3 shows the simplified survey results. Each group represents one language construct and indicates the most common gesture pattern and the number of participants that used the pattern. The direction of drawing is indicated by the small circle representing the final point of drawing.

*b) Results:* The results in the case of a class were influenced by the fact that the respondents come from Slovakia, where class is called "trieda" and begins with the letter "T". Many of the participants also included "N" as in "new", but the number was not as significant. Both were mainly beginners or at least advanced programmers and from our experience, sooner or later older and more experienced programmers begin to in English language explicitly and do not perceive the class declaration and instantiation as the same concept anymore. In case of variable declaration, the results were unclear, therefore we decided not to include the gesture without further research. In case of cycles, as for the shape, the results jointly indicate a circle, as for the direction, the results are unclear.

*2) Second Survey - Suitability of Collected Patterns:* Because the results from the first survey were very variable, we decided to perform a second survey. The goal was to determine which shape of the previously identified ones was the most appropriate for the given language construct.

The second survey was performed on 65 new participants. In the first survey we covered different levels of programmer experience to get universal results but this decision has proven to be wrong. In the second round we rather decided to target

[2] More information about semantic zooming can be found in Microsoft official guidelines: https://msdn.microsoft.com/en-us/library/windows/apps/hh465319.aspx

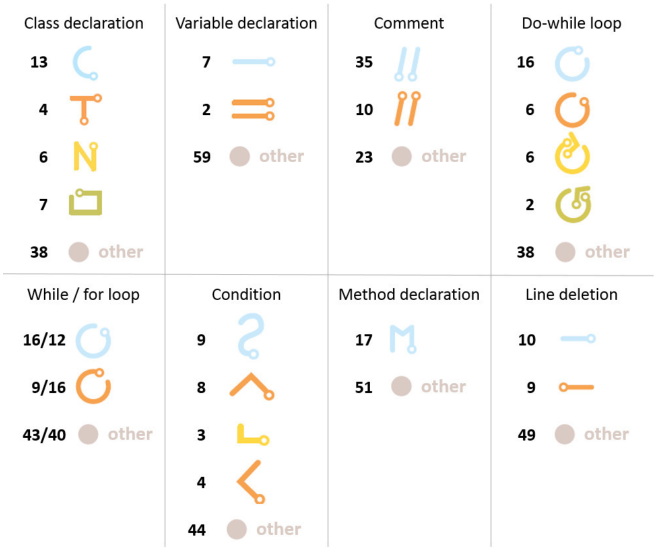


Fig. 3. The results of the first survey

the survey explicitly to last-year master students of informatics to achieve more relevant results. We used a questionnaire that contained only questions related to gestures. We included the most frequent answers from the first survey so the users could select from existing choices.

*a) Results:* The results of the second survey were much clearer. In case of class declaration, a significant number of 69% respondents have chosen the shape of the "C" letter which indicates it is an appropriate representation of a the gesture. In case of loops, the results indicate that the direction of drawing is not significant. There are multiple research papers that try to reveal the reason of the direction. A cross-cultural study by Amenomori et al. [15] identified a direct impact of culture and learning methods on the direction. Based on the facts we conclude that the gesture should be unified for all three loop constructions and should support both drawing directions. To unify all loop types, we need to add an additional step after gesture drawing, where the user selects the particular loop construction (while, do-while or for) according to his/her choice.

The results for condition statement were not as definite as in the case of class gesture, however a slightly more significant number of users selected the shape resembling a question mark. Since branching can again be noted by multiple constructs such as if, if-else, switch or shortened if-then-else statement (:?), the gesture can unify all possible constructs and enable a choice after, similarly to the loop gesture.

As for the final question, most of the respondents agreed on a horizontal line for line deletion.

## IV. Implementation of the Gesture DSL

At this point we will describe one of the ways of implementing recognition of the described gestures. We chose Android as a model platform mainly because of its accessibility, openness and good support of developing and recognizing custom gestures. However, the mechanics of gestures and the described procedures are common for any platform. The goal is not to point out to the details but to show the method of developing a gesture recognition engine.

We developed two prototypes of a simple gesture-driven IDE with the goal to design the gesture recognition algorithms and tested it with industrial developers. We focused on maximal usability. We deesigned and implemented recognition of both complex GPGs and drawn gestures.

Recognition of GPGs is implemented via native platform libraries. Recognition of drawn gestures is more complex and uses machine learning and recognition probabilities. Due to the limited space, we will only describe the implementation of domain-specific *drawn gestures*.

### A. Recognition of Drawn Gestures

To recognize drawn gestures we used the Android `com.android.gesture` package containing classes such as `Gesture`, `GestureOverlayView` and `Prediction` enabling possibilities to record, store, read and recognize gestures. The package is undocumented, therefore we will describe the design in more detail. `GestureOverlayView` is a graphical component that enables the actions of gesture recording and displaying it as curves. The recorded and predefined gestures are stored in a binary file located in application memory and contain a list of points and a single name may be assigned to each gesture.

Recognition works on the principle of comparing a drawn gesture with the whole file content and the result is a list of identified patterns with assigned probabilities. However, the default recognition is not very accurate, especially if the stored patterns are slightly similar to each other or if the drawn gesture did not imitate the target pattern too precisely. The more patterns are stored in the database, the less precise is the recognition. In case there are multiple gesture patterns, the recognition rate can be less than 40% which is unacceptable. Taking this fact into account, we needed to adjust the default recognition process so that each pattern could have a *database of multiple possible shapes*. We recommend to create the databases by recording pattern shapes from different users and multiple times and also to record the gestures drawn by the IDE users during their first use by means of a tutorial.

The *recognition probability* $p$ of match is determined by a positive number. The exact meaning of $p$ was not clear since the package is documented, thus the value was determined empirically for each pattern as stated below. In general, the gesture is successfully recognized $p > 1$ and the highest probability is considered as correctly recognized. This applies only after verifying the number of traces forming the gesture. If a circle is recognized after drawing two mirror-inverted C characters, the result will not be considered correct, even though it has the highest $p$ value.

The database for each pattern was created by drawing multiple slightly deformed shapes and in both drawing directions. Because of significant differences between probabilities for patterns with different complexity, an optimal $p$ interval was determined empirically for each pattern as follows:

- *Loop gesture* - is considered to be recognized if $p > 4$.
- *Condition gesture* - the most complex shape of all because of two changes in the direction. Recognized when $p > 3$.
- *Class gesture* - a higher value has proven to be optimal. This gesture is recognized if $p >= 5$.
- *Comment gesture* - a simple gesture of two consequent lines. Recognized if $p > 10$
- *Line deletion gesture* - the most simple gesture, almost identical to comment, but with only one line and different direction. Valid recognition if $p >= 10$.

The values were determined by re-entering gestures multiple times and observing the $p$ value after recognition. We verified both positive and negative cases and with the stated $p$ values, we achieved almost 100% success rate.

Since the recognition database is built using a simple form of machine learning and the recognition thresholds are determined empirically by applying probabilities, it is important to consider the number of available patterns in the database. Low number of predefined gestures that are not very complex and similar to each other is optimal. Also, it is very important *not* to continue the machine learning process during application use, because the user might draw incorrect shapes and hinder the recognition of the predefined gestures. Instead, we recommend to allow the user to create his/her own gestures and teach the correct and incorrect shapes to the recognition algorithm. This is also a problem of many systems for recognizing handwritten text. Each person has his/her own writing style and when supporting as many characters as the whole alphabet, the recognition can be incorrect multiple times and lead to reluctance and rejection of using the system. Moreover, the user is only able to remember a limited number of gesture shortcuts and forcing his/her to remember a whole database of patterns can reduce usability.

## V. Gesture DSL - Usability Test

To evaluate the gesture DSL, we created a prototype of a simple gesture-driven IDE for Android according to the design described in this paper and performed a usability evaluation.

Our target group were Java programmers, experienced in using IDE and minimally in using a touch-enabled smartphone or a tablet. For the reasons indicated by Nielsen [16], [7], we performed the usability evaluation with five participants and we targeted at qualitative evaluation. The participants were given tasks related to tutorial, GPGs and 2 tasks targeted at general work with the IDE.

Each testing was performed in the presence of one member of the research team. The sample was represented by five experienced developers with 1-5 years of practice in industry, using mostly IntelliJ IDEA. To avoid influence, the test was performed separately with each participant. In all cases we used Google Nexus 7 tablet and each time the application was reset to its original settings. Two of the respondents were given an external keyboard.

At the beginning, the participants were briefly familiarized with the research and with the testing process and they were given an in-app introductory tutorial. After finishing the tutorial, the participants were to perform the tasks on a sample class and to express their opinion.

### A. Observations and Conclusions

Gesture discoverability was 100%, each participant was able to perform the tasks successfully. The feedback received from the subjects indicated that almost all participants liked the idea of manipulating with the editor using GPGs and generating parts of code using drawn gestures. However, the idea of using it on a tablet was not accepted, each participant expressed the wish to use the gesture language on a common IDE which indicates the potential of our approach on hybrid devices. We concluded that the negative opinion towards tablets is related to writing code using a virtual keyboard. The results were definitely better in case of participants that used an external keyboard.

The gesture for folding all methods was marked most useful, however the users expressed the idea of subsequent use of the gesture multiple times to fold methods to regions. All of them liked the semantic zoom feature and indicated that it could replace the common IDE feature of code structure preview. They suggested to add a filter for constructors, member variables and a filter for displaying all method names.

The possibility to define custom gestures was described as a feature with high potential. Each developer mentioned an example of use, such as definition of custom predefined code parts and templates or combining with macros.

The design of new gestures is a major challenge. The gestures have to be simple, easy to learn and quick to write. This holds for the set presented in this paper, but with a growing number of gestures, their complexity might rise significantly, which could complicate their use. It might be beneficial to explore multiple combinations of gestures and patterns [17] [18] of their use.

## VI. Related Work

Biegel et al. [19] propose an inspirational approach to "touchifying" an IDE. They highlight multiple issues of common IDEs related to supporting touch handling in general. Authors present a solution on Eclipse IDE. Compared to our gesture set, they used already existing GPGs to support multiple functionalities. In their study, they did not include domain-specific drawn gestures. The fact that programmers rather prefer the combination of a traditional PC, touch display and a mouse shown in their paper *supported our findings* and confirms that this approach has a potential to be used on hybrid devices.

There are multiple works that focus on code refactoring triggered by gestures, such as the works by Murphy-Hill et al. [20] and Raab et al. [21]. Compared to their solution, we use a small set of gestures for commonly used features and for the sake of usability we never use multiple composed gestures to trigger a single operation. Lee et al. [22] propose a solution where refactoring can be triggered via drag-and-drop gestures but in contrast to our work they optimized the gestures for using a mouse.

Hesenius et al. [23] introduced an environment for tablets with advanced features but focused primarily on creating a new environment, not adaptation to hybrid devices. Delimarschi et al. [3] describe a natural UI for graphical IDE using a Kinect voice and physical gesture commands. Although we favour the potential of voice commands as means to speeding-up the developer work, voice recognition is still not mature enough to fully support multiple languages. As for physical gestures, the potential is less definite. Edwards and Barnette [24] unsuccessfully used tablet PCs with a pen in a laboratory programming course without adapting the software to this novel device.

Begel et al. [5] mentioned that it is possible that developers are less efficient by using a natural interface instead of typing. This is related to the fact that typing on a physical keyboard is already ingrained in every developer's habits, while more natural interaction might not be as standard in the development process. We agree that in the case of IDEs, not each natural interaction type is beneficial, however we argue for touch input being advantageous based on our observations and user feedback. Direct manipulation characteristics of touch helps to reduce the cognitive load [19]. Further investigation is needed to support this claim.

## VII. Conclusion

We introduced the concept of a DSL of gestures in the context of IDEs. We performed multiple studies to explore the visual representations of conceptual models of programming constructs and developed a prototype tool which served to conduct a pilot study to assess the feasibility of our approach. Despite the challenges and current limitations, we believe that it has the potential to improve and speed-up development and ease of use. To make the approach more accessible and verifiable in the context of hybrid devices, we have set forth the following goals: 1. Implement a prototype into a common desktop IDE such as Eclipse, IntelliJ IDEA or NetBeans in form of a plug-in. 2. Integrate learning capabilities that would allow the system to adapt to the user and his/her drawing style and allow the user to add new custom gestures via the existing IDE-native code template engine. 3. Generalize the approach beyond Android and Java, for a variety of operating systems and programming languages. 4. Experimentally identify the optimal number, complexity and combinations of drawn gestures.

## Acknowledgment

This work was supported by VEGA Grant No. 1/0341/13 Principles and methods of automated abstraction of computer languages and software development based on the semantic enrichment caused by communication.

## References

[1] A. Schade. Large touchscreens: What is different? [Online]. Available: http://goo.gl/jCsnDJ

[2] I. Zayour and H. Hajjdiab, "How much integrated development environments (ides) improve productivity?" *JSW*, vol. 8, no. 10, pp. 2425–2431, 2013. doi: 10.4304/jsw.8.10.2425-2431

[3] D. Delimarschi, G. Swartzendruber, and H. Kagdi, "Enabling integrated development environments with natural user interface interactions," in *Proc. 22nd Intern. Conf. on Program Comprehension*, ser. ICPC 2014. New York, NY, USA: ACM, 2014. doi: 10.1145/2597008.2597791. ISBN 978-1-4503-2879-1 pp. 126–129.

[4] S. L. Greene, "Characteristics of applications that support creativity," *Commun. ACM*, vol. 45, no. 10, pp. 100–104, Oct. 2002. doi: 10.1145/570907.570941

[5] A. Begel and S. L. Graham, "An assessment of a speech-based programming environment," in *IEEE Symp. on Vis. Languages and Human-Centric Comp., VL/HCC'06*, Sept 2006. doi: 10.1109/VLHCC.2006.9 pp. 116–120.

[6] Usability partners, "ISO standards in usability and user-centered design". [Online]. Available: http://usabilitypartners.se/about-usability/iso-standards

[7] J. Nielsen, *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 0125184050

[8] D. A. Norman, *The Design of Everyday Things*. New York, NY, USA: Basic Books, Inc., 2002. ISBN 9780465067107

[9] M. Bačíková and J. Porubän, "Ergonomic vs. domain usability of user interfaces," in *2013 The 6th Intern. Conf. on Human System Interaction (HSI)*, June 2013. doi: 10.1109/HSI.2013.6577817. ISSN 2158-2246 pp. 159–166.

[10] D. Wigdor and D. Wixon, *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*, 1st ed. San Francisco, CA, USA: Elsevier (Morgan Kaufmann Publishers Inc.), 2011. ISBN 978-0-12-382231-4

[11] C. Abras, D. Maloney-krichmar, and J. Preece, "User-centered design," in *Encyclopedia of Human-Computer Interaction, Bainbridge, W.* Thousand Oaks: Sage Publications, 2004.

[12] Google 2014: Material design guidelines. [Online]. Available: http://www.google.com/design/spec/material-design

[13] L. Wroblewski. (2010) Touch gesture reference guide. [Online]. Available: http://www.lukew.com/ff/entry.asp?1071

[14] Apple 2014: Mac basics: Multi-touch gestures, official apple support". [Online]. Available: https://support.apple.com/en-us/HT4721

[15] M. Amenomori, A. Kono, J. S. Fournier, and G. A. Winer, "A cross-cultural developmental study of directional asymmetries in circle drawing," *Journ. of Cross-Cultural Psychology*, vol. 28, no. 6, pp. 730–742, 1997. doi: 10.1177/0022022197286005

[16] N.-N. Group. How many test users in a usability study? [Online]. Available: http://www.nngroup.com/articles/how-many-test-users/

[17] J. Kollár et al., "Plero: Language for grammar refactoring patterns," *FedCSIS '13*, pp. 1503–1510, 2013.

[18] M. Nosál' and J. Porubän, "Xml to annotations mapping definition with patterns," *Comp. Sci. and Inf. Syst.*, vol. 11, no. 4, pp. 1455–1478, 2014. doi: 10.2298/CSIS130920049N

[19] B. Biegel et al., "U can touch this: Touchifying an ide," in *Proc. 7th Int. Works. on Coop. and Human Asp. of Soft. Eng.*, ser. CHASE 2014. New York, NY, USA: ACM, 2014. doi: 10.1145/2593702.2593726. ISBN 978-1-4503-2860-9 pp. 8–15.

[20] E. R. Murphy-Hill, M. Ayazifar, and A. P. Black, "Restructuring software with gestures," in *VL/HCC*. IEEE, 2011. doi: 10.1109/VL-HCC.2011.6070394. ISBN 978-1-4577-1246-3 pp. 165–172.

[21] F. Raab, C. Wolff, and F. Echtler, "Refactorpad: Editing source code on touchscreens," in *Proc. 5th ACM SIGCHI Symp. on Eng. Interactive Comp. Syst.*, ser. EICS '13. New York, USA: ACM, 2013. doi: 10.1145/2494603.2480317. ISBN 978-1-4503-2138-9 pp. 223–228.

[22] Y. Y. Lee, N. Chen, and R. E. Johnson, "Drag-and-drop refactoring: Intuitive and efficient program transformation," in *Proc. 2013 Intern. Conf. on Soft. Eng.*, ser. ICSE '13. Piscataway, NJ, USA: IEEE Press, 2013. ISBN 978-1-4673-3076-3 pp. 23–32.

[23] M. Hesenius, C. D. O. Medina, and D. Herzberg, "Touching factor: Software development on tablets," in *Soft. Comp.*, ser. Lect. Notes in Comp. Sci., vol. 7306. Springer, 2012. doi: 10.1007/978-3-642-30564-1_10. ISBN 978-3-642-30563-4 pp. 148–161.

[24] S. H. Edwards and N. D. Barnette, "Experiences using tablet pcs in a programming laboratory," in *Proc 5th Conf. on Inf. Techn. Edu.*, ser. CITC5 '04. New York, USA: ACM, 2004. doi: 10.1145/1029533.1029573. ISBN 1-58113-936-5 pp. 160–164.

DOI: 10.15439/2015F325
ACSIS, Vol. 5

# A New Algorithm for the Determinisation of Visibly Pushdown Automata

Radomír Polách, Jan Trávníček, Jan Janoušek, Bořivoj Melichar
Department of Theoretical Computer Science
Faculty of Information Technology
Czech Technical University in Prague
ul. Thákurova 9, 160 00 Prague 6, Czech Republic
Email: Radomir.Polach@fit.cvut.cz, Jan.Travnicek@fit.cvut.cz, Jan.Janousek@fit.cvut.cz, Borivoj.Melichar@fit.cvut.cz

***Abstract*—Visibly pushdown automata are pushdown automata whose pushdown operations are determined by the input symbol, where the input alphabet is partitioned into three parts for push, pop and local pushdown operations. It is well known that nondeterministic visibly pushdown automata can be determinised. In this paper a new algorithm for the determinisation of nondeterministic visibly pushdown automata is presented. The algorithm improves the existing methods and can result in significantly smaller deterministic pushdown automata. This is achieved in a way that only necessary and accessible states and pushdown symbols are computed and constructed during the determinisation.**

***Index Terms*—Pushdown automata, visibly pushdown automata, deterministic pushdown automata, determinisation of visibly pushdown automata.**

## I. Introduction

PUSHDOWN automata, which accept context–free formal languages, are one of the fundamental models of computation of the Theory of formal languages and automata [8]. Every nondeterministic finite automaton, which accepts a regular language, can be determinised and the theory of the determinisation of finite automata is simple and well–researched: states of an equivalent deterministic finite automaton represent so–called deterministic subsets of states of a given nondeterministic finite automaton [8]. The general determinisability does not hold for the case of all types of nondeterministic pushdown automata. The class of deterministic context–free languages is a proper subclass of context–free languages, ie for some nondeterministic pushdown automata their equivalent deterministic versions do not exist. Generally, it is not known how to decide for a given nondeterministic pushdown automaton whether there exists a deterministic equivalent or not. There is a lack of results in the theory of the determinisation of nondeterministic pushdown automata, although such results would be usable, eg when constructing practical deterministic algorithms from nondeterministic pushdown automata.

Visibly pushdown automata [3] are an important and well motivated subclass of pushdown automata, where pushdown operations are determined by the input symbol: the input alphabet is partitioned into three parts $A_c$, $A_r$ and $A_l$ for push, pop and local pushdown operations, respectively. This relates to function calls, for example. A function call is represented by push operation, local operations executed in the context of the called function are represented by local transitions, and, finally, the return from the function is represented by pop operation. The push, pop and local operations are sometimes referred as call, return and internal operations. Visibly pushdown automata are widely used, researched and known to be used in many practical applications, such as XML processing for example [1], [2], [5], [6], [7], [11].

It is well known that nondeterministic visibly pushdown automata can be determinised [3], [12]. These determinisations use principles that are also used in the well-known determinisation of finite automata: states of the equivalent deterministic automata are represented by so-called deterministic subsets [8]. Alur and Madhusudan [3] presented the proof of the determinisability of a given nondeterministic visibly pushdown automaton with $n$ states by creating a cartesian product consisting of all possible states and then creating deterministic subsets, which resulted in $2^{n^2+n}$ states of the deterministic version of the pushdown automaton. This was improved in [12], where the upper bound for the number of states was lowered from $2^{n^2+n}$ to $2^{n^2}$ and the upper bound for the number of pushdown store symbols was lowered from $|A_c|2^{n^2+n}$ to $|A_c|2^{n^2}$.

In this paper a new algorithm of the determinisation of nondeterministic visibly pushdown automata is presented. The algorithm improves the existing methods and can result in significantly smaller deterministic pushdown automata in many practical examples. Only necessary and accessible states and pushdown symbols of the deterministic pushdown automaton are computed and constructed during the determinisation, which is done by analysing which states are used in transitions on the same level of the nesting of pushdown operations and which pushdown store symbols can appear at the top of the pushdown store for each state.

The paper is organized as follows. Section 2 defines basic notions. Section 3 contains information on related works. Section 4 presents the new incremental algorithm for visibly pushdown automata determinisation. An example of the use of the presented algorithm is presented in Section 5. Finally, the conclusion of the paper is in Section 6.

This research has been supported the Czech Science Foundation as project No. 13-03253S.

978-83-60810-66-8/$25.00©2015, IEEE

## II. Basic notions

Basic notions are defined as in standard texts, such as [8].

### A. Alphabet, string

An *alphabet* $A$ is a finite nonempty set of *symbols*.

A *string* $s$ is a sequence of $n$ symbols $a_1a_2a_3\dots a_n$ from a given alphabet, where $n$ is the length of the string. A sequence of zero symbols is called empty string. Empty string is denoted by symbol $\varepsilon$ and its length is 0.

### B. Language

$A^*$ denotes the set of all strings over an alphabet $A$ including the empty string. Set $A^+$ is defined as $A^+ = A^* \setminus \{\varepsilon\}$. A *language* $L$ over an alphabet $A$ is a set $L \in A^*$. Similarly, for string $x \in A^*$, symbol $x^m$, $m \geq 0$, denotes the $m$-fold concatenation of $x$ with $x^0 = \varepsilon$. Set $x^*$ is defined as $x^* = \{x^m : m \geq 0\}$ and $x^+ = x^* \setminus \{\varepsilon\} = \{x^m : m \geq 1\}$.

### C. Pushdown automata

A *nondeterministic pushdown automaton* (nondeterministic PDA) is a seven-tuple $M = (Q, A, G, \delta, q_0, Z_0, F)$, where $Q$ is a finite set of *states*, $A$ is an *input alphabet*, $G$ is a *pushdown store alphabet*, $\delta$ is a mapping from $Q \times (A \cup \{\varepsilon\}) \times G^*$ into a set of finite subsets of $Q \times G^*$, $q_0 \in Q$ is an initial state, $Z_0 \in G$ is the initial pushdown store symbol, and $F \subseteq Q$ is the set of final (accepting) states.

Triplet $(q, w, x) \in Q \times A^* \times G^*$ denotes the configuration of a pushdown automaton. Top of the pushdown store $x$ is written on its left hand side. The initial configuration of a pushdown automaton is a triple $(q_0, w, Z_0)$ for the input string $w \in A^*$.

The relation $\vdash_M \subset (Q \times A^* \times G^*) \times (Q \times A^* \times G^*)$ is a *transition* of a pushdown automaton $M$. It holds that $(q, aw, \alpha z) \vdash_M (p, w, \beta z)$ if $(p, \beta) \in \delta(q, a, \alpha)$, where $z, \alpha, \beta \in G^*$. The $k$-th power, transitive closure, and transitive and reflexive closure of the relation $\vdash_M$ is denoted $\vdash_M^k$, $\vdash_M^+$, $\vdash_M^*$, respectively.

A pushdown automaton $M$ is a *deterministic* pushdown automaton (deterministic PDA), if it holds:

$$\forall q \in Q, \forall a \in A, \forall z \in G, \delta(q, \varepsilon, z) = \emptyset : |\delta(q, a, z)| \leq 1,$$
$$\forall a \in A, |\delta(q, \varepsilon, z)| \geq 1 : \delta(q, a, z) = \emptyset.$$

### D. Visibly pushdown automata

Visibly pushdown automata are defined as in [3], [12].

Let $A = A_c \cup A_r \cup A_l$ be a partition of $A$. The intuition behind the partition is: $A_c$ is the finite set of call (push) symbols, $A_r$ is the finite set of return (pop) symbols, and $A_l$ is the finite set of local symbols.

A *visibly pushdown automaton (VPA)* $M$ over $A$ is a seven-tuple $(Q, A, G, \delta, Q_0, \perp, F)$, where $Q$ is a finite set of *states*, $A = A_c \cup A_r \cup A_l$, $G$ is a finite *pushdown store alphabet*, a special symbol $\perp \in G$ represents the bottom-of-pushdown-store, which can be popped from the pushdown store unlimited number of times, $\delta = \delta_c \cup \delta_r \cup \delta_i$ is the transition mapping, where $(Q, G \setminus \{\perp\}) \in \delta_c(Q, A_c, \varepsilon)$, $(Q, \varepsilon) \in \delta_r(Q, A_r, G)$, and $(Q, \varepsilon) \in \delta_i(Q, A_l, \varepsilon)$, $Q_0 \subseteq Q$ is a set of initial states, $\perp \in G$ is initial pushdown store symbol, and $F \subseteq Q$ is a set of final (accepting) states.

## III. Related works

Visibly pushdown automata were introduced in [3]. Moreover, it was shown that any nondeterministic visibly pushdown automaton can be transformed into an equivalent deterministic one. The determinisation principle is similar to the the determinisation principle of finite automata [8].

In [3], states of the resulting deterministic visibly pushdown automaton consist of two components $(S, R)$. Component $R \in \mathcal{P}(Q)$ is an element of powerset of the states of the original automaton. Component $S = \mathcal{P}(Q \times Q)$ is a powerset of pairs of states of the original nondeterministic pushdown automaton that keeps tracking beginning states on path from push transitions to all states listed in $R$ component. We note that, given the union of states in second parts of pairs in $S$ component is equal to $R$ component, the $R$ component can be omitted but for keeping the automata hierarchy simple we maintain this $R$ component in the following definition as a connection to finite automata [12], where the states of the determinized automata correspond to the $R$ component.

Let $M = (Q, A, G, \delta, Q_0, \perp, F)$ be a nondeterministic VPA. For $A = A_c \cup A_r \cup A_l$, an equivalent deterministic VPA $M' = (Q', A, G', \delta', q'_0, \perp, F')$ can be constructed as follows: $Q' = 2^{Q \times Q} \times 2^Q$, $q'_0 = (Id_{Q_0}, Q_0)$ where $Id_X = \{(x, x) | x \in Q\}$, $F' = \{(S, R) | R \cap F \neq \emptyset\}$, $G' = 2^{Q \times Q} \times 2^Q \times A_c \cup \{\perp\}$, the transition relation $\delta' = \delta'_l \cup \delta'_c \cup \delta'_r$ is given by:

Local: For every
$l \in A_l, ((S', R'), \varepsilon) \in \delta'_l((S, R), l, \varepsilon)$ where

$$\begin{aligned} S' = \quad & \{(q, q') | \exists q'' \in Q : (q, q'') \in S, \\ & (q', \varepsilon) \in \delta_l(q'', l, \varepsilon)\}, \\ R' = \quad & \{q' | \exists q \in R : (q', \varepsilon) \in \delta_l(q, l, \varepsilon)\}. \end{aligned}$$

Push: For every
$c \in A_c, ((Id_{R'}, R'), (S, R, c)) \in \delta'_c((S, R), c, \varepsilon)$ where

$$R' = \quad \{q' | \exists q \in R : (q', \gamma) \in \delta_c(q, c, \varepsilon)\}.$$

Pop: For every $r \in A_r$,

- if the pushdown store is empty : $((S', R'), \varepsilon) \in \delta'_r((S, R), r, \perp)$ where $S' = \{(q, q') | \exists q'' \in Q : (q, q'') \in S, (q', \varepsilon) \in \delta_r(q'', r, \perp)\}$ and $R' = \{q' | \exists q \in R : (q', \varepsilon) \in \delta_r(q, r, \perp)\}$.
- otherwise: $((S'', R''), \varepsilon) \in \delta_r((S, R), r, (S', R', c))$, where

$$\begin{cases} R'' & = & \{q' | \exists q \in R' : (q, q') \in U\}, \\ S'' & = & \{(q, q') | \exists q_3 \in Q : (q, q_3) \in S', (q_3, q') \in U\}, \\ U & = & \{(q, q') | \exists q_1 \in Q, q_2 \in R : (q_1, q_2) \in S, \\ & & (q_1, \gamma) \in \delta_c(q, c, \varepsilon), \\ & & (q', \varepsilon) \in \delta_r(q_2, r, \gamma)\}. \end{cases}$$

The equivalent deterministic automaton has at most $2^{n^2+n}$ states and at most $|A_c|2^{n^2+n}$ pushdown store symbols. The size of the transition relation can be at most $|A_l|\left(2^{n^2+n}\right)^2 + |A_c|\left(2^{n^2+n}\right)^3 + |A_r||A_c|\left(2^{n^2+n}\right)^3$.

In 2009 an improved upper bound of the number of states has been found by Nguyen Van Tang [12]. In that paper two optimizations for Alur-Madhusudan's determinisation of visibly pushdown automata were introduced. First, the set of summaries $S$ component of a state pair for some special cases concerning initial states was minimized. Second, $R$ component of the state pair was removed. By removing $R$ component of determinised visibly pushdown automaton the upper bound for the number of states was lowered from $2^{n^2+n}$ to $2^{n^2}$ and there were at most $|A_c|2^{n^2}$ pushdown store symbols. The optimization is based on the observation that information stored in $R$ component of a state pair is already contained in $S$ component of the state pair [12]. However, that determinisation algorithm is still not practical. As pointed by Nguyen Van Tang in its implementation of visibly pushdown automata determinisation library, named VPAlib, the determinisation was performed in an exhaustive way. Therefore, their determinisation easily gets stuck with visibly pushdown automata of small size [12].

## IV. Determinisation algorithm

This section presents our new algorithm for the determinisation of nondeterministic visibly pushdown automata. The algorithm improves the original determinisation algorithm [3], [12]. As stated in the Introduction our algorithm computes and constructs only necessary and accessible states and pushdown symbols of the deterministic pushdown automaton. The basic idea of this improvement is analysing and tracking pushdown symbols that can appear on the top of the pushdown store for a particular state. With this information the explored pop transitions for the state can be reduced to only those that correspond to the possible pushdown store top symbols. Also, it is shown below that this information on the possible pushdown store top symbols for a state has certain interesting properties, which can be exploited for an effective way of calculation and storing this information for all states in the automaton.

We will use $\mathcal{T}_q$ to denote pushdown store top symbols. The *pushdown store top symbols* of state $q$, $\mathcal{T}_q \subseteq G'$ are the set of all pushdown store symbols that could appear at the top of the pushdown store for a state $q \in Q'$.

We will also use symbol $\lambda$ for a local connection: State $q'$ is locally connected to $q''$ if there is a sequence of transitions from state $q''$ to state $q'$, the pushdown store depth in both states is the same and the pushdown store depth for all other states along the sequence of transitions is greater than the pushdown store depth in $q''$. This relation between two states is not symmetric but is transitive. See Figure 1 shows an example of various local connections. Note that for example the path $1 \rightarrow 7 \rightarrow 8$ is not a local connection. The notation $a|\alpha \mapsto \beta$ denotes a transition that reads symbol $a \in A$ and replaces $\alpha \in G'^*$ with $\beta \in G'^*$ on the top of the pushdown store. This notation will be used in figures throughout the paper.

It can be easily seen that the pushdown store top symbols are shared between locally connected states.

With local connection from $q'$ to $q$ and local connection from $q$ to $q''$, the $q'$ is locally connected to $q''$ by transitive closure. If $\mathcal{T}_q$ is the set of pushdown store top symbols of state $q$, then $\mathcal{T}_{q'} \subseteq T_q$ and $\mathcal{T}_q \subseteq \mathcal{T}_{q''}$ and also $\mathcal{T}_{q'} \subseteq \mathcal{T}_{q''}$.

The local closure of state $q$ is $\lambda^*(q)$. See Figure 2 as an example of a local closure. See Figure 1 and note that the path $1 \rightarrow 7 \rightarrow 8$ mentioned before is in fact a local closure.

Closing all states under the $\lambda^*(q)$ connects all $\mathcal{T}_q$. See Figure 3.

We define these notions formally:

*Definition 1:* A local connection $\lambda(q)$ of state $q$. Given a deterministic visibly pushdown automaton $M' = (Q', A, G', \delta', q'_0, \bot, F')$, where $A = A_c \cup A_r \cup A_l$, states $q, q' \in Q'$, then $\lambda(q) = \{q' : (q, uw, \gamma_0) \vdash^k (q', w, \gamma_k), u \in A^*, w \in A^*, \gamma_0, \gamma_k \in G'^*, 1 \leq k, |\gamma_0| = |\gamma_k|, |\gamma_0| < |\gamma_i|, 1 \leq i < k\}$.

*Definition 2:* A local connection closure $\lambda^*(q)$ of state $q$. The local connection closure $\lambda^*$ is defined by these equalities:

$$\lambda^0(q) = q, \tag{1}$$

$$\lambda^{i+1}(q) = \bigcup_{\forall q' \in \lambda^i(q)} \lambda(q'), \tag{2}$$

$$\lambda^*(q) = \bigcup_{i \geq 0} \lambda^i(q). \tag{3}$$

*Definition 3:* A set of pushdown store top symbols $\mathcal{T}_q$ of state $q$. Given a deterministic visibly pushdown automaton $M' = (Q', A, G', \delta', q'_0, \bot, F')$, where $A = A_c \cup A_r \cup A_l$, states $q, q', q'' \in Q'$, then $T_q = \{g : (q', g) \in \delta(q'', c, \varepsilon), c \in A_c, g \in G', q' \in \lambda^*(q)\}$.

Due to convenient properties of $\lambda^*(q)$, $\mathcal{T}_q$ can be stored in a space optimal way. Given $q, q' \in Q'$, then $\forall q' \in \lambda^*(q)$ holds that $\mathcal{T}_{q'} \subseteq \mathcal{T}_q$, ie parts of $T_q$ can be shared between locally connected and locally closed states. See Figure 4 as an example of this process.

Further, we show by induction on the length of an input sentence that pushdown store top symbols are given by the $\lambda^*(q)$ and states that are source and target of the appropriate push and pop transition.

Step one:

$$\varepsilon : \bot \in \mathcal{T}_{q'_0}. \tag{4}$$

Then, assume that pushdown store top symbols are given by the local closure for first $i$ symbols of input word. Symbol $a$ is a $(i+1)$-th symbol of input word. Given $q_1, q_2, q_3, s_1, s_2, s_3 \in Q', a \in A_c$, then we have three distinct cases:

$$\lambda(r_1) = \{q_1 : (q_1, aw, \gamma) \vdash (r_1, w, \gamma)\}, \tag{5}$$

$$\mathcal{T}_{r_2} = \{(q_2, a) : (q_2, aw, \gamma) \vdash (r_2, w, (q_2, a)\gamma)\}, \tag{6}$$

$$\lambda(r_3) = \{q_2 : (q_3, aw, (q_2, a)\gamma) \vdash (r_3, w, \gamma)\}. \tag{7}$$

Notice that $\gamma$ does not change between states $q_1$ and $r_1$. Pushdown store top symbol $(q_2, s)$ was pushed in state $q_2$ and popped in state $q_3$ so $\gamma$ does not change between states $q_2$ and $r_3$ either. Pushdown store top symbols are given by the $\lambda^*(q)$ and states that are source and target of appropriate push and pop transition for first $i+1$ symbols of input word. The induction holds for $i+1$.

More informally: The deterministic automaton is constructed from the initial state. The deterministic subset of the

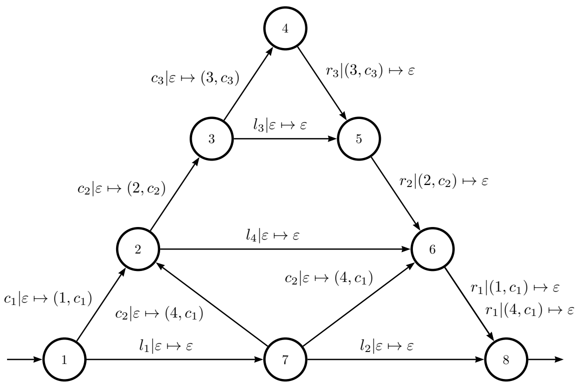

4
$c_3|\varepsilon \mapsto (3, c_3)$
$r_3|(3, c_3) \mapsto \varepsilon$
3
$l_3|\varepsilon \mapsto \varepsilon$
5
$c_2|\varepsilon \mapsto (2, c_2)$
$r_2|(2, c_2) \mapsto \varepsilon$
2
$l_4|\varepsilon \mapsto \varepsilon$
6
$c_2|\varepsilon \mapsto (4, c_1)$
$r_1|(1, c_1) \mapsto \varepsilon$
$c_1|\varepsilon \mapsto (1, c_1)$
$c_2|\varepsilon \mapsto (4, c_1)$
$r_1|(4, c_1) \mapsto \varepsilon$
1
$l_1|\varepsilon \mapsto \varepsilon$
7
$l_2|\varepsilon \mapsto \varepsilon$
8

Fig. 1. Various $\lambda$ relations of state 8 to state 1.

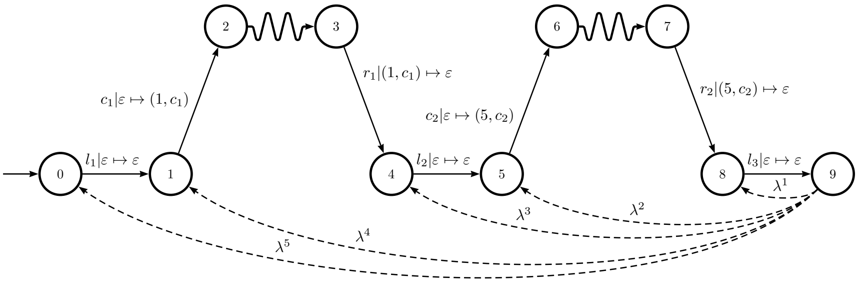

2
3
6
7
$r_1|(1, c_1) \mapsto \varepsilon$
$c_1|\varepsilon \mapsto (1, c_1)$
$r_2|(5, c_2) \mapsto \varepsilon$
$c_2|\varepsilon \mapsto (5, c_2)$
0
$l_1|\varepsilon \mapsto \varepsilon$
1
4
$l_2|\varepsilon \mapsto \varepsilon$
5
8
$l_3|\varepsilon \mapsto \varepsilon$
9
$\lambda^1$
$\lambda^3$
$\lambda^2$
$\lambda^5$
$\lambda^4$

Fig. 2. The $\lambda^*$ relation computed for state 9.

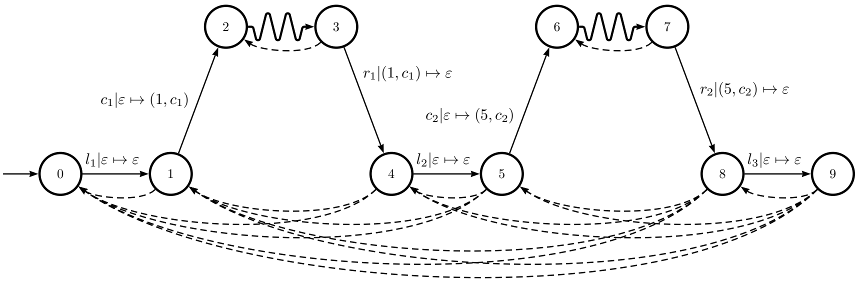

2
3
6
7
$r_1|(1, c_1) \mapsto \varepsilon$
$c_1|\varepsilon \mapsto (1, c_1)$
$r_2|(5, c_2) \mapsto \varepsilon$
$c_2|\varepsilon \mapsto (5, c_2)$
0
$l_1|\varepsilon \mapsto \varepsilon$
1
4
$l_2|\varepsilon \mapsto \varepsilon$
5
8
$l_3|\varepsilon \mapsto \varepsilon$
9

Fig. 3. The $\lambda^*$ relation computed for all states.

initial state is created from all initial states of the original automaton. Initial pushdown store symbol $\bot \in \mathcal{T}_{q'_0}$ forms the pushdown store top symbols set.

In every iteration, all local and push transitions are explored for the known states. Base set of possible pushdown store top symbols $\mathcal{T}_q$ of the pushdown store for given state $q \in Q'$ is given by push transitions. Then, we track pushdown store top symbols for each state.

Any two states $q$, $q'$, where a local transition exists from state $q'$ to state $q$, share part of $\mathcal{T}_{q'}$ in form of that everything from $\mathcal{T}_{q'}$ is in $\mathcal{T}_q$.

Any two states $q$, $q'$, where a pop transition popping symbol $(q', r)$ exists from state $x$ to state $q$, share part of $\mathcal{T}_{q'}$ in form of that everything from $\mathcal{T}_{q'}$ is in $\mathcal{T}_q$. Given $q, q', q'' \in Q', l \in A_l, c \in A_c, r \in A_r$, then for $\mathcal{T}$ the following properties hold:

$$\bot \in \mathcal{T}_{q'_0}, \tag{8}$$

$$\forall (q, \varepsilon) \in \delta(q', l, \varepsilon) \Rightarrow \mathcal{T}_{q'} \subseteq \mathcal{T}_q, \tag{9}$$

$$\forall (q, \varepsilon) \in \delta(q'', r, (q', c)) \Rightarrow \mathcal{T}_{q'} \subseteq \mathcal{T}_q, \tag{10}$$

$$\forall (q, (q'', c)) \in \delta(q'', c, \varepsilon) \Rightarrow (q'', c) \subseteq \mathcal{T}_q. \tag{11}$$

The algorithm of the determinisation is formally described by Algorithm 1. Given $q, q'' \in Q'$, its main part can be written in an abstract way as follows:

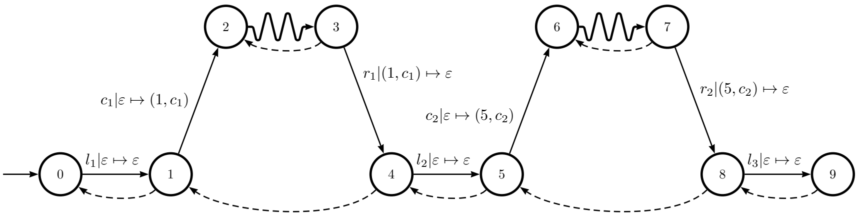

2
3
6
7
$c_1|\varepsilon \mapsto (1, c_1)$
$r_1|(1, c_1) \mapsto \varepsilon$
$c_2|\varepsilon \mapsto (5, c_2)$
$r_2|(5, c_2) \mapsto \varepsilon$
0
$l_1|\varepsilon \mapsto \varepsilon$
1
4
$l_2|\varepsilon \mapsto \varepsilon$
5
8
$l_3|\varepsilon \mapsto \varepsilon$
9

Fig. 4. Connecting $T_q$ between locally connected states.

1) create the initial state,
2) while $\nexists(q', \varepsilon) \in \delta(q, r, \gamma)$, where $r \in A_r$ and $\gamma \in \mathcal{T}_q$, do create pop transition $(q, r, \gamma)$,
   while $\nexists(q', \varepsilon) \in \delta(q, l, \varepsilon)$, where $l \in A_l$, do create local transition $(q, l, \varepsilon)$,
   while $\nexists(q', \gamma) \in \delta(q, c, \varepsilon)$, where $c \in A_c$, do create push transition $(q, c, \varepsilon)$,
3) set final states.

Only a pair of states $(q', q)$, where $q' \in \lambda^*(q)$, can appear as an element in $S$ component. As it is described in II, the new pairs of $S$ component are only created from push and local transitions. The push transitions only yield elements of $S$ component based on identity. On the other hand, the local transitions yield exactly the pairs that conform to local closure, because they connect appropriate targets of push and sources of pop operations.

Let $L$ be the set of all distinct pairs of locally closed states of the nondeterministic automaton. Then, $2^{|L|}$ is the maximum number of states of the deterministic automaton. The $|L|$ is at most $n^2$ when all states in the nondeterministic automaton are locally closed.

## V. Example

In this section the determinisation of a simple nondeterministic visibly pushdown automaton is demonstrated. For the sake of clarity, the component $R$ of states and pushdown store symbols is omitted, because the information it holds is already contained in the component $S$ (it is the second value of each pair), as it was described in [12].

*Example 5.1:* The nonderterministic visibly pushdown automaton $a_1$ is shown in Figure 6.

Let $a_1 = (Q, A_c \cup A_r \cup A_l, G, \delta, Q_0, \bot, F)$ be a nondeterministic visibly pushdown automaton. An equivalent deterministic visibly pushdown automaton $d_1 = (Q', A_c \cup A_r \cup A_l, G', \delta', q_0', \bot, F')$ can be constructed as follows.

The initial state is constructed as powerset of all identity pairs of initial states of automaton $a_1$, so $q_0' = \{(0, 0)\}$.

The push and the local transitions can be easily deduced from determinisation rules from Section IV. All transitions are shown in Figure 5.

In each push transition, the pushdown store top symbol is tracked for target state of the push transition. When a local transition occurs, all pushdown store top symbols are shared from a source state of the local transition to a target state of the local transition and all other locally connected states. The tracking of all locally connected states could be achieved by creating virtual transitions serving as a transitive closure.

Then, the pop transitions are created based on known input symbols (from the nondeterministic automaton) and tracked pushdown store top symbols. The transitions are created according to the determinisation rules from Section IV.

For an illustration of the determinisation see Figure 5. The pushdown store top symbols are tracked as follows. Push, local and pop transitions are marked green, gray and red, respectively. Arrows describe movements of tracked tops of the pushdown store. Dashed arrows represents source of top pushdown store symbols that are shared with target state when local transition occur.

The resultant deterministic PDA $d_1$ is shown in Figure 7.

The resultant deterministic PDA can be also reproduced by running Algorithm 1.

Given the nondeterministic pushdown automaton from the example above, the VPAlib algorithm [9] constructs a deterministic pushdown automaton with 45 states and 1206 transitions. We note that 45 (states) is not a power of 2, which is caused by the fact that the implementation of VPAlib library does not consider states in which components $R$ or $S$ are empty sets (in this way, it performs another optimization of the determinisation algorithm [12]). For comparison, our algorithm constructs an equivalent deterministic pushdown automaton with only 3 states and 8 transitions. This is a significant improvement over the previously existing determinisation algorithms.

## VI. Conclusion

A new incremental algorithm of the determinisation of nondeterministic visibly pushdown automata has been described. The algorithm creates only necessary states and pushdown symbols by analysing and tracking which states are achievable by computing transitions on the same levels of pushdown operations nesting. Possible tops of the pushdown store are stored for each state when a pop transition is in progress and then they are shared through local transitions with states on the same levels of the nesting. The behavior of the algorithm

```
Determinisation algorithm.
Input  : A nondeterministic visibly pushdown automaton $M_n(Q, A_c \cup A_l \cup A_r, G, \delta, Q_0, \perp, F)$.
Output: An equivalent deterministic visibly pushdown automaton $M_d(Q', A_c \cup A_l \cup A_r, G', \delta', q_0', \perp, F')$.
Method:
set $q_0' = \{(x, x) : x \in Q_0\}, Q' = \{q_0'\}, G' = \emptyset, \delta' = \emptyset, F' = \emptyset$;
create queue $Dirty = ((q_0', \perp))$, // the set of pairs (state, pushdown store symbol);
create set $Clean = \emptyset$, // the set of states;
while $Dirty\ is\ not\ empty$ do
    $(state, symbol)$ = dequeue $Dirty$;
    create set $States = \emptyset$;
    if $state \notin Clean$ then
        forall the $l \in A_l$ do
            forall the $(q, q_2) \in state$ do
                $state2$ = $\{(q, q_1) : (q_1, \varepsilon) \in \delta(q_2, l, \varepsilon)\}$; if $state2 = \emptyset$ then continue;
                add $state2$ to $Q'$; add $(state2, \varepsilon)$ to $\delta'(state, l, \varepsilon)$; add $state2$ to $States$;
                update $T_x, x \in \lambda^*(state2)$;
            end
        end
        forall the $c \in A_c$ do
            forall the $(q, q_2) \in state$ do
                $state2$ = $\{(q_1, q_1) : (q_1, g) \in \delta(q_2, c, \varepsilon)\}$; if $state2 = \emptyset$ then continue;
                add $state2$ to $Q'$; add $(state, c)$ to $G'$;
                add $(state2, (state, c))$ to $\delta'(state, c, \varepsilon)$; add $state2$ to $States$;
                update $T_x, x \in \lambda^*(state2)$;
            end
        end
        add $state$ to $Clean$;
    end
    forall the $r \in A_r$ do
        if $symbol\ is\ \perp$ then
            forall the $(q, q_2) \in state$ do
                $state2$ = $\{(q, q_1) : (q_1, \varepsilon) \in \delta(q_2, r, \perp)\}$; if $state2 = \emptyset$ then continue;
                add $state2$ to $Q'$; add $(state2, (state, \varepsilon))$ to $\delta'(state, r, symbol)$; add $state2$ to $States$;
                update $T_x, x \in \lambda^*(state2)$;
            end
        else
            create set $Update = \emptyset$;
            forall the $(q_1, q_2) \in state$ do
                $pairs$ = $\{(q, q_I) : (q_I, g) \in \delta(q_2, r, g),\ (q_1, \varepsilon) \in \delta(q, c, g),\ c = second(symbol)\}$;
                add $pairs$ to $Update$;
            end
            forall the $(q, q_3) \in first(source)$ do
                $state2$ = $\{(q, q_2) : (q_3, q_2) \in Update\}$; if $state2 = \emptyset$ then continue;
                add $state2$ to $Q'$; add $(state2, (state, \varepsilon))$ to $\delta'(state, r, symbol)$; add $state2$ to $States$;
                update $T_x, x \in \lambda^*(state2)$;
            end
        end
    end
    forall the $s \in States$ do
        forall the $g \in T_{state} \setminus T_s$ do  enqueue $(s, g)$ to $Dirty$ ;
    end
end
forall the $q \in Q'$ do
    if $exists\ (q_1, q_2) \in q\ such\ that\ q_2 \in F$ then  add $q$ into $F'$ ;
end
```

**Algorithm 1:** Determinisation

Fig. 5. Construction of deterministic automaton $d_1$.

| | **State** $q$ | $\{(0,0)\}$ | $\{(0,0),(1,1)\}$ | $\{(0,1)\}$ |
|---|---|---|---|---|
| | **Pushdown store**<br>**top symbols** $T_q$ | $\bot$ | $\bot$, $(a,\{(0,0)\})$,<br>$(a,\{(0,0)\})$, $(a,\{(0,0),(1,1)\})$ | $(a,\{(0,0)\})$,<br>$(a,\{(0,0),(1,1)\})$ |
| 1. | $a\vert\varepsilon \mapsto (a,\{(0,0)\})$ | $\{(0,0),(1,1)\}$ | | |
| 2. | $c\vert\varepsilon \mapsto \varepsilon$ | $\{(0,1)\}$ | $\{(0,1)\}$ | |
| 3. | $a\vert\varepsilon \mapsto (a,\{(0,0),(1,1)\})$ | | $\{(0,0),(1,1)\}$ | |
| 4. | $d\vert(a,\{(0,0),(1,1)\}) \mapsto \varepsilon$ | | $\{(0,1)\}$ | |
| 5. | $d\vert(a,\{(0,0)\}) \mapsto \varepsilon$ | | $\{(0,1)\}$ | |
| 4. | $b\vert(a,\{(0,0),(1,1)\}) \mapsto \varepsilon$ | | | $\{(0,1)\}$ |
| 6. | $b\vert(a,\{(0,0)\}) \mapsto \varepsilon$ | | | $\{(0,1)\}$ |

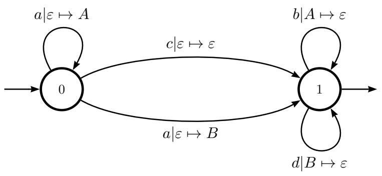

$a|\varepsilon \mapsto A$
$b|A \mapsto \varepsilon$
$c|\varepsilon \mapsto \varepsilon$
0
1
$a|\varepsilon \mapsto B$
$d|B \mapsto \varepsilon$

Fig. 6. The nondeterministic visibly pushdown automaton $a_1$ from Example 5.1.

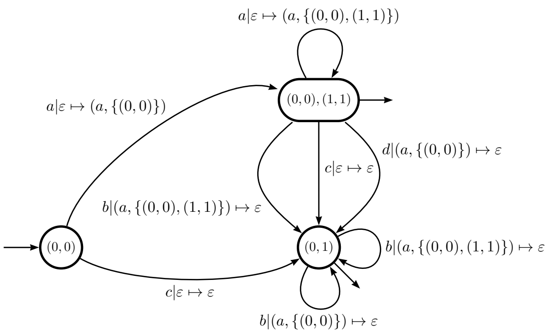

$a|\varepsilon \mapsto (a,\{(0,0),(1,1)\})$
$a|\varepsilon \mapsto (a,\{(0,0)\})$
(0,0),(1,1)
$d|(a,\{(0,0)\}) \mapsto \varepsilon$
$c|\varepsilon \mapsto \varepsilon$
$b|(a,\{(0,0),(1,1)\}) \mapsto \varepsilon$
(0,0)
(0,1)
$b|(a,\{(0,0),(1,1)\}) \mapsto \varepsilon$
$c|\varepsilon \mapsto \varepsilon$
$b|(a,\{(0,0)\}) \mapsto \varepsilon$

Fig. 7. The resultant deterministic PDA $d_1$ from Example 5.1 created by determinisation of automaton $a_1$.

is inspired by the behavior of the incremental construction of the deterministic finite automaton.

The algorithm has been implemented as a part of an experimental automata library [13].

Although the number of states of the deterministic automaton for the worst case is still $2^{n^2}$ for a given nondeterministic visibly pushdown automaton with $n$ states, it has been shown that the upper bound of the number of states of the deterministic automaton is dependant on the number of distinct pairs of locally connected states. For those and other reasons in many practical cases our algorithm provides significantly smaller deterministic automata than the previously existing determinisation algorithms.

Furthermore, a similar approach can be adapted for the determinisation of Height-deterministic pushdown automata [10].

Further information can be found on [4].

## References

[1] R. Alur. Marrying words and trees. In J. Meseguer and G. Rosu, editors, *Algebraic Methodology and Software Technology, 12th International Conference, AMAST 2008, Urbana, IL, USA, July 28-31, 2008, Proceedings*, volume 5140 of *Lecture Notes in Computer Science*, page 1. Springer, 2008.

[2] R. Alur, V. Kumar, P. Madhusudan, and M. Viswanathan. Congruences for visibly pushdown languages. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, editors, *Automata, Languages and Programming, 32nd International Colloquium, ICALP 2005, Lisbon, Portugal, July 11-15, 2005, Proceedings*, volume 3580 of *Lecture Notes in Computer Science*, pages 1102–1114. Springer, 2005.

[3] R. Alur and P. Madhusudan. Visibly pushdown languages. In L. Babai, editor, *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 202–211. ACM, 2004.

[4] Arbology www pages. Available on: http://arbology.fit.cvut.cz/, 2015. May 2015.

[5] V. Bárány, C. Löding, and O. Serre. Regularity problems for visibly pushdown languages. In B. Durand and W. Thomas, editors, *STACS 2006, 23rd Annual Symposium on Theoretical Aspects of Computer Science, Marseille, France, February 23-25, 2006, Proceedings*, volume 3884 of *Lecture Notes in Computer Science*, pages 420–431. Springer, 2006.

[6] D. Debarbieux, O. Gauwin, J. Niehren, T. Sebastian, and M. Zergaoui. Early nested word automata for xpath query answering on XML streams. *Theor. Comput. Sci.*, 578:100–125, 2015.

[7] E. Filiot, J. Raskin, P. Reynier, F. Servais, and J. Talbot. Properties of visibly pushdown transducers. In P. Hlinený and A. Kucera, editors, *Mathematical Foundations of Computer Science 2010, 35th International Symposium, MFCS 2010, Brno, Czech Republic, August 23-27, 2010. Proceedings*, volume 6281 of *Lecture Notes in Computer Science*, pages 355–367. Springer, 2010.

[8] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to automata theory, languages and computation*. Addison-Wesley, second edition, 2001.

[9] H. Nguyen. Visibly pushdown automata library. Available on: http://www.emn.fr/z-info/hnguyen/vpa/, 2015. May 2015.

[10] D. Nowotka and J. Srba. Height-deterministic pushdown automata. In L. Kučera and A. Kučera, editors, *Mathematical Foundations of Computer Science 2007, 32nd International Symposium, MFCS 2007, Český Krumlov, Czech Republic, August 26-31, 2007, Proceedings*, volume 4708 of *Lecture Notes in Computer Science*, pages 125–134. Springer, 2007.

[11] J. Srba. Visibly pushdown automata: From language equivalence to simulation and bisimulation. In Z. Ésik, editor, *Computer Science Logic, 20th International Workshop, CSL 2006, 15th Annual Conference of the EACSL, Szeged, Hungary, September 25-29, 2006, Proceedings*, volume

4207 of *Lecture Notes in Computer Science*, pages 89–103. Springer, 2006.

[12] N. V. Tang. A tighter bound for the determinization of visibly pushdown automata. In A. Legay, editor, *Proceedings International Workshop on Verification of Infinite-State Systems, INFINITY 2009, Bologna, Italy, 31th August 2009.*, volume 10 of *EPTCS*, pages 62–76, 2009.

[13] J. Trávníček. ALT: Automata (Algorithms) Library Toolkit, 2015.

DOI: 10.15439/2015F289
ACSIS, Vol. 5

# A Review of Source Code Projections in Integrated Development Environments

Ján Juhár and Liberios Vokorokos
Department of Computers and Informatics
Technical University of Košice
Letná 9, 0420 00 Košice, Slovakia
Email: {jan.juhar, liberios.vokorokos}@tuke.sk

***Abstract*—The term Projectional editor is commonly used for tools that can work directly with the program's abstract syntax tree. They are able to provide different views of the program, according to the specific editor used. The ability to look at the program from multiple views is often requested as a mean to simplify program comprehension. During their evolution, the Integrated Development Environments were equipped with tools that provide such possibilities. Many of them already work with the parsed abstract syntax tree of the code and thus can be considered for projections. In this paper we review projections available in 6 widely used IDEs. The review categorizes existing projections and shows that significant number of IDE tools depend on the knowledge of program structure, but also that data from other integrated tools are used to enhance the projections.**

## I. Introduction

THE ability to evolve a software system depends on the programmer's ability to comprehend its source code. The source code comprehension (also program comprehension) is a cognitive process that involves analysis of the source code and retrieval of information and knowledge about the analyzed system. This process tends to take up to a half of a programmer's time during the software development and maintenance [1], [2].

The main hindrance that programmers face while comprehending the source code is the wide semantic gap that exists between the problem domain and the solution domain. Due to this gap, the mapping of a specific program feature to the code that implements this feature (or vice versa) is not straightforward. Factors like programmer's personality, experiences and skills, combined with the nature of the task at hand have a strong influence on this process [2]. Program represented as a source code enforces the single structure chosen by its author onto all programmers that will work with it later, regardless of whether they will intend to reuse, extend or modify it. This is the reason why the tools that are able to provide different, context-driven views of the code are requested during the comprehension process [3].

The goal to aid programmers in the program comprehension process stands behind many tools. The Integrated Development Environments (IDEs) represent the most complex toolset for working with a source code. The ultimate purpose of an IDE and all of its tools is to cover all phases of the software life cycle and to increase the productivity of programmers. For this reason, an IDE typically contains code editors, browsers and analyzers, refactoring and build automation tools, debuggers, and other tools. The research on the comprehension strategies of professional programmers by Maalej et al. [2] points out the importance of integrated environments: comprehension tools that are not a part of the IDEs are virtually not used at all in the practice.

The main advantage modern IDEs have over pure text editors (even the advanced ones) is that they can "understand" the structure of the source code. After loading the source code contained in text files, they parse it into an abstract representation of the code – a form of its abstract syntax tree (AST). Many operations that IDE performs, e.g., refactoring or contextual code completion, are performed against this AST [4]. Whenever the code or the AST changes, the other is accordingly updated to stay in sync.

There are IDEs that take this idea to the next level and use an AST as a base program representation. The program is stored in the files as a serialized form of this AST, e.g., using XML notation. After loading such file into the editor, a *projection* of this base representation is presented to the programmer. There is no need for parsing and when the programmer is editing the program, he or she is basically directly editing the AST [5]. Such base representation abstracted from concrete language notation can be presented through projections in multiple forms and consequently the notation that programmer deals with can be tailored to best suit a particular domain. Editors of such IDEs are therefore called *projectional editors* [4], [5]. The most notable example of such IDE is the JetBrains Meta Programming System (MPS) which is used not only to develop programs in domain specific languages, but also to create these languages along with the appropriate language notations and editors.

In the case of MPS-like tools the idea of projections is well-defined. It is the core of their functionality. As we already hinted, modern source-based IDEs also operate on the abstract representation of the code, even though it needs to be parsed from the textual notation and it exists only during code editing. Yet, this gives an opportunity to exploit the idea of projections even by these IDEs.

We believe that source code projections are used by source-based IDEs on a large scale. However, the term *projection* is not well established in their context. This may be the reason why tools that use them are seldom recognized or referred to

978-83-60810-66-8/$25.00©2015, IEEE

as projectional. Our goal in this paper is to identify, review and categorize projections that are used by popular IDEs.

## II. Integrated Development Environments and Projections

There are many tools that programmers can use during development for code editing. They range from pure text editors that provide the basic text manipulation operations, through the advanced ones that add support for syntax highlighting or word completion, to fully featured IDEs with all the tools mentioned in the previous section. However, there is no generally recognized definition of what exactly is an IDE. Due to a great number of existing tools and an effort to differentiate themselves, there is no clear boundary between what is still "only" a text editor and what is already an IDE.

For our purposes, we have already laid out some requirements on what we consider for an IDE, mostly with regard to projections we want to explore. We will understand an integrated development environment as an application that is designed to support the majority of the software development life cycle – that is, at least the implementation, testing and maintenance phases – by providing an environment for programmers that includes tools for the associated tasks. In addition, we will require that the included tools (not necessarily all of them) are able to take advantage of the program structure – they should operate on the abstract representation of the source code.

As for the term *projection*, in the context of the MPS-like projectional editors it is used to represent an interactive, customizable rendering of the AST [5]. This can be true even for the source-based IDEs that operate on a parsed AST, though possibilities are limited here by the need to preserve parsability of the base source code. Through projections, source code can be conveyed in multiple views. The source code projections can be thus considered for a mapping between a set of base source code structures and a set of dynamically created views [6]. Such views usually focus on some aspect of the system and convey it in a concise way. Different projections can be useful in different contexts, but generally they can have positive effects on program comprehension by providing a higher-level view of the system.

## III. A Review of Source Code Projections

In order to select which IDEs will be used for the review of existing code projections, we referred to their popularity ranking *Top IDE Index*[1] that was compiled from Google search trends as of April, 2015. We focused on the first ten ranks of the index, as listed in the table I.

However, the criteria according to which the tools were added to the index were less strict than our understanding of an IDE. As a result, the index also contains tools like Vim, Emacs and SublimeText. Although these are very flexible and extensible tools for programming, we consider them for code editors, because they work with the source code only at the textual level. Thus, these were ruled out of the review. Additionally, we had to exclude XCode as it is available only for Apple Mac platform, to which we did not have access. Six remaining IDEs marked in the "reviewed" column of the Table I were used in the following review of existing projections. The reviewed versions of used IDEs are also listed in the table.

TABLE I
Selection of IDEs for the review

| Rank[1] | Name | Category | Reviewed | Version |
|---|---|---|---|---|
| 1 | Eclipse | IDE | ✓ | 4.4.2 |
| 2 | Visual Studio | IDE | ✓ | 2013 Ultimate |
| 3 | Vim | Editor | – | – |
| 4 | NetBeans | IDE | ✓ | 8.0.2 |
| 5 | XCode | IDE | – (n/a) | – |
| 6 | SublimeText | Editor | – | – |
| 7 | Komodo | IDE | ✓ | 9.0 |
| 8 | IntelliJ IDEA | IDE | ✓ | 14.1 |
| 9 | Emacs | Editor | – | – |
| 10 | Xamarin | IDE | ✓ | 5.9 |

The review of projections was conducted by evaluating features of main application menu and code editor of each selected IDE for projectional properties. Within the single IDE, the features were checked for multiple languages and project types.

Four categories of projections were identified. In the following subsections we describe these categories and list the relevant projections along with their requirements.

### A. In-editor Projections

The code editor is a very significant part of each IDE. It was already mentioned that it projects editable representation of the AST constructed by parsing the source code. Of course, this projection will work only for languages that are supported by the IDE. Code editor displays the loaded file with exactly the same content as persisted on the storage medium. The projectional properties manifest themselves in the additional features that augment the text, and these are reviewed below.

The first is the *code highlighting*. It is able to convey information not only about the code syntax (e.g., by highlighting the keywords of the language), but also about its structure. It takes into account the scope of the variables – it distinguishes between local and global ones even if they have the same name. Furthermore, it can highlight all occurrences of the same program element (see fig. 1), identify unused statements and more. This is where the structural information provided by the parsed AST are exploited. However, the level to which the IDE is capable of these features depends on the detailness of the parsed AST and also on the language properties. The more a language has "dynamic"[2] properties, the less complete (or reliable) knowledge about actual program structure during

[1] http://pypl.github.io/IDE.html, accessed April 2015

[2] Dynamic languages are mostly defined by support for dynamic typing or powerful reflection that allows extensive run-time modifications of a program behavior.

runtime can be obtained from the static structure of the AST. These properties have influence on all projections that utilize the AST.

A related projection is the *error highlighting*, which brings another informational layer to the code editor. The simplest to detect are syntax errors. Strongly-typed languages can benefit from possibility to check type assignment and visualize the errors.

Another common in-editor projection is the *code completion*, visualized as a pop-up list of available program elements in the specific context. Yet again, the knowledge of structural relations of elements in the edited source code is required for realization of this feature and the level of list completeness depends on the AST detailness and language properties.

Apart from above described in-editor projections that are available in every selected IDE, there are some that exist only in one of them. IntelliJ IDEA is able to project method or class implementation in the pop-up window over the selected class or method name with action called *Quick Definition* (see fig. 1). It is a different form of the *go to definition* projection (described in section III-C) that causes less navigational overhead, as the pop-up can be easily closed and user does not need to switch to other editor tab or window. The projected code is, however, not editable. Visual Studio provides feature called *Peek definition*, which creates a projection with similar purpose. This one is inserted directly in-line within the editor, as can be seen in fig. 2, and is fully editable. The use of this projection recursively inside the in-lined view replaces the whole view and adds so-called “breadcrumbs” for switching among the already opened views.

Source code in the editor can be further augmented by metadata available in the AST. To give an example, the presence of documentation annotation `@deprecated` in a Java code can be projected by crossing out any usage of the annotated element in the editor to visually warn the programmer about usage of a deprecated API.

In-editor projections are also used to project information not obtained from the AST. While running the debugging session, the IntelliJ IDEA projects the actual values of the variables next to their occurrences in the code editor as the user steps through the program execution. *CodeLens*, a feature of Visual Studio, decorates declarations of methods with a reference counter, as displayed in fig. 2. In addition to this, it can show a unit tests passing score and a code changes history, if the data are available. Only the reference counter value can be obtained by analyzing the AST of the program. The later two pull required data from the integrated test runner and version control system, respectively.

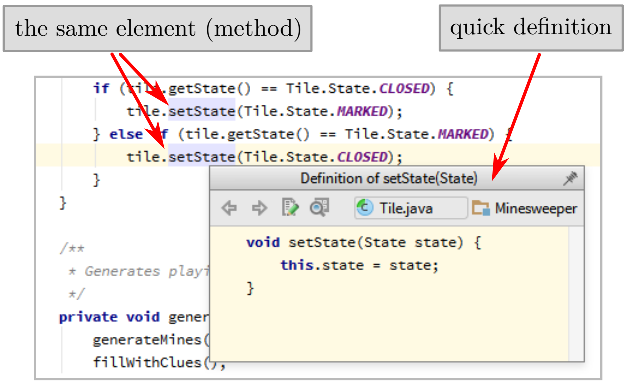
### B. Structure Projections

Each of the selected IDEs contains tools that provide overview of the project structure. The most basic one – the file browser – can be considered for an *identity* projection of the project’s files, with exactly the same structure as stored on a storage medium.

With the exception of the Komodo IDE, the selected IDEs contain panels that show hierarchical, tree-like project structure according to packages, namespaces, modules, classes, or other structural elements of the programs. These high-level views of the program structure are augmented with graphical symbols, informing the user about the visibility of the structural elements or distinguishing their type (class, abstract class, interface, and others).

The view of the same tool in a particular IDE differs depending on the used language or the project type. For example, in IntelliJ IDEA, *Project* panel shows more detailed structure in the case of Java files than, say, Python or JavaScript files. When working on a web project in the NetBeans IDE, the *Projects* panel extends the tree structure to include even the remote files that are linked from inside of the HTML pages.

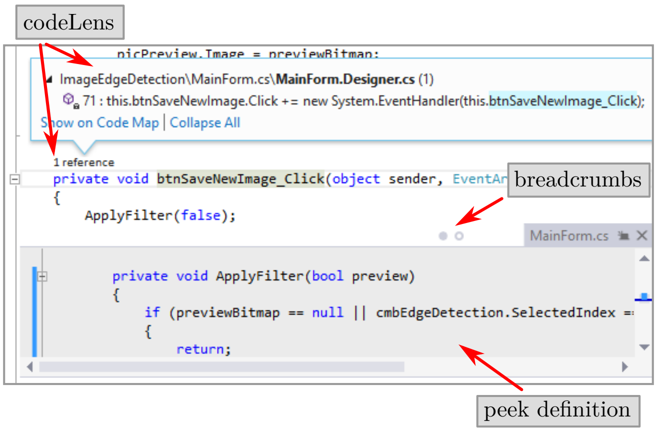
Individual IDEs also differ in the level of the details displayed by these tools. Particularly detailed is the Package Explorer of Java projects that is available in the Eclipse IDE. As shown in fig. 4, it goes down to the level of individual class members and also provides their signature. At the file level the view is further extended with repository information of the latest edit. Another tool that goes deeper into the program structure is the *Architecture Explorer* in Visual Studio. In addition to listing the class members, it can recursively show



```
    if (tile.getState() == Tile.State.CLOSED) {
        tile.setState(Tile.State.MARKED);
    } else if (tile.getState() == Tile.State.MARKED) {
        tile.setState(Tile.State.CLOSED);
    }
}

/**
 * Generates playi
 */
private void gener
    generateMines(
    fillWithClues();
```

```
void setState(State state) {
    this.state = state;
}
```

Fig. 1. code highlighting and *quick definition* in IntelliJ IDEA



```
        picPreview.Image = previewBitmap;
```

ImageEdgeDetection\MainForm.cs\**MainForm.Designer.cs** (1)
71 : this.btnSaveNewImage.Click += new System.EventHandler(this.btnSaveNewImage_Click);

```
1 reference
private void btnSaveNewImage_Click(object sender, EventAr
{
    ApplyFilter(false);
```

```
        private void ApplyFilter(bool preview)
        {
            if (previewBitmap == null || cmbEdgeDetection.SelectedIndex =
            {
                return;
```

Fig. 2. *codeLens* and *peek definition* in Visual Studio all the methods called inside the implementation of a particular method. Moreover, it allows to create a graph that represents the selected call hierarchy.

Apart from tools projecting structure of the whole projects, all selected IDEs contain more focused tools that display detailed structure of the file currently viewed in the editor. This is intended to speed-up the navigation inside a single file. A related to this projection is the *breadcrumbs* panel that shows context of the currently edited part of the file based on the position of the caret in the editor. Eclipse and IntelliJ IDEA contain also projections that can display *type hierarchy* of classes and *call hierarchy* of methods.

### C. Search and Go-to Projections

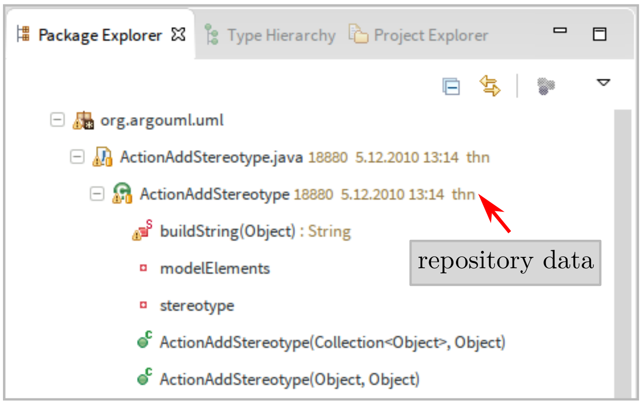
The next group of projections relate to searching for specific program elements across the project. The simple text-based search is by the IDEs extended to take into account the source code structure. This helps to connect related elements and distinguish between the different elements represented with the same name (the same text).

One of the tools providing this kind of projection is used to search for all references of the same program element. This tool is called differently in each IDE, with names like *find usages* or *find all references*. It is useful for quick navigation and for discovering code dependencies. Another related tool deals with the comments. It displays all code locations where comment starts with the word "todo", or other configured pattern, which makes possible to track tasks across the project (see fig. 4).

Similar are tools for quick navigation between different program element occurrences and within the inheritance hierarchy. These include tools with self-explanatory names like *go to declaration*, *go to implementation*, *go to super implementation* and *go to type declaration*. If they can navigate to only one place in the source code, there is no "view" created and the action implied by the tool's name is immediately performed.

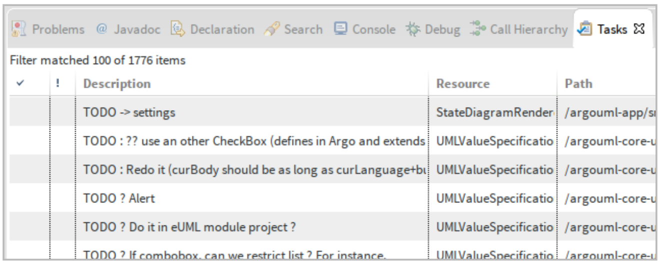
### D. Domain-specific Projections

The projections reviewed so far drew the required information for their construction from the parsed AST. Many general-purpose IDEs (in our case Eclipse, IntelliJ IDEA, NetBeans and Visual Studio) have support for a number of software frameworks. And there are IDEs that are built specifically to support some frameworks (e.g., Xamarin for mobile application development). These IDEs take advantage of common application structures, conventional configurations and domain-specific APIs in order to simplify development of applications with supported frameworks. As a result, code projections that support specific framework features were found in the reviewed tools as well.



org.argouml.uml
ActionAddStereotype.java 18880 5.12.2010 13:14 thn
ActionAddStereotype 18880 5.12.2010 13:14 thn
buildString(Object) : String
modelElements
stereotype
ActionAddStereotype(Collection<Object>, Object)
ActionAddStereotype(Object, Object)

Fig. 3. Package explorer in Eclipse



Filter matched 100 of 1776 items

| ✓ | ! | Description | Resource | Path |
|---|---|---|---|---|
| | | TODO -> settings | StateDiagramRender | /argouml-app/s |
| | | TODO : ?? use an other CheckBox (defines in Argo and extends | UMLValueSpecificatio | /argouml-core-u |
| | | TODO : Redo it (curBody should be as long as curLanguage+bu | UMLValueSpecificatio | /argouml-core-u |
| | | TODO ? Alert | UMLValueSpecificatio | /argouml-core-u |
| | | TODO ? Do it in eUML module project ? | UMLValueSpecificatio | /argouml-core-u |
| | | TODO ? If combobox, can we restrict list ? For instance, | UMLValueSpecificatio | /argouml-core-u |

Fig. 4. TODO tasks in Eclipse

Applications that use a particular framework have (to some extent) common structure and this is exploited by these IDEs. Based on the known structure, program element across different languages can be interconnected. Web applications created with the Spring or Play frameworks in IntelliJ IDEA or with ASP.Net framework in Visual Studio can have code completion working for dynamic segments of page templates because the IDE "knows" which program element represents the context of the template.

An example of a projection based on a framework configuration is the ability of IntelliJ IDEA to view and edit classes annotated with the Java Persistence API's annotations in the form of entity relationship diagram.

The common example of domain-specific projection among the reviewed Java-supporting IDEs – Eclipse, IntelliJ IDEA and NetBeans – is the graphical user interface builder for the Swing framework. Eclipse can project class inheriting from one of the Swing window components directly to its graphical representation and any edits made in this graphical view are reflected back to the code. To achieve the same functionality, IntelliJ IDEA and NetBeans use intermediary XML file that represents the layout of user interface components. Graphical representation is the result of projecting this XML file. In the other direction, the source code is generated from the XML. IntelliJ IDEA by default postpones this code generation to the compile time, while NetBeans updates code on each change of the design. Similar projections are available also in Visual Studio for Windows Forms framework and in the Xamarin IDE for creating graphical layouts for mobile applications. All these projections require the IDE to "understand" the API of the particular framework.

## IV. Projectional Tools in the Research

The research in the area of projectional tools is mostly associated with the tools like the already mentioned JetBrains

MPS. In their work [5], [7], Voelter et al. focus primarily on this language workbench. They discuss the concept of projections and possibilities they brig to the domain-specific language development, but also the associated issues of direct AST manipulation. They also deal with projections in the context of composition and extension of programming languages.

There is an ongoing endeavor in the research community to create tools that supports program comprehension. Few of those are generally viewed as projectional tools. However, many of the tools are implemented as IDE extensions and have projectional properties similar to those we described in the review.

Among these tools, the *Fluid source code views* [8] tool implemented for the Eclipse IDE is similar to the *peek definition* feature of Visual Studio that was reviewed in this paper. *Registration-based abstractions* presented by Davis et al. [9] are another example of IDE projections – in this case they change language syntax in the Eclipse IDE editor to achieve easier comprehension of frequently used coding patterns.

Another type of tools deal with the program concerns. The *Sieve Source Code Editor* [10] for the NetBeans IDE uses the structured code comments to create concern-oriented views of the source code. The *Code Bubbles* [11] and the *Code Canvas* [12] are alternative code editors for the Eclipse and Visual Studio IDEs, respectively. These provide projections of the source code that are abstracted from traditional file-based views and allow programmers to create free-form arrangements of the code they are working with.

## V. Concluding Remarks

In this paper we presented a review of code projections that are available in today's integrated development environments. We identified four main categories of the tools that feature projectional behavior.

The *in-editor* projections enhance the textual notation of the source code with code highlighting or completion. The *structure* projections are available as a separate panels that create views of the system concentrated on its structure, providing thus higher-level view that is easier to grasp than the full source code with all of its details. The *search and go-to* projections are provided by the tools for searching the related program elements or navigating the class hierarchies. In the most cases, these three categories require for their functionality a knowledge that can be extracted from abstract syntax tree analysis. We found also projections that use other information sources, exploiting advantages of the tools that are available in the integrated environments. There are data projected from debuggers, version control systems or test runners. The last category contains *domain-specific* projections that rely on conventions of supported frameworks, such as common application structures and APIs designed for a particular domain. As a result, they can provide code completion across languages, project annotated classes as relational diagrams or enable graphical builders of user interfaces.

We have shown in the review that the projections are not only a matter of projectional language workbenches as they are extensively used in many aspects of modern source-based integrated development environments. Their infrastructure is well-prepared for such functionality, which is also exploited by many research tools that are implemented as extensions of these IDEs. And as the MPS is approaching the usability of the standard source-based editing [5], the distinction between the AST-based and the source-based editors is becoming less apparent.

## Acknowledgment

This work was supported by the project KEGA 008TUKE-4/2013 and by the Slovak Research and Development Agency under the contract No. APVV-0008-10.

## References

[1] T. Kosar, M. Mernik, and J. Carver, "The impact of tools supported in integrated-development environments on program comprehension," in *33rd International Conference on Information Technology Interfaces (ITI'11)*, 2011. ISSN 1330-1012 pp. 603–608.

[2] W. Maalej, R. Tiarks, T. Roehm, and R. Koschke, "On the Comprehension of Program Comprehension," *ACM Transactions on Software Engineering and Methodology*, vol. 23, no. 4, pp. 1–37, Aug. 2014. doi: 10.1145/2622669. [Online]. Available: http://dx.doi.org/10.1145/2622669

[3] M.-A. Storey, "Theories, Methods and Tools in Program Comprehension: Past, Present and Future," in *13th International Workshop on Program Comprehension (IWPC'05)*. IEEE, 2005. doi: 10.1109/WPC.2005.38. ISBN 0-7695-2254-8 pp. 181–191. [Online]. Available: http://dx.doi.org/10.1109/WPC.2005.38

[4] M. Fowler, "Projectional Editing," 2008. [Online]. Available: http://martinfowler.com/bliki/ProjectionalEditing.html

[5] M. Voelter, J. Siegmund, T. Berger, and B. Kolb, "Towards User-Friendly Projectional Editors," ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8706, pp. 41–61. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11245-9_3

[6] M. Nosáľ, J. Porubän, and M. Nosáľ, "Concern-oriented source code projections," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, Kraków, 2013, pp. 1541–1544.

[7] M. Voelter, B. Kolb, and J. Warmer, "Projecting a Modular Future," *IEEE Software*, pp. 1–1, 2014. doi: 10.1109/MS.2014.103. [Online]. Available: http://dx.doi.org/10.1109/MS.2014.103

[8] M. Desmond, M. Storey, and C. Exton, "Fluid source code views," in *Program Comprehension, 2006. ICPC 2006. 14th IEEE International Conference on*, 2006. doi: 10.1109/ICPC.2006.24 pp. 260–263. [Online]. Available: http://dx.doi.org/10.1109/ICPC.2006.24

[9] S. Davis and G. Kiczales, "Registration-based language abstractions," *ACM SIGPLAN Notices*, vol. 45, no. 10, p. 754, Oct. 2010. doi: 10.1145/1932682.1869521. [Online]. Available: http://dx.doi.org/10.1145/1932682.1869521

[10] J. Porubän and M. Nosáľ, "Leveraging Program Comprehension with Concern-oriented Source Code Projections," *3rd Symposium on Languages, Applications and Technologies. OpenAccess Series in Informatics (OASIcs)*, vol. 38, pp. 35–50, 2014. doi: 10.4230/OASIcs.SLATE.2014.3. [Online]. Available: http://dx.doi.org/10.4230/OASIcs.SLATE.2014.35

[11] A. Bragdon, R. Zeleznik, S. P. Reiss, S. Karumuri, W. Cheung, J. Kaplan, C. Coleman, F. Adeputra, and J. J. LaViola, "Code bubbles: a working set-based interface for code understanding and maintenance," in *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*. ACM Press, Apr. 2010. doi: 10.1145/1753326.1753706 pp. 2503–2512. [Online]. Available: http://dl.acm.org/citation.cfm?id=1753326.1753706

[12] R. DeLine and K. Rowan, "Code Canvas: Zooming towards Better Development Environments," in *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - ICSE '10*, vol. 2. ACM Press, Jan. 2010. doi: 10.1145/1810295.1810331 p. 207. [Online]. Available: http://dx.doi.org/10.1145/1810295.1810331

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 929–934
DOI: 10.15439/2015F238
ACSIS, Vol. 5

# Profile-driven Source Code Exploration

Emília Pietriková, Sergej Chodarev
Technical University of Košice, Department of Computers and Informatics
Letná 9, 04200 Košice, Slovak Republic
Email: {emilia.pietrikova, sergej.chodarev}@tuke.sk

***Abstract*—The following study deals with static analysis of Java source codes and it is dedicated to those readers who are interested in techniques aiming at evaluation of programming abilities of job candidates or students. In our case, the goal of the static analysis is to assemble the most significant and interesting data about source code author (programmer). If properly visualized, such assembled data may form programmer's profile which, to impartial observer, may further determine author's real programming abilities and his/her habits, both good and the bad ones. The present study represents first experiments attempting to form programmer's profile by static analysis of language element frequency. Conclusion offers a broader view, combining also other techniques as a future plan to generate knowledge profiles more precisely.**

## I. Introduction

KNOWLEDGE, skills and their level are often the focus of attention in many disciplines. In order to be successful, people are often compared with each other. In the area of programming it is similar, however, the range of skills-tracking possibilities is quite limited. In the following study, we present early stages of profile-driven source code analysis where our interest is focused on source code exploration with the intention of knowledge profile generation. Such a profile represents an objective evaluation of current knowledge and skills, individual progress compared to the past, or possible deficiencies to be addressed.

Knowledge profile may be beneficial for both beginners and experienced programmers as well as for lecturers. Profiles can be helpful during overall student assessment, moreover, they can be used when identifying course drawbacks towards improvement of the course. In labor market, job candidates may find programming profile generators beneficial as well. That is, this study is dedicated to those researchers who deal with source code analysis, focusing on author of the code.

There exists a large variety of automated tools dedicated to source code analysis. These tools deal with code from various perspectives, e.g. security evaluation, quality, design etc. Outputs of such tools mostly include reports reflecting various metrics, graphs or warnings. They, however, do not collect a profile of the programmer knowledge [1], [2]. More from the related work can be found in Section V.

In this study, our intention is to generate a programming knowledge profile from source code with a possibility of its comparison with different profiles. This includes comparison of the current profile with a profile which was actual in the past. This way, the profile report may point out author's progress. Profiles of the group of programmers can be compared with each other to reveal possible differences in their knowledge. It should be also possible to compare a personal profile to some explicitly defined knowledge level (e.g. needed to fulfill specific task). We consider tracking and comparing source code in the form of summarizing profiles as a contribution to a new view of knowledge, to a better analysis and filtration of irrelevant data. Yet, to our best knowledge, such a profile-driven tool has not been developed.

In the following sections, we describe the concept of knowledge profiles (Section II) and we introduce an initial prototype proposed and developed as a source code exploring tool (Section III). This tool operates on the basis of static analysis and it represents a partial solution of the presented task. The tool works with Java language constructs and it visualizes knowledge profiles based on various statistics and metrics. We discuss results generated by the tool on a medium-sized project as well as on a large project (Section IV).

## II. Knowledge profiles

In general, we understand knowledge profile as a description of knowledge and bindings between its elements necessary to handle a specific task.

In our study, we focus on knowledge profiles in an area where it is possible to formally define such a profile and to construct it automatically from particular input artifacts. Primarily, we deal with an area of programming where the artifacts are represented by source code and a profile is formally defined over a language in which the source code is created. We distinguish two types of profiles: subject and object profile.

*Subject profile* represents an expression of what the subject (author of the code) knows, how deep is his/her knowledge, what kind of issues is the subject capable to solve. In programming, this means that the subject (programmer) knows, for example, how to use `if` command, how to call or declare a function, how to use generic programming [3]. That is, the subject profile represents the range of tasks actually solvable by the programmer.

*Object profile* represents a profile of knowledge necessary to handle a specific task (or tasks) over some object. A programming book may define knowledge profile of prerequisites, i.e. what any reader should know before reading the book in order to understand its contents. There can even be a differential object profile determining what is the reader supposed to learn (know after understanding the book contents). Such a differential profile can be determined for each book chapter

978-83-60810-66-8/$25.00©2015, IEEE

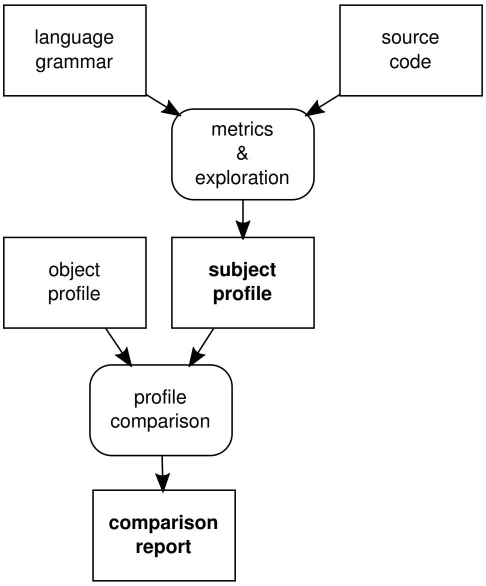


Fig. 1. Simplified knowledge profile generator scheme

as well. That is, the object profile represents the range of tasks which are supposed to be known by the programmer (but the actual state may be different).

The profile should allow to verify whether a specific programmer has sufficient knowledge to solve a task. Moreover, it should identify missing knowledge. For this reason, the profile needs to be structured. Such an assumption is supported by the fact that each programming task or its solution are structured as well [4]. In other words, if an actual (incomplete) task solution is structured than it is possible to assume that the same or similar knowledge, which has already been applied, will be necessary to complete the task.

In an early stage of our research the profile prototype is represented by a simple table, later we assume its transformation to a tree or a graph with annotated nodes or edges [5].

### *A. Profile Construction*

A profile can be constructed manually, however, an important part of our research is to generate profiles automatically from artifacts (source code). That is, one profile is supposed to be constructed after processing a finite number of source code files through their analysis. This way it is possible to generate an object profile and also a subject profile provided source code created by the subject is available. For experimental purposes, object profiles may be created manually. In order to construct a subject profile, it is necessary to analyze source code synthesized (created) by the subject.

The idea is depicted using the scheme in Fig. 1. The object profile is optional, so it does not have to be necessarily present. However, language and source code are compulsory. Without these two artifacts, subject profile cannot be generated. If both subject and object profiles are present, a comparison profile can be generated. Such a profile can be bind to a specific task through the object profile.

Source code analysis can be performed through parser of a particular language. An assumption is that particular grammar rules define concepts and the rules used by the code author mean that the programmer understands language constructs describing and defining the language. Complete language syntax is not necessary when processing source code, however, syntax definition should be accustomed to required knowledge expression. An appropriate form of rules should be as expressed in Eq. 1 not 2.

$$If \rightarrow \text{"}\texttt{if}\text{"}\ \text{"("}\ Expression\ \text{")"}\ Statement \tag{1}$$

$$A \rightarrow \text{"}\texttt{if}\text{"}\ \text{"("}\ B\ \text{")"}\ C \tag{2}$$

That is, the form should be human-interpretable, e.g. in order to understand *if*, one should understand expressions and statements.

Obviously, such a naive approach is not sufficient when creating a complex profile. The fact that the code author who called a function might indicate that he understands it, however, one function call does not provide a clear evidence that the subject perfectly understands every detail related to this function. This is why there is a need for metrics definition, based also on empirical observation. In the metrics, we may take into account multiplicity of one method use assuming the more is one method used, the more the subject understands it. Moreover, we may assess method complexity (code length and documentation length) [6]. Such metrics definition represents a separate part of the research regarding knowledge profile generation.

### *B. Use cases*

Presented approach towards knowledge profiles generation can find its practical benefits in the following:

- book profile (object profile) – based on subject profile, one may select the most helpful book,
- candidate selection (subject profile) – regarding a task or group of tasks (object profile) supposed to be solved,
- determination of skills necessary to handle some task (object profile) – based on subject profile,
- statistical evaluation of what people frequently use/not use – may indicate the difficulty of use (subject profiles)
- determination of language constructs complexity or library complexity.

The verification of the proposed method of knowledge profile generation can be done within the educational process, e.g. by creating a record of changes in student profile after passing a programming course. Such a record may be beneficial during the exam, indicating student improvement.

As stated in [7], static analysis tools generate lots of data. Therefore, in addition to appropriate techniques of profile creation, two other topics are related and represent a separate part of the research: usability and visualization of knowledge profiles. Assembled data regarding subject or object profile cannot be beneficial if the way of their visualization as well as the user interface are disarranged or too complicated to make any sense.

## III. Prototype

To evaluate the concept, a prototype has been implemented, that allows to analyze program source code written in Java language. The prototype represents only the first iteration of our research in the area. It uses the counts of language constructs used in the code to generate a profile. The profile itself is represented as a table containing the counts for each source code file and also summary data. The table is serialized in JSON format.

The data are then visualized in different ways to allow their further examination and comparison.

The tool provides four ways to display profile data:

1) *Detailed tables* — display counts of the used language constructs for each source code file. Constructs are divided into several logical groups (e.g. arithmetic operators or control flow statements) that are displayed in separate tables.
2) *Summary tables* — display summary counts for all files with values of statistical variables like arithmetic mean, modus, median, standard deviation etc, that characterize distribution of a language construct between source code files.
3) *Heat maps* — represent a matrix with total counts for each language construct, where cells of the matrix are colored according the counts (darker color means higher occurrence) and additional statistical data is displayed in a tooltip window (see Fig. 2 that displays comparison of several profiles).
4) *Box plots* also called box-and-whisker plots [8] — visually display summary data together with their distribution (see Fig. 3).

The tool can display simple profiles – data collected for some set of source code files that are produced by single person (subject profile) or are part of a single project (object profile). In addition, there are two compound types of profiles that consist of several simple profiles:

- *group profiles* that display data for several profiles and allows to summarize and compare them,
- *comparison profiles* that allow to compare several profiles with a single master profile.

For the comparison purposes, the most valuable type of display turned out to be *the heat map*. It allows to display a large set of data in a compact form which is easy to explore and therefore allows to visually find anomalies that may indicate significant results.

To implement the parser of Java language, ANTLR parser generator [9] was used. The visualization is based on web technologies such as AngularJS framework[1] and HighCharts interactive graph plotting library[2].

## IV. Experiments

We have performed several experiments regarding the developed tool.

### A. Analysis of language constructs used in large projects

The tool has been tested on several existing projects, both medium and large. The goals was to assess how Java language constructs are used in them and to test extraction of object profiles.

The results show that in medium-sized projects there is a lot of language constructs that are not used at all. For example, one of the tested projects – YAJCo parser generator [10] – did not adopt any bitwise and bit shift operators, a large part of the arithmetic operators and some other constructs.

Even in a large project, like Google Guava library[3], there are constructs that are never used, including some bit shift operators, try-with-resource blocks, and a default argument for annotation parameters. Bit shift operators, however, have been used in a form of compound assignment operators. On the other hand, try-with-resource was added to the language in version 7, so it has probably been avoided for compatibility reasons.

On the other hand, in the large project most of the Java language constructs was used at least once. This means that profiles based solely on language construct counting would not be comprehensive enough for such projects. For this reason we plan to extend the prototype with advanced analysis and additional metrics.

### B. Comparison of student assignments

To evaluate the comparison of profiles, we planned to compare projects of similar size and in the same domain. For this reason we have chosen source code developed by our students as part of their assignments. We used the assignments from the Object-Oriented Programming course. For most students, this course is a first introduction into Java language and object-oriented methodology. This way we were able to compare subjects with similar starting knowledge working on the same problem. We also added a solution developed by a teacher to the comparison.

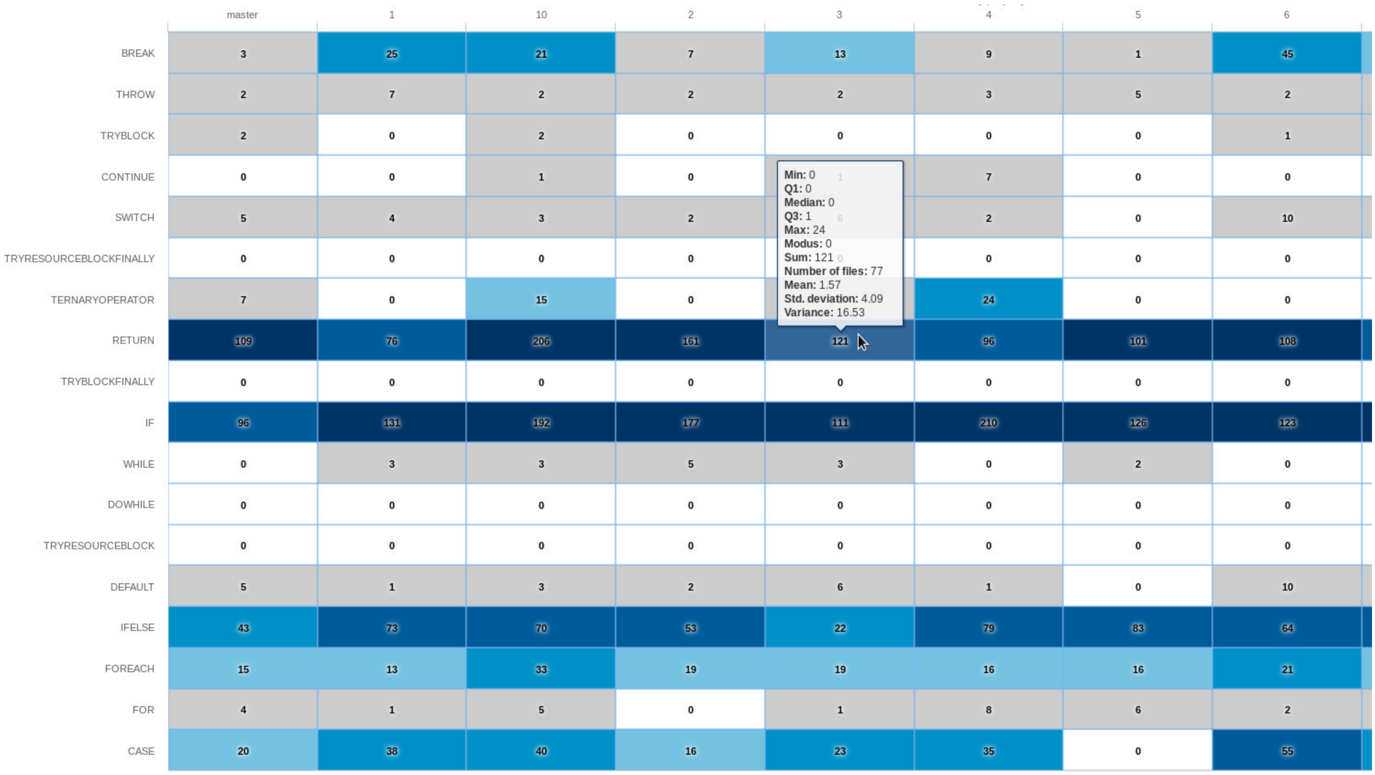
Fig. 2 shows a fragment of the comparison results. Rows correspond to different language constructs, for example *break* statement or *try* block. Columns represent different projects. Students' projects are identified by numbers while teacher's project is called *master*. The table contains a total number of construct occurrences within analyzed codes. After pointing to some table cell, a tooltip window appears containing statistical parameters that represent distribution of the language construct per source code file (see return statement for student 6 in Fig. 2).

The comparison has show that in general the results of both students' and teacher's solution are quite similar. There are, however, some notable differences. For example, student 3 used the largest number of different language constructs,

[1] https://angularjs.org/

[2] http://www.highcharts.com/

[3] https://github.com/google/guava



| | master | 1 | 10 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|
| BREAK | 3 | 25 | 21 | 7 | 13 | 9 | 1 | 45 |
| THROW | 2 | 7 | 2 | 2 | 2 | 3 | 5 | 2 |
| TRYBLOCK | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 1 |
| CONTINUE | 0 | 0 | 1 | 0 | 1 | 7 | 0 | 0 |
| SWITCH | 5 | 4 | 3 | 2 | 6 | 2 | 0 | 10 |
| TRYRESOURCEBLOCKFINALLY | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| TERNARYOPERATOR | 7 | 0 | 15 | 0 | [illegible] | 24 | 0 | 0 |
| RETURN | 109 | 76 | 206 | 161 | 121 | 96 | 101 | 108 |
| TRYBLOCKFINALLY | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| IF | 96 | 131 | 192 | 177 | 111 | 210 | 126 | 123 |
| WHILE | 0 | 3 | 3 | 5 | 3 | 0 | 2 | 0 |
| DOWHILE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| TRYRESOURCEBLOCK | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DEFAULT | 5 | 1 | 3 | 2 | 6 | 1 | 0 | 10 |
| IFELSE | 43 | 73 | 70 | 53 | 22 | 79 | 83 | 64 |
| FOREACH | 15 | 13 | 33 | 19 | 19 | 16 | 16 | 21 |
| FOR | 4 | 1 | 5 | 0 | 1 | 8 | 6 | 2 |
| CASE | 20 | 38 | 40 | 16 | 23 | 35 | 0 | 55 |

**Min:** 0
**Q1:** 0
**Median:** 0
**Q3:** 1
**Max:** 24
**Modus:** 0
**Sum:** 121
**Number of files:** 77
**Mean:** 1.57
**Std. deviation:** 4.09
**Variance:** 16.53

Fig. 2. Fragment of the students assignments comparison displayed as a heat map

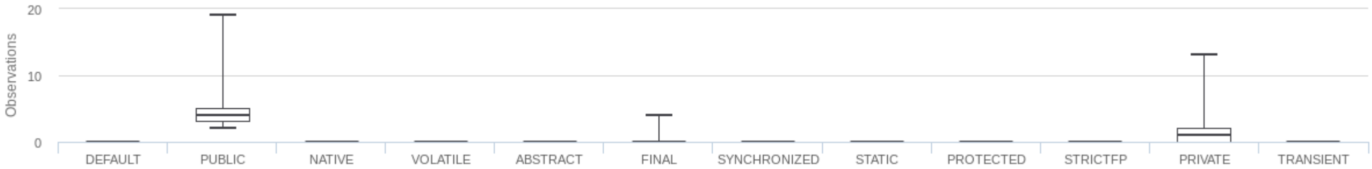


Fig. 3. Example of the box plot displaying counts of modifiers per source code file

even the ones not used by the teacher. Student 7, on the other hand, probably encountered problems with understanding the principles of object-oriented programming, since he used the *static* modifier much more often than the other subjects.

Some students missed language constructs used by all the others. For example, student 1 did not use *float* and *long* types, student 5 did not use *switch* statement. This may indicate that they did not understand these types or constructs, or they simply selected a different implementation strategy. Therefore, exploration of the source codes themselves is needed in both cases.

On the other hand, the fact that student 6 did not use *final* modifier quite clearly indicates that he does not understand the importance of immutability in programs.

## V. Related Work

There is a number of studies dealing with source code analysis. Most of them are focused on software security, detecting bugs, defects and potential vulnerabilities. Two of such studies are [11] and [12], both dedicated to static analysis of C/C++ source code. Static analysis tools which are the most popular usually explore static code and identify a large variety of bugs and bad programming practice [13].

Usually, static analysis refers to methods of automated determination of a program behavior during compile time. Static analysis tools have become part of modern compilers, however, these tools can only identify elementary errors [14]. E.g. traditional tools cannot identify the presence of deadlocks, having their own research branch [15], [16], or breaking mutual exclusion in concurrent applications [17], [18].

A method dedicated to collecting, comparing, and combining program semantics is refered to as abstract interpretation and it has been successfully used to derive run time properties of a program which can be used for program optimization. Other objectives of static analysis tools are mostly concept location [19], [20], code transformation [21], [22], security [23], [24], or reverse engineering [25].

When dealing with techniques of the static analysis, one may refer to [26], published a decade ago but still actual, focused on various approaches in software testing based on automata theory. One may also refer to a newer publication

[27], of which authors claim that empirical code evaluation plays an important role in software analysis.

A technique presented in [28] locates computational units typical for a set of related features through execution profiles. In order to detect the most feature-specific computational units, concept analysis is performed [28]. This is combined with static analysis using the feature-specific computational units to detect additional units along with the dependency graph.

Static software analysis has been also covered by a number of surveys, e.g. [29] or [30].

An interesting source of data for programming profile generation may be software repositories, produced and archived throughout software development [31]. In order to explore and examine software repositories, mining software repositories (MSR) have been created. As stated in [32], MSR exploration used to be subjected on industrial systems in the past. However, with an extensive increase of open-source software, this research has become a new challenge. MSR researchers mostly focus on clearer understanding of software evolution [33], development of tools, methods and processes.

Metadata analysis differs with particular exploration objectives and software repositories. The most common issues addressed by MRS researchers are [34]:

- detection of change patterns,
- prediction of changes,
- detection of bugs,
- analysis of bug-fixing change,
- source code exploration,
- identification of software developers.

All of the mentioned issues have one main objective in mind: To augment traditional software engineering techniques in order to guide decision processes in modern software projects [35]. That is, while MSR researchers focus on programming targets (programming result – software), our attention is paid to the source (software author). Since the aim of this study is to assess quality of the code author, source code exploration and developer identification [36] are the most related issues.

## VI. Conclusion

Having been first described as a prototype, this study has dealt with an exploration tool aimed at the Java code of various programmers. The main objective is to automatically generate programmer assessment profile. The analysis indicates the topic is quite extensive and little explored. This is why we described only a few of the potential methods for the profile generation.

The proposed tool has been developed and experimentally evaluated on several code samples including a medium-sized and large software project. The tool is intended to analyze knowledge through counting the language constructs. In order to clarify the results, the method utilizes elements of the descriptive statistics [37]. Within the experiment, various possibilities of source code processing have been implemented. For all the processing types, results are available in JSON meta-form as well as in various types of graphical web-based representation.

Since we are still in early stages, the described code-exploring tool and its continuing maturation will include the capability of treating more complex code solutions and utilization of additional metrics. The future plan is to detect and evaluate more advanced language usage e.g. nested loops, or programming idioms [38]. The tool should also track used library classes and method in addition to built-in language constructs. In order to support this, it would be required to implement processing of references in a programming language [39]. In distinction to general search-based techniques, future work will also involve model-based deductive evaluation, similar to [17]. Moreover, if combined with automated code-functionality evaluation during the educational process [40], knowledge profiles may become a significant contribution to student assessment.

Even with the current implementation it is hard to manually analyze and compare large number of profiles. This means that growing amount of data in the profile would require advanced methods of its visualization and automated analysis.

Apparently, automated knowledge evaluation might not be completely accurate. In order to achieve more precise profile results, it will be necessary to perform a lot of experiments over a large group source code and to combine several types of metrics or statistics. However, interesting results will be visible immediately.

## Acknowledgment

This work was supported by project VEGA 1/0341/13 “Principles and methods of automated abstraction of computer languages and software development based on the semantic enrichment caused by communication”.

## References

[1] D. Mihályi and V. Novitzká, “Towards the Knowledge in Coalgebraic model of IDS,” *Computing and Informatics*, vol. 33, no. 1, pp. 61–78, 2-14.

[2] J. Paralič, F. Babič, and M. Paralič, “Process-driven Approaches to Knowledge Transformation,” *Acta Polytechnica Hungarica*, vol. 10, no. 5, pp. 125–143, 2013.

[3] R. Garcia, J. Jarvi, A. Lumsdaine, J. G. Siek, and J. Willcock, “A Comparative Study of Language Support for Generic Programming,” *SIGPLAN Notices*, vol. 38, no. 11, pp. 115–134, 2003. doi: 10.1145/949343.949317

[4] J. Kollár and P. V. Jaroslav Porubän, “Separating concerns in programming: Data, control and actions,” *Computing and Informatics*, vol. 24, no. 5, pp. 441–462, 2005.

[5] M. Nosáľ and J. Porubän, “XML to Annotations Mapping Definition with Patterns,” *Computer Science and Information Systems*, vol. 11, no. 4, pp. 1455–1477, 2014. doi: 10.2298/CSIS130920049N

[6] M. Nosáľ, J. Porubän, and M. Nosáľ, “Concern-oriented Source Code Projections,” in *Federated Conference on Computer Science and Information Systems (FEDCSIS)*. IEEE, 2013. ISBN 978-1-4673-4471-5 pp. 1541–1544.

[7] S. Heckman and L. Williams, “A Comparative Evaluation of Static Analysis Actionable Alert Identification Techniques,” in *International Conference on Predictive Models in Software Engineering*. ACM, 2013. doi: 10.1145/2499393.2499399 pp. 4:1–4:10.

[8] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[9] T. Parr and K. Fisher, “LL(*): the foundation of the ANTLR parser generator,” *SIGPLAN Notices*, vol. 46, no. 6, pp. 425–436, 2011. doi: 10.1145/1993316.1993548

[10] J. Porubän, M. Forgáč, M. Sabo, and M. Běhálek, "Annotation based parser generator," *Computer Science and Information Systems*, vol. 7, no. 2, pp. 291–307, 2010. doi: 10.2298/CSIS1002291P

[11] R. Huuck, "Technology transfer: Formal analysis, engineering, and business value," *Science of Computer Programming*, vol. 103, pp. 3–12, 2015. doi: 10.1016/j.scico.2014.11.003

[12] V. Ivannikov, A. Belevantsev, A. Borodin, V. Ignatiev, D. Zhurikhin, and A. Avetisyan, "Static analyzer Svace for finding defects in a source program code," *Programming and Computer Software*, vol. 40, no. 5, pp. 265–275, 2014. doi: 10.1134/S0361768814050041

[13] Q. Hanam, L. Tan, R. Holmes, and P. Lam, "Finding Patterns in Static Analysis Alerts: Improving Actionable Alert Ranking," in *Working Conference on Mining Software Repositories*. ACM, 2014. doi: 10.1145/2597073.2597100 pp. 152–161.

[14] V. Djukić, I. Luković, A. Popović, and V. Ivančević, "Model Execution: An Approach based on extending Domain-Specific Modeling with Action Reports," *Computer Science and Information Systems*, vol. 10, no. 4, pp. 1585–1620, 2013. doi: 10.2298/CSIS121228059D

[15] P. T. Breuer and S. Pickin, "One Million (LOC) and Counting: Static Analysis for Errors and Vulnerabilities in the Linux Kernel Source Code," in *Reliable Software Technologies – Ada-Europe*, ser. Lecture Notes in Computer Science. Springer, 2006, vol. 4006, pp. 56–70.

[16] M. Tomášek, "Language for a Distributed System of Mobile Agents," *Acta Polytechnica Hungarica*, vol. 8, no. 2, pp. 61–79, 2011.

[17] Z. Lu and S. Mukhopadhyay, "Model-Based Static Source Code Analysis of Java Programs with Applications to Android Security," in *Computer Software and Applications Conference (COMPSAC)*. IEEE Computer Society, 2012. doi: 10.1109/COMPSAC.2012.43 pp. 322–327.

[18] S. Šimoňák, "Verification of Communication Protocols Based on Formal Methods Integration," *Acta Polytechnica Hungarica*, vol. 9, no. 4, pp. 117–128, 2012.

[19] D. Poshyvanyk, M. Gethers, and A. Marcus, "Concept Location Using Formal Concept Analysis and Information Retrieval," *ACM Transactions on Software Engineering Methodology (TOSEM)*, vol. 21, no. 4, pp. 23:1–23:34, 2013. doi: 10.1145/2377656.2377660

[20] A. Marcus, V. Rajlich, J. Buchta, M. Petrenko, and A. Sergeyev, "Static Techniques for Concept Location in Object-Oriented Code," in *International Workshop on Program Comprehension*. IEEE Computer Society, 2005. doi: 10.1109/WPC.2005.33 pp. 33–42.

[21] F. Catthoor, K. Danckaert, S. Wuytack, and N. Dutt, "Code transformations for data transfer and storage exploration preprocessing in multimedia processors," *Design Test of Computers*, vol. 18, no. 3, pp. 70–82, 2001. doi: 10.1109/WPC.2005.33

[22] A. C. Murray, R. V. Bennett, B. Franke, and N. Topham, "Code Transformation and Instruction Set Extension," *ACM Transactions on Embedded Computer Systems (TECS)*, vol. 8, no. 4, pp. 26:1–26:31, 2009. doi: 10.1145/1550987.1550989

[23] A. Baláž, *Computer Systems Security*, 2nd ed., 2015. ISBN 978-80-553-1948-3

[24] L. Vokorokos, A. Baláž, and N. Ádám, "Secure web server system resources utilization," *Acta Polytechnica Hungarica*, vol. 12, no. 2, pp. 5–19, 2015.

[25] H. M. Kienle and H. A. Müller, "Rigi — An Environment for Software Reverse Engineering, Exploration, Visualization, and Redocumentation," *Science of Computer Programming*, vol. 75, no. 4, pp. 247–263, 2010. doi: doi:10.1016/j.scico.2009.10.007

[26] G. J. Holzmann, "Software Analysis and Model Checking," in *Computer Aided Verification*, ser. Lecture Notes in Computer Science. Springer, 2002, vol. 2404, pp. 1–16.

[27] M. B. Dwyer, J. Hatcliff, R. Robby, C. S. Pasareanu, and W. Visser, "Formal Software Analysis Emerging Trends in Software Model Checking," in *Future of Software Engineering*. IEEE Computer Society, 2007. doi: 10.1109/FOSE.2007.6 pp. 120–136.

[28] T. Eisenbarth, R. Koschke, and D. Simon, "Locating features in source code," *IEEE Transactions on Software Engineering*, vol. 29, no. 3, pp. 210–224, 2003. doi: 10.1109/TSE.2003.1183929

[29] P. Emanuelsson and U. Nilsson, "A Comparative Study of Industrial Static Analysis Tools," *Electronic Notes in Theoretical Computer Science*, vol. 217, pp. 5–21, 2008. doi: 10.1016/j.entcs.2008.06.039

[30] S. Heckman and L. Williams, "A Systematic Literature Review of Actionable Alert Identification Techniques for Automated Static Code Analysis," *Information and Software Technology*, vol. 53, no. 4, pp. 363–387, 2011. doi: 10.1016/j.infsof.2010.12.007

[31] S. O. Olatunji, Y. S. Al-Ghamdi, and J. S. A. Al-Ghamdi, "Mining Software Repositories – A Comparative Analysis," *International Journal of Computer Science and Network Security*, vol. 10, no. 8, pp. 161–174, 2010.

[32] H. Kagdi, M. L. Collard, and J. I. Maletic, "A Survey and Taxonomy of Approaches for Mining Software Repositories in the Context of Software Evolution," *Journal of Software Maintenance and Evolution: Research and Practice*.

[33] J. Kollár and M. Forgáč, "Combined approach to program and language evolution," *Computing and Informatics*, vol. 29, no. 6+, pp. 1103–1116, 2010.

[34] K. Chaturvedi, V. Sing, and P. Singh, "Tools in Mining Software Repositories," in *Computational Science and Its Applications (ICCSA)*, 2013. doi: 10.1109/ICCSA.2013.22 pp. 89–98.

[35] A. Hassan, "The road ahead for Mining Software Repositories," in *Frontiers of Software Maintenance*, 2008. doi: 10.1109/FOSM.2008.4659248 pp. 48–57.

[36] S. Koch and G. Schneider, "Effort, cooperation and coordination in an open source software project: Gnome," *Information Systems Journal*, vol. 12, no. 1, pp. 27–42, 2002. doi: 10.1046/j.1365-2575.2002.00110.x

[37] W. Trochim, "Research methods knowledge base: Descriptive statistics," http://www.socialresearchmethods.net/kb/statdesc.php, 2006, Accessed: 2015-04-30.

[38] A. Sutton, R. Holeman, and J. I. Maletic, "Identification of idiom usage in C++ generic libraries," *International Conference on Program Comprehension*, pp. 160–169, 2010. doi: 10.1109/ICPC.2010.37

[39] D. Lakatoš, J. Porubän, and M. Bačíková, "Declarative specification of references in DSLs," in *Federated Conference on Computer Science and Information Systems (FedCSIS 2013)*. IEEE, 2013. ISBN 978-1-4673-4471-5 pp. 1527–1534.

[40] M. Biňas, "Improving reliability of arena platform for automated assessments," in *Electrical Engineering and Informatics 5: Proceedings of FEEI*, 2014, pp. 115–118.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 935–946
DOI: 10.15439/2015F156
ACSIS, Vol. 5

# A step towards genuine declarative language-integrated queries

Radosław Adamus
Institute of Applied Computer Science, Lodz University of Technology
90-924 Lodz, ul. Stefanowskiego 18/22, Poland
Email: r.adamus@iis.p.lodz.pl

Tomasz Marek Kowalski
Institute of Applied Computer Science, Lodz University of Technology
90-924 Lodz, ul. Stefanowskiego 18/22, Poland
Email: t.kowalski@iis.p.lodz.pl

Jacek Wiślicki
Institute of Applied Computer Science, Lodz University of Technology
90-924 Lodz, ul. Stefanowskiego 18/22, Poland
Email: j.wislicki@iis.p.lodz.pl

***Abstract*—*Native functional-style querying extensions for programming languages (e.g., LINQ or Java 8 streams) are widely considered as declarative. However, their very limited degree of optimisation when dealing with local collection processing contradicts this statement. We show that developers constructing complex LINQ queries or combining queries expose themselves to the risk of severe performance deterioration. For an inexperienced programmer, a way of getting an appropriate query form can be too complicated. Also, a manual query transformation is justified by the need of improving performance, but achieved at the expense of reflecting an actual business goal. As a result, benefits from a declarative form and an increased level of abstraction are lost.***

***In this paper, we claim that moving of selected methods for automated optimisation elaborated for declarative query languages to the level of imperative programming languages is possible and desired. We propose an optimisation method for collection-processing constructs based on higher-order functions through factoring out of free expressions in order to avoid unnecessary multiple calculations. We have implemented and verified this idea as a simple proof-of-concept LINQ optimiser library.***

## I. Introduction

SINCE the release of LINQ (Language-Integrated Query) for the Microsoft .NET platform in 2007, there has been a significant progress in the topic of extending programming languages with native querying capabilities [1]. Programming languages are mostly imperative; their semantics relies on the program stack concept. They operate on volatile data and the meaning of collections is rather secondary. On the other hand, query languages are usually declarative and their semantics often bases on some forms of algebras or logics; these languages operate mostly on collections of persistent data. Declarativity of a query language reveals itself mostly when considering operators for collections. In the case of an imperative language, operating on a collection takes a form of an explicit loop iterating over collection elements in a specified order, while in query languages one declares a desired result (e.g., a sub-collection containing elements of a base collection matching a given selection predicate) and an algorithm of filtration itself is not an element of an expression representing the query. Based on characteristics of data structures, a database state and existence of additional auxiliary structures (e.g., indices), an execution environment can choose the most promising algorithm (a plan) for evaluation of the query. Declarativity allows one to postpone selection of an algorithm even to the moment of an actual query execution. In this paper we discuss to what extent solutions for processing of collections within programming languages are actually declarative. To do so, we made an extensive research on query optimisation. In databases it is a crucial process that allows a programmer to be relieved from thinking about details of a processing control flow, auxiliary data structures and algorithms.

LINQ seems to be the most robust solution introducing a promise of declarative collection processing within an imperative programming environment. It is commonly used for direct processing of collections and as a mapper to resources devoid of a robust declarative query API or query optimisation. When encountering performance issues, developers are forced to manually optimise LINQ expressions or partly resign from declarative constructs in favour of an imperative code.

```
var ikuraQuery =
  from p in products
  where (
    from p2 in products
    where p2.productName == "Ikura"
    select p2.unitPrice).Contains(p.unitPrice)
  select p.productName;
```

Listing 1. Example 1 – query expression syntax.

Consider a LINQ query expression in Listing 1 (the database diagram including the Products table is available at *http://northwinddatabase.codeplex.com/*) whose purpose is to find names of products with a unit price equal to a price of a product (or products) named Ikura. If the query addresses a native collection of objects, its execution is severely inefficient as the nested subquery, searching for prices of products named Ikura, is unnecessarily evaluated for each product addressed by the outer query. Although this task could be resolved in a time linearly proportional to the collection's cardinality, the LINQ engine induces an outer loop and a nested loop, both iterating over the products' collection. Using this example, in further sections we show that manual optimisation of complex LINQ queries is not an easy task.

LINQ enables to express the same goal in many different ways. However, evaluation times of two semantically equivalent queries may differ by several orders of magnitude. In particular, in the context of the LINQ query

978-83-60810-66-8/$25.00©2015, IEEE

expressions' declarative syntax, it violates the declarative programming principle. Without knowledge on how a query engine works in a context of given data, the optimisation process is too complex and time-consuming. This is particularly true if a programmer wants to preserve semantics and properties of his query construct.

To the best of our knowledge, the problem of automated global optimisation of LINQ queries for direct processing of collections of objects has not been addressed in the literature so far. By global optimisation we understand the ability to define an efficient query execution plan based on the whole query structure as opposed to the local optimisation that usually only targets a single operator. Below we prove that global optimisation can be done automatically making LINQ genuinely declarative.

Nonetheless, the problem that this paper deals with is not limited to LINQ. Surprisingly, it extends to dozens of programming environments that support functional-style operations on collections of elements, such as filter, map or reduce. Pipelines and streams introduced in Java 8 are a solution equivalent to LINQ to Objects [2]. The main difference lies in the naming convention of new operators corresponding to their functional prototypes (e.g., *map* and *filter* instead of LINQ's *Select* and *Where*). Furthermore, list comprehension constructs are examples of a shorthand syntax for specifying operations of projection and selection (filtering). Consequently, discussed issues concern many imperative languages exploiting this feature (e.g., Python). Fowler summarises such a functional-style programming pattern using a term collection pipeline [3]. Examples given in LINQ can be expressed in many imperative and functional programming languages. While we extend the conclusions of our work to the universe of imperative programming, they do not directly apply to functional languages (e.g., Haskell) since their principles of program evaluation are significantly different [4].

The rest of the paper is organised in the following way. First, we present a brief description of the state of the art followed by characteristics of language-integrated query constructs. Next, we describe issues with nested independent subqueries and free expressions revealing a huge optimisation potential. Finally, we present our solution followed by measured results and principles of our optimisation approach, being the core of the paper. The paper is concluded with a short summary.

## II. Related Work and the State-of-the-Art

Databases are the area of the computer science where declarative programming and query optimisation have developed extensively. Over 40 years of the research on relational systems resulted in various optimisation techniques [5][6] and numerous solutions are incorporated in available commercial products. Our research presented in the paper particularly addresses query optimisations analogous to query unnesting, dating back to the early 80s [7]. This topic is constantly appearing in the context of arising database technologies. Different approaches to handle nested queries evaluation have been proposed for object-oriented databases [8][9] and XML document-based stores [10]. However, NoSQL solutions marginalise the topic of query languages and usually rely on a minimalistic programming interface and domain-specific optimisations, mostly implemented by high redundancies and storing data in the form matching assumed queries. Most attention from the scientific community concentrates on the topic of distributed data-parallel computing using the Map/Reduce paradigm (like Hadoop or Dryad for Azure). This paradigm can be transparently used in declarative collection processing. The Dryad programming environment based on LINQ [11] takes advantage of mechanisms similar to Map/Reduce in order to write scalable, parallel and distributed programs. To increase sharing of computations in a data centre, Dryad can benefit from the Nectar system [12]. It is able to cache results of frequently used queries and incrementally update them. The use of cached results is achieved through automatic query rewriting. Robust query and program optimisations have been developed for solutions based on the functional paradigm. According to Fegaras [13], an optimisation framework for a functional lambda-DB object-oriented database relies on mathematical bases, i.e. the monoid comprehensions calculus. It generalises many unnesting techniques proposed in the literature.

Glasgow Haskell Compiler (GHC) for the Haskell non-strict purely functional language introduces many methods based on code rewriting. They range from relatively simple rules that can be used to improve efficiency of programs through modifications on a high syntactic level to more complex low-level core language transformations (e.g., let-floating, beta reduction, case swapping, case elimination) [14]. In particular, a procedure called full laziness (or fully lazy lambda lifting) has been proposed to avoid reevaluation of inner expressions for which result could be pre-calculated only once [15][16].

Currently, due to introduction of lambda abstractions into object-oriented languages, functional style of programming became ubiquitous. Stream and collection processing constructs derived from functional languages can be naturally evaluated in parallel using multiple processor cores. Therefore, the most popular solutions, like Java 8 streams, LINQ or ScalaBlitz, enable such optimisation through various libraries or frameworks [17].

In the field of functional-style queries integrated into a programming language, the topic of query optimisation seems the most advanced in LINQ. A LINQ provider library can implement direct processing of data (e.g., LINQ to Objects, LINQ to XML) or delegate processing to a remote external resource by sending a translated query (e.g., LINQ to SQL, LINQ to Entities). To be precise, a mixture of both approaches can be used, e.g. when the query language of a remote resource cannot completely express the semantics of a LINQ query. In the case when LINQ sends a translated query, it also delegates the responsibility for query optimisation. Consequently, if the external resource engine provides optimisation, developers can fully rely on a declarative style of programming. However, in the context of LINQ to SQL, the problem of analysing and normalising of LINQ queries in order to provide minimal and cohesive mapping to SQL has drawn attention of the scientific community. This is caused mostly by some drawbacks of the original Microsoft's solution that in some cases may fail or produce a so-called "query avalanche" [18][19].

The issue of performance deficiencies while processing collections of objects has not passed unnoticed by the LINQ community. In order to cope with the shortage in optimisation comparing to database engines, the i4o project (abbr. index for objects) solution adapted the idea of indexing to native objects' collections [20]. It is implemented as an alternative for the LINQ to Objects provider library. Utilising the concept of secondary access structures, i4o can produce several orders of magnitude of a gain in performance for queries filtering data at the cost of a data modification overhead.

Another examples of LINQ query optimisation tools are Steno [21] and LinqOptimizer [22] provider libraries. Their authors focused on a significant performance deficiency of LINQ queries in contrast to the equivalent manually optimised code that can be several times faster. Experiments have shown that Steno allows one to obtain up to 14-fold increase in processing of sequential data and 2-fold comparing to a problem processed by the DryadLINQ distributed engine [11]. The main idea behind Steno is to eliminate the overhead introduced by virtual calls to iterators that are the fundamental mechanism used by the LINQ engine. This problem has been solved by automatic generation of an imperative code omitting iterators. The optimisation addresses mainly implementation of individual operators. This also concerns the case of nested loops' optimisation when Steno has to analyse a series of operators only to preserve the order of iteration induced by the LINQ to Objects library implementation. This is justified by the loop fusion efficiency and consideration of side effects that are allowed in LINQ. Steno is also capable of higher-level optimisation giving an example of the GroupBy-Aggregate optimisation. It involves a local term rewriting, addressing a pair of neighbouring operators, i.e. GroupBy followed by Aggregate. When encountering such a sequence of operators, Steno replaces it by a dedicated GroupByAggregate operator that saves memory by storing per-key partial aggregates instead of the whole collection of group values. This optimisation takes advantage of LINQ declarativity by changing the course of evaluation. As a result, introducing side effects would cause its incorrectness. Being aware of a difficulty of automatic reasoning about side effects within queries, Steno's authors suggest developer-guided optimisation. Optimisation similar to GroupBy-Aggregate is considered in the SkyLINQ project [23] that develops an alternative operator called Top. This operator can be used to substitute a sequence of OrderBy and Take method calls (i.e. an operation to get top k elements). The significance of LINQ grew up with introducing LINQ to Events, an extension enabling declarative programming according to the reactive paradigm [24]. The solution derives from Functional Reactive Programming and is well suited for composing asynchronous and event-based programs [25]. Recently, this approach has attracted attention of commercial and scientific communities and, as a programming paradigm, faces efficiency issues indicating possible areas for optimisation [26][27].

Other current research on LINQ strives to allow seamless integration of heterogeneous data sources [28]. As a result, users can transparently process and modify data shared among contributing resources. Because of complex multilayer architecture, such an environment is not efficiency-oriented. LINQ is generally focused on local optimisation performed at a data source layer. In processing of heterogeneous and distributed data, it is unlikely that such optimisation is provided by each contributing resource. Therefore, it raises a need for global optimisation performed at the level of a LINQ query itself.

Declarative functional-style constructs in general-purpose object-oriented languages are not pure. As a result, decisions concerning optimisation have to be made by programmers. Transparent and aggressive compile-time optimisations can be achieved by introducing a query language extension into a programming language compiler [29].

One of numerous examples of extending compilers of existing languages with declarative constructs is SBQL4J [30]. It enables seamless integration of SBQL queries with language instructions and executing them in a context of Java collections. SBQL4J is based on the Stack-Based Architecture (SBA) approach instead of the functional approach and offers capabilities comparable to the LINQ technology [31][32]. What distinguishes it from other programming language-integrated queries is incorporation of several automatic optimisation methods developed for SBA. One of these methods, i.e. factoring out independent subqueries [8], enables SBQL4J to cope with optimisation of queries equivalent to examples discussed in this paper. It belongs to the group of optimisation methods that are based on query rewriting. Factoring out concerns a subquery (that in SBA represent any subexpression) that is processed many times in loops implied by so called non-algebraic operators despite that in subsequent loop cycles its result is the same. In SBQL4J rewriting is applied at a compile-time and a resulting performance improvement can be very significant, sometimes giving query response times shorter by several orders of magnitude.

## III. Characteristics of Language-Integrated Query Constructs

Declarative style programming (especially in the context of databases) is often associated with the select-from-where syntactic sugar known from SQL that was adapted into LINQ. The query in Listing 1 is expressed using the LINQ query expression syntax. That form lacks explicit information on an order of performed operations and virtually a compiler could translate it to any semantically equivalent lower-level code that could be considered a query execution plan. Consequently, programmers must be particularly careful about potential side effects within declarative constructs in order to avoid the risk of unpredicted violations. Technically, query expressions are syntactic sugar over the implementation layer using lambda expressions, higher-order functions and, so called, extension methods [33]. An executable query, after removing the LINQ syntax sugar, will take the form presented in Listing 2.

```
var ikuraQuery = products.
  Where(p => products.
    Where(p2 => p2.productName == “Ikura”).
    Select(p2=>p2.unitPrice).Contains(
      p.unitPrice)).
  Select(p => p.productName);
```

Listing 2. Example 1 – de-sugared.

The translated query uses the traditional, non-declarative object-oriented programming syntax. When processing collections or XML documents directly, the most crucial LINQ library extension methods (e.g., Select and Where) expose iterators that perform a specified operation on elements of a given collection. Lambda expressions are used to express details concerning such an operation, e.g. the selection predicate for the Where operator. Despite of similarity of Listing 2 to the original query expression, such composition of method calls on the products collection determines the order of evaluation.

Due to the specific implementation based on iterators and lambda abstractions, the execution strategy of LINQ queries is deferred. Execution is performed in presence of functions or instructions forcing iteration over elements specified by a query. However, a result of an iteration is not saved or cached, so each execution reevaluates a query against a given (current) data state.

The approach used in the LINQ to the objects' library implementation is generally ubiquitous (however, not uniform) in numerous programming languages (e.g., Python, Java 8, Elixir, Ruby) [3]. A good summary describing the possible set of properties of functional-style constructs can be found in the documentation of Java 8 streams [2].

The functional nature makes a language construct a good candidate for optimisation due to intelligible querying. All operations in a query processing chain produce a new queryable result instead of modifying original data; hence they do not introduce side effects. For example, during filtration on a list, no element is actually removed. Even though filter and map (common functional-style operators) are often used to directly process elements of a local in-memory collection, in reality elements can be obtained one by one from any so called queryable data source, e.g., a data structure, a generator function, an iterator, an I/O channel and a chained pipeline of collection operations. Such generality allows, usually time-consuming, querying of remote data sources, additionally making optimisation desirable.

The above properties are common in implementations of language-integrated query mechanisms. However a programmer must be sensitive to possible differences in various programming languages. For example, some languages implement consumable evaluation of queries. In such a strategy, elements of a queryable data source instance can be visited only once during its life. As a result, each query instance can be evaluated only once. The last property is present in Java 8 streams whereas LINQ operators are not consumable. The laziness-seeking property has the most profound impact on evaluation and semantics of language-integrated queries. It is connected with the lazy evaluation strategy assuming that a next element is returned for further processing only if necessary. Usually, a place of a lazy construct definition does not determine an actual moment of query execution (i.e. deferred execution strategy). Actual query execution occurs when its result is required, for example elements referred by a query are iterated or counted. Operations like selection, projection, and removal of duplicates are often implemented lazily. Consequently, to ensure coherence execution of eager constructs (e.g., grouping or ordering) is also deferred. Lazy evaluation usually results in better performance. It is cache-friendly since an element is processed by a chain of collection pipeline operations before proceeding to the next element. Moreover, in cases when the desired query result has been reached before visiting all elements it is not necessary to continue iterating (e.g., a query finding the first product with name “Ikura”).

In the context of query optimisation, it is important to preserve properties of optimised constructs. In a general case, any change in this matter can affect semantics of application code. Switching an expression evaluation strategy from lazy to eager or forcing immediate execution can have serious consequences. Only laziness-seeking constructs can deal with possibly unbounded data sources. Eager evaluation of selection and projection operators on an infinite data source would require infinite computational resources and time, while lazy evaluation can return partial results.

In the next section, we show that preserving deferred execution, which is implied by the lazy evaluation strategy, is the factor impeding query optimisation.

## IV. Performance Pitfalls

### A. *Evaluation of Independent Subqueries*

Analysing the expression from Listing 2, it becomes obvious that the nested query selecting products named Ikura will be executed multiple times, since it is a part of a lambda abstraction (specifying a selection predicate) called against each product (the external *Where* operator induces a loop iterating over elements of products collection). This form is not efficient and makes the computational complexity quadratic (i.e. $O(n^2)$). However, searching for products named Ikura is independent of the parent query and could be evaluated just once. In order to improve query performance, a programmer must transform it. A natural way for optimisation seems to be factoring out the problematic subquery to a separate instruction and assigning it to a new variable (see Listing 3). The changes could also be presented on the LINQ query expression, but because the form with extension methods is actually executed, it will be a basis for this study.

```
var nestedQuery = products.
  Where(p2 => p2.productName == "Ikura").
  Select(p2=>p2.unitPrice);
var ikuraQuery = products.
  Where(p => nestedQuery.Contains(p.unitPrice)).
  Select(p => p.productName);
```

Listing 3. Example 1 – loops in separate instructions.

A result of the transformation shown in Listing 3 may seem effective; however, the expected goal will not be achieved. The problem lays in the execution strategy of LINQ queries. The *nestedQuery* variable holds an instance of a non-executed query that will be evaluated – like in the case of the non-transformed expression (Listing 2) – at every traversal of the loop induced by the *ikuraQuery Where* operator.

In Java 8 streams proper execution of corresponding queries generally would become impossible due to the consumable property of streams. After the transformation, the selection predicate of the *ikuraQuery* would share the same instance of a *nestedQuery* stream. Evaluation of the nested query would be performed only once, at the first traversal of the loop induced by the *ikuraQuery Where* operator, whereas the following iterations would result in terminating query evaluation and throwing an exception.

Solving the above problems requires eliminating deferred execution of the nested query. There exist several techniques to force immediate execution of a LINQ query. For example, the *ToList* method returns a list containing a materialised query result. Applying it to the nested query makes the solution more efficient (linear computational complexity) than the query in Listing 2. However, a part of the original query is executed and the other part remains deferred to the moment of an actual demand. It is possible that data in a collection may change between creation of a query and its evaluation. The original query form (and the programmer's intention) is insusceptible to it – the query is always completely executed on a current data state. After immediate execution of the nested query one cannot be sure about it – the *ikuraQuery* can be evaluated when data needed for calculating the *nestedQuery* subquery got already modified. As a result of the transformation, there occurred a change of the query semantics that is very difficult to detect by a programmer or tests. Ultimately, a programmer is forced to resign from deferred execution of the whole query, which is shown in Listing 4.

```
var nestedQuery  = products.
  Where(p2 => p2.productName == "Ikura").
  Select(p2=>p2.unitPrice).ToList();
var ikuraQuery = products.
  Where(p => nestedQuery.Contains(p.unitPrice)).
  Select(p => p.productName).ToList();
```

Listing 4. Example 1 – fully immediate execution.

Due to explicit materialisation, reusing the optimised query against a different data state becomes troublesome. For an inexperienced programmer, a way of getting an appropriate query form can be too complicated. Without deeper knowledge on the LINQ internal semantics in a context of object data, obtaining an optimal structure of code is a tricky, time-consuming and error-prone task. The example shows a lack of real independence of LINQ from a type of a data source. Despite the fact that LINQ allows unified processing on various types and sources of data, an actual execution plan relies on them. It seems that the basis for elaborating this layer of the language was mostly the integration of the object-relational mapping with a type system of a programming language (what also shows at the level of the LINQ query expression syntax and implementing providers [30]).

### B. Factoring Out Constructs Executed Immediately

Although LINQ queries execution is deferred, an execution strategy of some expressions comprising LINQ queries can be immediate. Such expressions are evaluated locally in the place of the definition. Some operator, like aggregate functions returning a single value instead of a queryable data source, force immediate execution of a query. The query in Listing 5 contains such an expression determining the greatest unit price in the products' collection. However, it will not be evaluated until the execution of the *maxQuery* since it is contained in a lambda expression defining a selection predicate for the *Where* method.

```
var maxQuery = products.
  Where(p => products.
    Max(p2=>p2.unitPrice) == p.unitPrice).
  Select(p => p.productName);
```

Listing 5. Example 2 – extension methods syntax
(quadratic computational complexity).

Similarly to the subquery determining the price of the Ikura product, it should be evaluated only once during execution of the *maxQuery* and therefore needs to be factored out. Let us call such constructs free expressions. Using the same procedure as presented in the previous section, we break the query into two instructions and perform immediate execution (see Listing 6).

The *ToList* operation does not need to be applied to the expression defining maxPrice (actually, it cannot be applied because it returns a value), due to its inherent immediate execution.

```
var maxPrice = products.Max(p2=>p2.unitPrice);
var maxQuery = products.
  Where(p => maxPrice == p.unitPrice).
  Select(p => p.productName).ToList();
```

Listing 6. Example 2 – immediate execution
(linear computational complexity).

### C. Consequences of Changing the Evaluation Order

There exist some subtle consequences concerning evaluation after the manual optimisation. In the original forms of example queries (Listing 2 and Listing 5), nested expressions would be evaluated only if the products' collection is not empty. In the optimised forms (Listing 4 and Listing 6) it will be unnecessarily evaluated also when the collection is empty. This is particularly important for performance when a nested query operates on a collection different than the external query does. The current example concerns just one collection, but it is easy to imagine a situation when collections are distinct (e.g., products from other shops kept in separate collections). In extreme cases, if calculation of a factored out expression is time-consuming, this can worsen overall query performance. Aside from performance issues, the transformation presented in previous sections can have dangerous impact on query semantics. In the second example (Listing 5) in the case of an empty collection of products, the selection predicate is not evaluated at all and the final result is simply an empty collection of product names. After optimisation (Listing 6), the expression *products.Max(p2 => p2.unitPrice)* is always evaluated at the beginning. The *Max* method applied to an empty collection throws an exception. Consequently, the behaviour of the optimised query is unsafe and inconsistent with the intent of a programmer.

To make optimisation immune to the described risk, the original order of evaluation should be restored. This could be achieved by applying the lazy loading pattern to the free expression determining *maxPrice*. In Listing 7 we introduce an improved transformation.

```
var maxPriceThunk =
  new Lazy<Double>(products.Max(p2=>p2.unitPrice));
var maxQuery = products.
  Where(p => maxPriceThunk.Value == p.unitPrice).
  Select(p => p.productName).ToList();
```

Listing 7. Example 2 – immediate execution (linear computational complexity).

A *Lazy* class instance is a simple thunk – an object in memory representing an unevaluated (suspended) computation, used in the call-by-need evaluation strategy. The argument of the Lazy constructor specifies a function that should be evaluated at most once, only if its value is requested for the first time. The request is signalised by accessing the *Value* property of the *Lazy* instance. Consequently, the original order of the query and the free expression evaluation are restored (except the free expression being processed at most once) making the optimisation semantically safe. It is achieved at the expense of overhead concerning access to the *Value* property.

In the next sections we present a general approach to optimisation that preserves semantics and characteristics of an original query while reducing its computational complexity.

## V. Factoring Out Free Expressions

The solutions presented for both examples share a common shortcoming; they do not preserve the deferred execution property. Our main aim is to propose a general query rewriting rule overcoming the problem. In order to keep the solution generic, additional constraints have been assumed: (1) transformation should not break a query into separate instructions (in contrary to what is shown, for example in Listing 7), (2) we express the rules in general terms rather than LINQ specific ones (e.g., operators common in functional programming). Obviously, we assume also that the intent of a programmer is simply to query and not to introduce side effects deliberately.

Generalisation of the factoring out procedure should take into account queries more complex than presented above. A nesting level of a lambda expression in the examples presented in Listing 2 and Listing 5 is shallow, but conceptually can be arbitrary with no need to modify the factoring out procedure. Free expressions can be bound either globally, i.e. to an environment independent from a query, or to a lambda expression at any nesting level lower than the lambda expression containing a free expression. The examples presented in Listing 2 and Listing 5 concern the former case. A generalising solution to the latter case can be achieved by treating any subquery as a separate query and the rest as a global environment.

The basic idea behind the transformation is, first, to identify free expressions that could be evaluated before a loop induced by an operator containing them, and next, to apply an appropriate rewriting rule. This is generalisation of the standard procedure called loop-invariant code motion known from the compilation theory [34]. An example of incorporating this idea to programming language-integrated queries can be found in the Stack-Based Architecture theory [8]. To optimise evaluation in functional languages, a similar procedure of fully lazy lambda lifting (called also full laziness) has been also proposed [16]. In both cases rewriting rules are straightforward and make use only of the basic set of language operators. Our attempt to generalise, in a similar manner, factoring out free expressions within LINQ queries using only methods supplied by LINQ has been unsuccessful. In particular, LINQ operators in presence of a queryable data source (e.g., a collection) cause iteration over elements, whereas factoring out requires treating empty, single or multiple elements as an individual result cached for reuse in further calculation.

### A. Formalising Optimisation

The procedure of factoring out free expressions can be applied to the following query pattern:

*queryUnoptimized* ::= *queryExpr*(λ(***freeExpr***))

where *queryExpr* denotes a query expression that includes a nested lambda abstraction λ(*freeExpr*) containing *freeExpr* that is free from any lambda abstraction within the query. Additionally, we assume that *freeExpr* should be evaluated several times during the execution in order to make factoring out profitable. This pattern is not restricted to a whole query. It can match any subquery.

The solution requires introducing transformation of factoring out a free expression before a loop using it and applying the lazy loading evaluation strategy. Several aspects need to be addressed to make such optimisation effective and general. (1) In imperative programming languages deferring execution is often achieved through enclosing code in a function or by introducing an iterator. In both cases, repeated execution (e.g., inside a loop implied by map or filter collection pipeline operators) causes repeated evaluation. If this applies to a factored out expression, then it is usually necessary to force materialisation of its result before entering a loop using it. (2) Moreover, materialisation solves the issue with factoring out consumable data sources since they cannot be evaluated more than once. Before factoring out, the problem does not exist since such constructs reside inside lambda abstractions (that are parameters of collection pipeline operators) and therefore are evaluated only once during single lambda call evaluation. (3) As stated earlier, it is possible that a free expression is skipped during evaluation of the original query. In a general case it is safe to preserve an order of evaluation by suspending materialisation of a factored out expression and prevent its immediate execution before entering a loop using it. (4) An instance of a mechanism used for suspending materialisation of a factored out expression should not be shared between query executions. To solve this problem, it can be additionally enclosed in a lambda abstraction. Otherwise, following executions would share a cached result determined during the first execution. (5) In order to prevent collection pipeline operators from iterating over a collection, it has to be nested into a new collection as a single element. The same procedure can be applied to a single result to enable usage of collection pipeline operators.

Let us denote the following abstract operations: ***Collection***(*arg*) – creates and returns a collection consisting of one element specified by an argument, e.g., if an argument is a collection it returns a nested collection), ***Immediate***(*expr*) – evaluates and materialises a result of an expression passed as an argument (except when the *expr* execution strategy is already immediate), ***Suspend***(*lambda*) – returns an instance of a mechanism for lazy loading of an expression specified by a lambda abstraction passed as an argument, ***Value***(*lazy*) – returns a lazily initialised value stored by a lazy loading mechanism instance specified by an argument.

Taking advantage of above operations, we introduce the following rewriting procedure:

*queryOptimized* ::=
***Collection***(() => ***Suspend***(() => ***Immediate***(*freeExpr*))).
map(*lambdaParam* => *lambdaParam*()).
map(*freeExprThunk* => *queryExpr*(
λ(***Value***(*freeExprThunk*)))).flatten()

where λ(***Value***(*freeExprThunk*)) is a nested lambda abstraction λ(*freeExpr*) with an occurrence of *freeExpr* expression substituted by ***Value***(*freeExprThunk*). This form ensures that execution of all components of the original query is deferred assuming that collection pipeline operators map and flatten have such an execution strategy.

The first part of the rewritten query
***Collection***(() => ***Suspend***(() => ***Immediate***(*freeExpr*)))
creates a collection consisting of a single element that is a lambda function creating an instance of a mechanism for suspended materialisation of the factored out free expression. The following map operator ensures execution of the lambda function. As a result, the following map operator will process
*queryExpr*(λ(***Value***(*freeExprThunk*)))
expression only once for *freeExprThunk* assigned the lazily loaded cached value of *freeExpr*. Therefore, the result of evaluation of *queryExpr*(λ(***Value***(*freeExprThunk*))) is equal to the result of evaluation of *queryExpr*(λ(*freeExpr*)). The flatten operator eliminates an outer collection implied by the *Collection* operator. Consequently, the final result of the optimised query is taken from evaluation of the
*queryExpr*(λ(***Value***(*freeExprThunk*)))
expression.

### B. Implementing Optimisation in C#

In C#, to simplify optimisation we introduce an auxiliary method *AsGroup* to take care of the *Collection* operation and suspended evaluation of a lambda expression returning a materialised value of the free expression. Listing 8 shows the implementation of the auxiliary operator.

```
static IEnumerable<TSource> AsGroup<TSource>(
                    Func<TSource> sourceFunc) {
  yield return new Lazy<TSource>(sourceFunc);
}
```

Listing 8. Auxiliary optimisation method.

The ***Suspend*** operation is achieved by a Lazy class constructor ***new*** *Lazy<TSource>*(*sourceFunc*). The ***yield*** return statement is a syntax sugar enabling creating a collection available through an iterator deferring any computations until iteration starts. In this way, a programmer avoids using a concrete type of a collection and enables a compiler to choose the best implementation on its own. *AsGroup* exposes an iterator that returns only one element, i.e. an instance of a mechanism for suspended materialisation of the factored out free expression. It is created directly before yielding replacing the projection *map*(*lambdaParam* => *lambdaParam*()). Consequently, the rewritten query in case of LINQ takes the following form:

*LINQ-deferredQueryOptimized* ::=
*AsGroup*(() => ***Immediate***(*freeExpr*)).
*SelectMany*(*freeExprThunk* =>
*queryExpr*(λ(*freeExprThunk.Value*)))

where the *SelectMany* LINQ operator substitutes map and flatten and *freeExprThunk.Value* realises the *Value*(*freeExprThunk*) operation.

The above transformation can be adapted to a situation when *queryExpr*(λ(*freeExprThunk.Value*)) is a construct executed immediately (e.g., when it returns a single value). In that case *SelectMany* needs to be replaced with two operations: Select realising projection and *First* responsible for flattening and immediate execution:

*LINQ-immediateExpressionOptimized* ::=
*AsGroup*(() => ***Immediate***(*freeExpr*)).
*Select*(*freeExprThunk* =>
*queryExpr*(λ(*freeExprThunk.Value*))).*First*()

The ***Immediate*** operation is required only in the case when *freeExpr* is a LINQ query deferred in execution. Explicit materialisation can be achieved using LINQ specific methods, e.g., *freeExpr.ToList*(). The transformation constitutes the general rewriting rule for optimisation of LINQ queries through factoring out free expressions. Applying it to the examples from Listing 2 and Listing 5 is shown in Listing 9 and Listing 10, respectively.

```
var ikuraQuery =
  AsGroup(() => products.
    Where(p2 => p2.productName == “Ikura”).
    Select(p2=>p2.unitPrice).ToList()).
  SelectMany(ikuraPriceThunk => products.
    Where(p => ikuraPriceThunk.Value.
    Contains(p.unitPrice)).Select(p => p.productName));
```

Listing 9. Example 1 – after factoring out suspended free expressions optimisation.

```
var maxQuery =
  AsGroup(() => products.Max(p2=>p2.unitPrice)).
    SelectMany(maxPriceThunk =>
      products.Where(p =>
            maxPriceThunk.Value == p.unitPrice).
      Select(p => p.productName));
```

Listing 10. Example 2 – after factoring out suspended free expressions optimisation.

The queries execution strategy after optimisation remains deferred and in the case of the second example (Listing 10), the problem of the exception while addressing an empty products' collection does not occur.

## VI. Performance Tests

We have evaluated the impact of factoring out of free expressions optimisation in C# by applying it manually to a number of problems: ***samePriceAs*** – given a collection of products, find products with the same price as the product specified by a name, ***maxPrice*** – given a collection of products, find products with the maximal price in the collection, ***promoProducts*** – given a collection of products, find names of products in the imaginary sale promotion, i.e. exactly *k* times more expensive than any other product, and ***pythagoreanTriples*** – from natural numbers between 1 and n find a number of triples satisfying the Pythagorean theorem.

In experimental tests, the collection of products ranged from 1 to 1,000,000 elements. The size of each product averaged to 175 bytes. Tests for ***samePriceAs***, ***maxPrice***, ***promoProducts*** and ***pythagoreanTriples*** problems have been conducted using queries in Listing 2, Listing 5, Listing 11, and Listing 12 accordingly. The problems have been solved relatively simply and each one has at least one free expression suitable for the factoring-out optimisation. Solutions to ***samePriceAs*** and ***maxPrice*** have free nested queries, whereas ***promoProducts*** and ***pythagoreanTriples*** introduce simple mathematical calculations that can be factored out. The tests include comparison with PLINQ and LinqOptimizer optimisation framework. We also combine them manually with our optimisation to explore limits and further opportunities.

```
var promoProducts =
  products.Where(p => products.
     Any(p2 => p2.unitPrice ==
             Math.Round(p.unitPrice / 1.2, 2))).
  Select(p => p.productName);
```

Listing 11. A query concerning the ***promoProducts*** problem before optimisation.

```
var pythagoreanTriples =
  Enumerable.Range(1, max + 1).SelectMany(a =>
    Enumerable.Range(a, max + 1 - a).SelectMany(b =>
      Enumerable.Range(b, max + 1 – b).Where(
        c => a * a + b * b == c * c))).Count()
```

Listing 12. A query concerning the ***pythagoreanTriples*** problem before optimisation.

We conducted our experiments on a workstation with a 4-core Intel Core i7 4790 3.6 GHz processor, 32 GB of DDR3 1600MHz RAM, hosting Windows Server 2012 R2. Benchmarks have been compiled for a x64 platform with enabled code optimisations using target .NET Framework v. 4.5. Tests results for following problems are presented in Fig. 1, Fig. 2, Fig. 3 and Fig. 4.

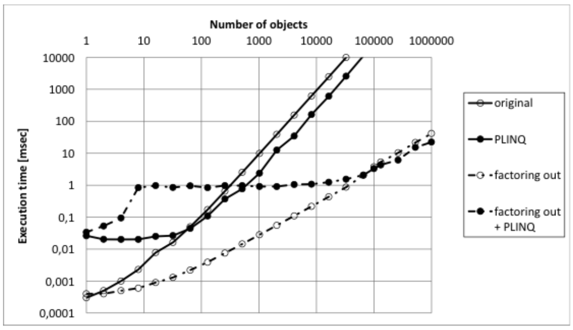


Fig 1. Query evaluation times for ***samePriceAs*** problem.

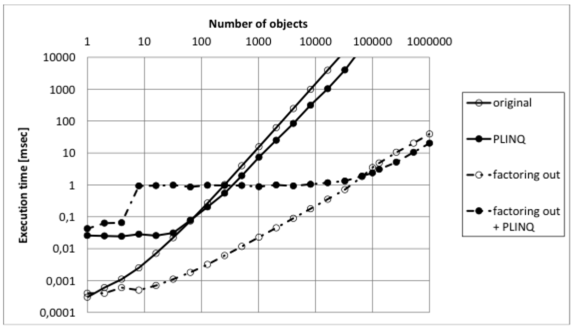


Fig 2. Query evaluation times for ***maxPrice*** problem.

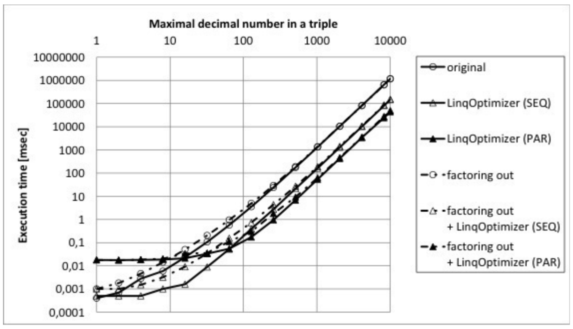


Fig 3. Query evaluation times for ***pythagoreanTriples*** problem.

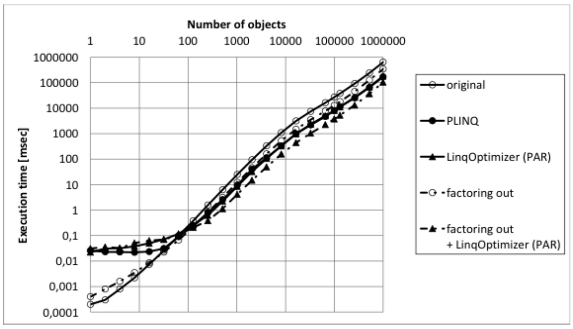


Fig 4. Query evaluation times for ***promoProducts*** problem.

The LinqOptimizer is used in two variants: sequential (denoted by SEQ) and parallel (denoted by PAR). The latter competes with PLINQ. Each query before and after factoring-out optimisation has been subjected to three further optimisation variants, i.e. PLINQ, LinqOptimizer sequential or parallel variant. The tests focus on query execution times and omit optimisation and compilation of a query. Most of the plots use logarithmic scales to more clearly reveal differences in performance for various collection sizes. To improve readability, the plots omit optimisation variants that are generally worse. In particular, the sequential variant of LinqOptimizer is shown only if it improved query performance in any collection size range, and the better alternative between PLINQ and parallel variant of LinqOptimizer is selected.

Results of the tests are as follows:

- Tests' results are consistent with an expected computational complexity. In ***samePriceAs*** and ***maxPrice*** problems it has been reduced from quadratic to linear, achieving a gain in orders of magnitude for large collections, e.g., in the case of the second example (Listing 5 and Listing 10) the query after factoring out is more than 30,000 times faster for 100,000 products (boost from ~115 s to ~3.8 ms).
- Except for the ***pythagoreanTriples*** problem, the profitability threshold of individual factoring-out optimisation is very low when comparing to PLINQ and LinqOptimizer. Even for a collection of 2 objects, optimised queries can work faster than original ones (e.g. ***samePriceAs*** and ***maxPrice***).
- The performance penalty in the case of a collection consisting of a single element is at most 0.6 μs which corresponds to a ~60% deterioration (the ***pythagoreanTriples*** problem).
- When processing large collections, the factoring-out transformation can give several times better performance by taking advantage of PLINQ (especially in the case of the ***promoProducts*** problem). For smaller collections, PLINQ imposes overhead significantly greater than factoring out.
- The ***pythagoreanTriples*** problem optimisation tests show that it may be difficult to obtain a significant gain when factoring out a simple expression (i.e. a * a + b * b). A ~3% gain is achieved for n equal to 10,000. • The LinqOptimizer framework seems to be designed for optimising queries involving numbers rather than complex objects. Only in the ***pythagoreanTriples*** problem optimisation, it outperforms both PLINQ and factoring out.
- In general, combining factoring out of free expressions with LinqOptimizer is not likely to produce the best solution. However, it seems that tuning of the LinqOptimizer algorithm should be possible. In the ***pythagoreanTriples*** problem, PLINQ is able to produce more efficient query after factoring out, whereas LinqOptimizer favours the original query. Unfortunately, the differences are too small to be seen on the plot.

C# libraries offer a *Lazy* class realising the Suspend operation, but considering performance, we have implemented our own lightweight version. We have experimented with different variants of performing *Collection*, *Suspend* and *Immediate* operations but the presented solutions generally resulted in performance better than others.

## VII. Automatic Optimisation

### A. Free Expression Detection

The transformation is justified by the need to increase effectiveness, which is achieved at the expense of reflecting the business goal. As a result, benefits from a declarative form and an increased level of abstraction are lost.

LINQ expression trees enable run-time analysing and dynamic building of LINQ queries [36]. This feature allows developing an optimisation method relying on rewriting of a LINQ abstract syntax tree. Automated detection of specified query patterns and transformation to an optimised form are required to make LINQ queries truly declarative. The previous part of this paper deals with the latter, i.e. the definition of efficacious rewriting rules for factoring out of a free expression. This section describes an algorithm for detection of free expressions within a query. The procedure does not address any details of implementation for the LINQ platform. It is general in terms of functional-style programming.

Let us establish a set of definitions concerning expressions and lambda abstractions (inspired by the definitions introduced by Hughes [16]):

- **Def. 1 (bound variables of lambda)**. An occurrence of a variable within lambda $\lambda_A$ is bound to $\lambda_A$ if and only if it is a parameter of $\lambda A_{A,}$
- **Def. 2 (bound expressions of lambda)**. An expression within lambda $\lambda_A$ is bound to $\lambda_A$ if and only if it contains a variable bound to $\lambda A_A$.
- **Def. 3 (native lambda of expression)**. The innermost lambda in which an expression *e* is bound is its native lambda. Let us denote this lambda *n*λ(e),
- **Def. 4 (free expressions in lambda)**. An expression *e* within lambda $\lambda_A$ is free in lambda $\lambda_A$ if $\lambda_A$ is nested in native lambda of expression *e*.
- **Def. 5 (maximal free expressions)**. A maximal free expression (MFE) is a free expression of some $\lambda_A$ that is not a proper subexpression of another free expression of $\lambda_A$.

Additionally, to simplify definitions and the algorithm description, we assume that names of variables are unique. Moreover, we implicitly treat a whole query as a lambda abstraction with all free variables (constituting a global environment) as its parameters. In the case of examples from Listing 2 and Listing 5 native lambda of each MFE is the whole query.

From the definitions above, it follows that any MFE e free in a lambda $\lambda_A$ can be determined before $\lambda_A$ evaluation. Precisely, it could be determined anytime during evaluation of *n*λ(*e*). The above statement is correct since:

1. *e* is a free expression (see definition 5).
2. $\lambda_A$ is inside *n*λ(*e*) (see definition 4).

3. *e* is not bound to $\lambda_A$ (see definition 3).
4. *e* does not contain variables bound to $\lambda_A$ (see definition 2).
5. $\lambda_A$ call does not introduce any variable (parameter) required by *e* (see definition 1) that makes *e* independent from $\lambda_A$.

Consequently, it is possible to factor out the expression *e* from $\lambda_A$ and evaluate it at the level of the *n*λ(*e*) lambda.

The algorithm uses the standard depth-first search approach and detects all MFEs during a single pass through a query expression tree. Expression visitation focuses on finding its bindings that we define as a set of lambda abstractions declaring variables (usually as lambda parameters) used in the expression. This information is further used to determine bindings of its parent. Usage of lambda abstraction parameters determines whether an expression is free or bound. Therefore, it is necessary to handle information about names of the parameters and lambda abstractions to which they are bound. This is a task of an auxiliary map called binders. To correctly manage parameters' binding, the procedure specifically handles lambda abstractions and terminal name binding expressions.

While visiting lambda abstraction, the binders' map is filled with its parameters. They are visible only within the lambda abstraction. This sets the right context for the recursive visitation of the lambda body in order to detect free expressions bound specifically to the current lambda. Finally, the bindings set is returned to the lambda parent except for the current lambda that is removed (information on binding to the current lambda is not relevant outside).

The binders' map is used when visiting name-binding terminal expressions. These expressions consist only of an identifier name. If a name is found in the binders' map, a corresponding lambda is returned (as a single-element bindings set). If a name is not bound to any lambda, then it is assumed to be a globally free variable.

The described behaviour does not concern a name on the right hand side of a member access operator (e.g., field names). Such a name is bound locally to its left side, therefore field member access bindings are inherited from their left side expression. In general, bindings for remaining types of expressions are simply inherited from their children (a sum of the sets).

In the implementation nesting level annotations for lambda abstractions and variables are introduced to simplify the binding analysis. Expression bindings provide sufficient information to determine all MFEs and their native lambdas.

To exemplify the algorithm let us consider the promoProduct problem shown in Listing 11. The query in its optimised form is presented in Listing 13. The expression determining a price *Math.Round*(*p.unitPrice* / 1.2, 2)) is unnecessarily evaluated multiple times during execution of the inner loop implied by the Any operator. What distinguishes this and previous examples is that the transformation applies not to the whole query but only to the *Where* predicate. Additionally, the predicate is not a LINQ query but an expression returning a Boolean value. Therefore, *Select* and *First* methods were used instead of *SelectMany*.

```
var promoProductsOptimized =
  products.Where(p =>
    AsGroup(() => Math.Round(p.unitPrice / 1.2, 2)).
    Select(priceThunk =>
      products.Any(p2 => p2.unitPrice ==
                   priceThunk.Value)).First()).
  Select(p => p.productName);
```

Listing 13. Example 3 – rewriting inside lambda abstraction.

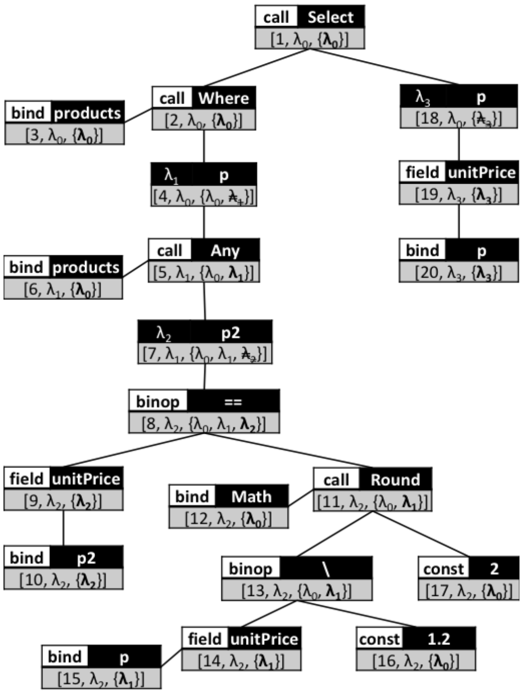

call Select
[1, $\lambda_0$, {$\boldsymbol{\lambda_0}$}]
call Where
[2, $\lambda_0$, {$\boldsymbol{\lambda_0}$}]
bind products
[3, $\lambda_0$, {$\boldsymbol{\lambda_0}$}]
$\lambda_1$ p
[4, $\lambda_0$, {$\lambda_0$, $\cancel{\lambda_1}$}]
call Any
[5, $\lambda_1$, {$\lambda_0$, $\boldsymbol{\lambda_1}$}]
bind products
[6, $\lambda_1$, {$\boldsymbol{\lambda_0}$}]
$\lambda_2$ p2
[7, $\lambda_1$, {$\lambda_0$, $\lambda_1$, $\cancel{\lambda_2}$}]
binop ==
[8, $\lambda_2$, {$\lambda_0$, $\lambda_1$, $\boldsymbol{\lambda_2}$}]
field unitPrice
[9, $\lambda_2$, {$\boldsymbol{\lambda_2}$}]
bind p2
[10, $\lambda_2$, {$\boldsymbol{\lambda_2}$}]
call Round
[11, $\lambda_2$, {$\lambda_0$, $\boldsymbol{\lambda_1}$}]
bind Math
[12, $\lambda_2$, {$\boldsymbol{\lambda_0}$}]
binop \
[13, $\lambda_2$, {$\lambda_0$, $\boldsymbol{\lambda_1}$}]
field unitPrice
[14, $\lambda_2$, {$\boldsymbol{\lambda_1}$}]
bind p
[15, $\lambda_2$, {$\boldsymbol{\lambda_1}$}]
const 1.2
[16, $\lambda_2$, {$\boldsymbol{\lambda_0}$}]
const 2
[17, $\lambda_2$, {$\boldsymbol{\lambda_0}$}]
$\lambda_3$ p
[18, $\lambda_0$, {$\cancel{\lambda_3}$}]
field unitPrice
[19, $\lambda_3$, {$\boldsymbol{\lambda_3}$}]
bind p
[20, $\lambda_3$, {$\boldsymbol{\lambda_3}$}]

Fig 5. Example abstract syntax tree algorithm nodes annotations.

Partial results of the algorithm work for the unoptimised query are presented in Fig. 5. Each abstract syntax tree node of the query is annotated with three values: (1) a number indicating an order of visitation, (2) a lambda expression directly including an expression, (3) bindings set including the bolded element denoting a native lambda of an expression. Lambda expressions have been assigned unique numbers to facilitate their identification. Bindings that are removed at the end of lambda node visitation are indicated by a strikethrough symbol.

Free expressions have their native lambda (bolded lambda in bindings set) different from a nearest lambda (denoted by the second annotation), i.e. expressions with visitation order ranks 6, 11-17. After omitting terminal expressions such as literals (constant type nodes ranks 16 and 17) and name bindings (bind type nodes ranks 6 and 15), the only MFE left to factor out is *Math.Round*(*p.unitPrice* / 1.2, 2)). Its native lambda is $\lambda_1$. Hence, factoring out should be applied to its indirect parent: the *Any* node with a visitation order rank 5 (presented in Listing 11). It is an expression inducing iteration over the products collection at the highest level within $\lambda_1$.

### *B. Applying Factoring Out*

The factoring out rewriting rule can be applied during visitation of lambda expressions. However, not all MFEs should be factored out. The conditions under which the optimisation promises well are described in analogous solutions [15][8], namely: (1) a free expression cannot be too simple (e.g., names and literals), (2) a free expressions' result should be used more than once. They can be verified during preparation to the transformation.

First, the complexity of an MFE can be examined. An appropriate threshold for applying transformation could be introduced, e.g., based on an arbitrarily set weight of language constructs comprising an MFE. Performance tests on the ***promoProducts*** problem involving factoring out a relatively simple expression have proven improvement in the case of collections consisting of at least 30 objects. For over 250 products optimised query was about twice as fast.

The second condition concerns a number of times that a MFE result is used in evaluation. An additional analysis may be necessary for confirming that $n\lambda$(MFE) contains a method that causes iteration over some collection that may require repeated evaluation of the MFE. For example, in LINQ this concerns mainly operators parameterised with a lambda abstraction (such as *Select*, *Where*, *Max*, etc). Operators operating on sets (e.g., *Contains*, *Union*) or custom ones are not any indication for the optimisation. The more detailed cardinality analysis is doubtful in case of a programming language environment and a lack of a cost model.

We have implemented a prototype LINQ provider library realising the mentioned optimisation (available at *https://github.com/radamus/OptimizableLINQ*). The analysis and the transformation are performed using the LINQ expression trees' representation available at runtime. Access to expression trees is provided though the *IQueryable*<T> interface that does not allow direct query execution. Instead, it exposes an abstract syntax tree of a query (in a form of a type-checked expression tree) to a data store provider. The provider makes use of this representation to build a query in a form (language) dedicated for a given data model (e.g., LINQ to SQL) [36].

Implementing optimisation in the form of a LINQ provider library gives a developer possibility to resign from aggressive, global query optimisation, e.g. when the order of evaluation is important considering some planned side effects. To enable automatic optimisation, the *AsOptimizable* extension method should be applied to a source collection. It is shown in Listing 14 for the Ikura product example.

```
var ikuraQuery = products.AsOptimizable().
  Where(p => products.Where(p2 =>
                  p2.productName == "Ikura").
    Select(p2=>p2.unitPrice).Contains(p.unitPrice)).
  Select(p => p.productName);
```

Listing 14. Example 1 – automatically optimised.

As a result, a rewritten query is compiled and becomes available for multiple use. One-time overhead occurring at the site of the definition is about a millisecond. A developer should consider runtime optimisation with caution when a query is used only once over a small collection. In contrast to LINQ, Java 8 streams operators are consumable, which prevents multiple usages of the same query. We are not aware of any mechanism enabling rewriting optimisations of Java 8 stream queries at runtime; nevertheless, in the case of consumable constructs the cost of optimisation done at runtime would burden each query execution.

## VIII. Summary

The proposed solution proves that it is possible to provide programming languages offering functional-style access to querying data collections with resource-independent static optimisation mechanisms. We proposed a formal method – factoring out of free expressions – based on higher-order functions rewriting. Its essence is to avoid unnecessary recurring calculations. Factoring out of a free expression that is complex to calculate generally produces a robust performance gain. Such optimisation can be fully automated and does not require any interference or implementation-specific knowledge from a programmer. Using simple examples, we emphasise the significance of the order of evaluation implied by semantics of functional-style operators. Finally, we elaborate general and safe optimisation, considering characteristics of functional-style querying in imperative programming languages.

In contrast to the Nectar system [12], which also uses term rewriting to increase sharing of computations, our work addresses functional-style queries in general, i.e. without context of application which would limit our optimisation. We take advantage of the similar approach to optimisation as Steno [21], LinqOptimizer [22], or SkyLinq [23]. However, we make an attempt to explore more aggressive, global optimisations comparable to optimisations of database query languages.

The presented approach was verified in Microsoft .NET environment and its Language-Integrated Query technology. However, the automated solution has not been straightforward to elaborate due to necessity of considering several variants implied by execution strategies of constructs comprising LINQ queries and complexity of implementing LINQ providers.

Our optimisation for LINQ can be combined automatically with other ones as long as they preserve queries in an expression trees form. In other cases, fusion of optimisations has to be done manually. For example, PLINQ enables to take advantage of multiple cores and achieve several times better efficiency in processing of large collections. Moreover, the optimiser in some cases could automatically (or by a programmer's decision) resign from suspending evaluation of a factored out expression and remove overhead that it imposes. The tests showed that it results in further improvement of performance, up to ~18%. Finally, it seems that transformations would be the most profitable if incorporated in a compiler. Considering source-to-source transformations already performed by the C# compiler on LINQ query expressions [33] this solution imposes itself.

We believe that our work is as a real step towards genuine declarative language-integrated queries. We conduct further works on optimisation of functional-style constructs processing collections. One branch of our research concerns the elaboration of methods that are aware of operators semantics, e.g., addressing complex queries taking advantage of the selection operation, which exposes a huge potential for optimisation (e.g., pushing selection [37]). We also consider adapting other methods, such as revealing weak dependencies within queries that enable performing further factoring out [38].

## References

[1] E. Meijer, "The world according to LINQ," Commun. ACM, vol. 54, no. 10, pp. 45–51, Oct. 2011. http://dx.doi.org/10.1145/2001269.2001285

[2] Oracle, Java API docs, "Package java.util.stream", http://docs.oracle.com/javase/8/docs/api/java/util/stream/package-summary.html, accessed: December 2014.

[3] M. Fowler, "Collection Pipeline", 28 July 2014, http://martinfowler.com/articles/collection-pipeline/, accessed: December 2014.

[4] Hudak, "Conception, Evolution, and Application of Functional Programming Languages". ACM Computing Surveys 21 (3), pp. 383–385, 1989. http://dx.doi.org/10.1145/72551.72554

[5] Chaudhuri, "An Overview of Query Optimization in Relational Systems", Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of data-base systems, Seattle, Washington, United States, pp. 34-43, 1998. http://dx.doi.org/10.1145/275487.275492

[6] M. Jarke, and J. Koch, "Query Optimization in Database Systems", ACM Computing Surveys 16(2), pp. 111-152, 1984. http://dx.doi.org/10.1145/356924.356928

[7] W. Kim, "On optimizing an SQL-like nested query", ACM Trans. on Database Systems, 7(3), pp. 443–469, 1982. http://dx.doi.org/10.1145/319732.319745

[8] J. Plodzien, and A. Kraken, "Object Query Optimization through Detecting Independent Subqueries". Information Systems 25(8), pp. 467-490, 2000.

[9] S. Cluet, and G. Moerkotte, "Nested Queries in Object Bases", In DBPL'93, pp. 226-242, 1993.

[10] N. May, S. Helmer, and G. Moerkotte, "Strategies for Query Unnesting in XML Databases", ACM Transactions on Database Systems (TODS), Volume 31 Issue 3, September 2006, pp. 968-1013, 2006. http://dx.doi.org/10.1145/1166074.1166081

[11] Y. Yu, M. Isard, D. Fetterly, M. Budiu, U. Erlingsson, P. K. Gunda, and J. Currey, "DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language", Symposium on Operating System Design and Implementation (OSDI), San Diego, CA, December 8-10, 2008.

[12] P. K. Gunda, L. Ravindranath, C. A. Thekkath, Y. Yu, and L. Zhuang. "Nectar: Automatic Management of Data and Computation in Datacenters", In OSDI, 2010.

[13] L. Fegaras, and D.Maier, "Optimizing object queries using an effective calculus", ACM Transactions on Database Systems (TODS), Volume 25 Issue 4, pp. 457-516, 2000. http://dx.doi.org/10.1145/1166074.1166081

[14] S. Jones, A. Tolmach, and T. Hoare, "Playing by the Rules: Rewriting as a practical optimisation technique in GHC", Proceedings of the 2001 Haskell Workshop, pp. 203-233, 2001.

[15] S. P. Jones, W. Partain, and A. Santos, "Let-Floating: Moving Bindings to Give Faster Programs", Proceedings of the 1996 ACM SIGPLAN International Conference on Functional Programming, 1996. http://dx.doi.org/10.1145/232629.232630

[16] J. Hughes, "The Design and Implementation of Programming Languages", Oxford University, D.Phil. Thesis, 1983.

[17] A. Biboudis, N. Palladinos, and Y. Smaragdakis, "Clash of the Lambdas", 9th ICOOOLPS (Implementation, Compilation, Optimization of OO Languages, Programs and Systems) workshop, Uppsala, Sweden, 2014.

[18] T. Grust, J. Rittinger, and T.Schreiber, "Avalanche-safe LINQ compilation", Proceedings of the VLDB Endowment, Volume 3 Issue 1-2, pp. 162–172, 2010. http://dx.doi.org/10.14778/1920841.1920866

[19] J. Chaney, S. Lindley, and P Wadler, "A practical theory of language-integrated query", ICFP '13 18th ACM SIGPLAN international conference on Functional programming, ACM SIGPLAN Notices - ICFP'13, Volume 48 Issue 9, pp. 403-416, 2013. http://dx.doi.org/10.1145/2500365.2500586

[20] i4o, "i4o - Indexed LINQ", http://i4o.codeplex.com, accessed: September 2014.

[21] D. G. Murray, M. Isard, and Y. Yu, "Steno: Automatic Optimization of Declarative Queries", PLDI'11 June 4–8, San Jose, California, USA, 2011. http://dx.doi.org/10.1145/1993498.1993513

[22] N. Palladinos, and K. Rontogiannis., "LinqOptimizer: an automatic query optimizer for LINQ to objects and PLINQ", http://nessos.github.io/LinqOptimizer/, accessed: December 2014.

[23] Sky LINQ, "Sky LINQ", https://skylinq.codeplex.com, accessed: December 2014.

[24] The Reactive Manifesto, "The Reactive Manifesto", http://www.reactivemanifesto.org, 23 September 2013, accessed: December 2014.

[25] E. Meijer, "Your mouse is a database", Commun. ACM, 55(5), pp. 66-73, 2012. http://dx.doi.org/10.1145/2160718.2160735

[26] I. Maier, and M. Odersky, "Higher-Order Reactive Programming with Incremental Lists", ECOOP 2013 – Object-Oriented Programming, Lecture Notes in Computer Science Volume 7920, pp. 707-731, 2013. http://dx.doi.org/10.1007/978-3-642-39038-8_29

[27] G. Schueller, and A. Begrend, "Stream fusion using reactive programming, LINQ and magic updates", 16th International Conference on Information Fusion (FUSION), pp. 1265-1272, 2013.

[28] Y. Wang, and X. Zhang, "The Research of Multi-source heterogeneous Data Integration Based on LINQ", Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on Computer Science and Electronics Engineering, pp.147-150, 2012. http://dx.doi.org/10.1109/ICCSEE.2012.437

[29] C. Reichenbach, Y. Smaragdakis, and N. Immerman. PQL, "A purely-declarative java extension for parallel programming". ECOOP '12, LNCS 7313, pp. 53–78, 2012. http://dx.doi.org/10.1007/978-3-642-31057-7_4

[30] E. Wcislo, P. Habela, and K. Subieta, "Implementing a Query Language for Java Object Database", ADBIS 2012, pp. 413-426, 2012. http://dx.doi.org/10.1007/978-3-642-33074-2_31

[31] R. Adamus., P. Habela, K. Kaczmarski., M. Lentner, K. Stencel, and K. Subieta, "Stack-Based Architecture and Stack-Based Query Language", ICOODB Proceedings of the First International Conference on Object Databases. Germany, Berlin, pp. 77-95, 2008.

[32] K. Subieta, "Stack-Based Approach (SBA) and Stack-Based Query Language (SBQL)", http://www.sbql.pl, 2011, accessed: December 2014.

[33] G. M. Bierman, E. Meijer, and M. Torgersen, "Lost In Translation: Formalizing Proposed Extensions to C#", Proceedings of the 22nd annual ACM SIGPLAN conference on Object-oriented programming systems and applications, pp. 479-498, 2007. http://dx.doi.org/10.1145/1297105.1297063

[34] A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman, "Compilers - principles, techniques and tools", Pearson Education, Inc., 2006.

[35] O. Eini, "The Pain of Implementing LINQ Providers", ACM Queue - Mobile Devices in the Enterprise, Volume 9 Issue 7, 2011. http://dx.doi.org/10.1145/1978542.1978556

[36] T. Petricek, "Building LINQ queries at runtime in C#", http://tomasp.net/blog/dynamic-linq-queries.aspx/, 2007, accessed: December 2014.

[37] M. Drozd, M. Bleja, K. Stencel, and K. Subieta, "Optimization of Object-Oriented Queries through Pushing Selections", ADBIS 2012, pp. 57-68, 2012. http://dx.doi.org/10.1007/978-3-642-32741-4_6

[38] M. Bleja, K.Stencel, and K. Subieta, "Optimization of object-oriented queries addressing large and small collections" IMCSIT 2009, pp. 643-650, 2009. http://dx.doi.org/10.1109/IMCSIT.2009.5352770

DOI: 10.15439/2015F99
ACSIS, Vol. 5

# Ruby Benchmark Tool using Docker

Richard Ludvigh, Tomáš Rebok
Faculty of Informatics, Masaryk University
Botanická 68a, 60200, Brno, Czech Republic
Email: 409737@mail.muni.cz, xrebok@fi.muni.cz

Václav Tunka, Filip Nguyen
Red Hat Czech, JBoss Middleware
Purkyňova 111, 61245 Brno, Czech Republic
Email: {vtunka,fnguyen}@redhat.com

***Abstract*—The purpose of this paper is to introduce and describe a new Ruby benchmarking tool. We will describe the background of Ruby benchmarking and the advantages of the new tool. The paper documents the benchmarking process as well as methods used to obtain results and run tests. To illustrate the provided tool, results that were obtained by running a developed benchmarking tool on existing and available official ruby benchmarks are provided. These results document advantages in using various Ruby compilers or Ruby implementations.**

## I. Introduction

RUBY IS A PURE OBJECT-ORIENTED interpreted language. The language itself has three major implementations: MRI written in C, JRuby written in Java, and Rubinius written in Ruby. These are often compared in different ways such as usage, performance, and memory requirements. The non-functional attributes of the implementations vary significantly.

Our goal was to develop a benchmarking tool for these implementations that would wrap all of the provided versions and run benchmarks against them. It is important to mention that this paper does not focus on developing benchmarks, but merely on providing a tool capable of taking any kind of existing benchmark, running it against all available Ruby versions, and testing its usability.

To ensure complete isolation of all tested Ruby versions, we used Docker [2] to pack each configuration inside a Docker container. Docker is a open-source tool that enables a Linux application and its dependencies to be packaged as a lightweight container. The benchmarking tool was then developed to handle these containers as well as to validate their content (correct versions, compilation flags, compilers used). It is also responsible for running selected benchmarks and storing their results. In section III, we describe the basic practices used while developing the benchmark tool as well as all of its responsibilities, methods used to collect data, and available Ruby versions.

In our research, we used official available benchmarks from the Ruby repository[1] and a parallelism benchmark from the Rubinius repository[2] to point out some basic characteristics of Ruby implementations and their power. The tool was run on a baremetal and virtual server to provide results from both environments. In section III-D, we describe both environments and their configuration in detail.

The results we present are also available online. The results are in three main areas:

- MRI Versions overview - we have compared multiple MRI Ruby versions to determine the progress mainly in memory usage as new MRI (2.2.0) has announced a new garbage collection algorithm.
- A comparison of MRI compilers determined the differences in using different C compilers to compile Ruby (2.2.0 used in benchmarks). Our results proved that the widely used, also default for a dominant group of linux distributions, version 4.8 of GCC is the best choice for prime performance.
- Running benchmarks on different Ruby implementations not only proved that MRI is best for short single-threaded executions (these tests were extremely short prohibiting to start the just-in-time compilers for JRuby and Rubinius) while JRuby and Rubinius are better for longer or multi-thread runs, but it also showed the progression and power of JRuby in handling parallel tasks and improving in performance from version to version.

## II. State of The Art

Ruby [1] is an interpreted object-oriented programming language which was designed and released by Yukihiro Matsumoto, known as Matz, in 1995. Ruby is a pure object-oriented language in which even values of primitive types (true, false, nil) are represented as objects. It is also suitable for functional programming and capable of powerful metaprogramming.

There are three major Ruby implementations.

The oldest - original implementation - is known as MRI ("Matz's Ruby Interpreter") or CRuby (since it is written in C). CRuby does support native threads, but it uses Global Interpreter Lock [3] (known as GIL) which allows data to be modified only one thread at the time, thus prohibiting true concurrency.

JRuby, as the title suggests, is a 100% Java implementation of Ruby. This allows for Ruby applications to be run on a Java Virtual Machine (JVM), thus utilizing its just-in-time (JIT) compiler, garbage collection, and mainly its concurrent threads. It also allows us to use any library that is compatible with JVM.

Access to the CERIT-SC computing and storage facilities provided under the programme Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ. 1.05/3.2.00/08.0144, is greatly appreciated.

[1] https://github.com/ruby/ruby

[2] https://github.com/rubinius/rubinius-benchmark

978-83-60810-66-8/$25.00©2015, IEEE

Rubinius is the Ruby version of Python's PyPy. Much of its source is written in Ruby making it easier to understand. It also includes a just-in-time compiler on a virtual machine written in C++.

In December 2013, Sam Saffron published a call for an official long-running Ruby benchmark [4]. At the beginning of development, in November 2014, Ruby still had no long term running benchmarks like Pypy speed center[3] for Python. At that time, there was just one Ruby benchmarking suite that was used by the community to test different configurations and experiment with their performance. The Ruby Benchmark Suite [5] by Antonio Cangiano was developed between the years 2008 and 2013. It uses a host OS and installed rubies to perform tests. During the development of our benchmark tool, Rubyfy.ME, Guo Xiang Tan[4] presented his own benchmark tool in cooperation with Sam Saffron. This became the official ruby long-term running benchmark[5]. His tool uses a single docker image containing all tested Ruby versions. It also provides tests for individual commits in the official Ruby repository acting more like performance CI (Continuous Integration) solution.

## III. Benchmark Tool

In this section we describe the benchmark tool that we have developed. The main aim during development was to ensure the separation of all Ruby versions so they did not share a library or resource that could affect the results. To ensure complete isolation, docker images were created for each Ruby version or compiler. Since this tool is written in Ruby, there are only two system requirements that are required to run the developed benchmarking tool:

- Ruby of version 2.0 or above
- Docker of version 1.0.1 or above

### A. Docker integration

Docker [2] is an open-source program which is capable of packing linux applications and their dependences as a container. Container-based virtualization isolates applications from each other. It runs in userspace on the host operating system utilizing resource isolation and allocation benefits while being much more efficient.

The tool automatically downloads all required docker images and validates a correct Ruby presence. This means that the correct Ruby version, the correct compiler and its version, and the correct compilation flags are present in each docker image. It is also responsible for running benchmarks in correct containers.

Table I lists Ruby versions and implementations that are present in the benchmarking tool.

### B. Running Benchmarks

All benchmarks must be located inside exactly one subfolder in the benchmarks folder. This hierarchy allows to track the origin or divide the benchmarks into groups. There is one special category of benchmarks located inside the benchmarks/custom folder (we will call them custom benchmarks from now) which is handled differently compared to the others. By default, official Ruby benchmarks are present in the benchmarks/ruby-official.

Every benchmark that is not from a custom category is, before its execution, wrapped inside an additional code that captures all of the needed information. Memory usage is stored before code execution. Then, using the official benchmark library for Ruby, time consumption is tracked. At the end, garbage collection is triggered manually and memory usage is tracked once more. This way, the needed time, used memory, and total executable memory are tracked for each benchmark.

Custom benchmarks do not undergo the described flow. Code is executed without modification and standard and error output is captured, thus allowing us to store any kind of information.

Rubinius and JRuby implementations of Ruby contain a just-in-time compiler which results in slower code execution during program startup. For these versions, it is important to warm up the virtual machine (JVM for JRuby) by running the code multiple times or for a short period of time before actually benchmarking. There is still no support because benchmarks need to be adapted to run in loops in order to ensure warm-up. For now, we do not use suitable benchmarks so all benchmarks are run without a warm-up, resulting in slower times for JRuby and Rubinius. However, we plan to add this feature after we ensure enough benchmarks that are able to run in loops.

### C. Storing and Publishing Results

Each run stores a record in a csv file named after the Ruby version used for that run. Each record contains information about the executable benchmark, Ruby version, compiler, and current time. Standard output and standard error output are stored for custom benchmarks while time and memory usage are stored for others.

After providing information about the web interface of this benchmark tool, the user is able to push results from csv files to Ruby on Rails application which provides complex results and graphs.

The feature to publish stored results allows us to build an easy and user friendly presentation. Using the highcharts[6] library, simple graphs were created to simplify collected results. These graphs were split into three main categories: an MRI versions comparison, an MRI C compilers comparison, and a Ruby implementations comparison. Overall, graphs were created for each category as well as for each benchmark.

This tool was developed for community use so its main characteristic is easy usability and extensibility allowing users

[3] http://speed.pypy.org/

[4] https://github.com/tgxworld

[5] http://rubybench.org/

[6] http://www.highcharts.com/

TABLE I
RUBY VERSIONS AND IMPLEMENTATIONS

| Implementation | Versions | Details |
|---|---|---|
| JRuby | 9.0.0.0.pre1 | OpenJDK 64-Bit Server, Virtual machine version 1.7.0_75-b13 |
| JRuby | 1.7.12, 1.6.8 | OpenJDK 64-Bit Server, Virtual machine version 1.7.0_65 |
| Rubinius | 2.2.10, 2.3.0, 2.4.0, 2.4.1 | |
| MRI Ruby | 2.2.0 | Multiple images with different compilers: GCC 4.8, GCC 4.9, Clang 3.3, Clang 3.4, Clang 3.5, all on both -O2 and -O3 flags |
| MRI Ruby | 1.6.8 - 2.1.5 (12 versions) | All compiled using GCC 4.8 -O2 |

to easily add new benchmarks, test their own code, or even add more Ruby versions. This tool also comes with its own Rails web application which allows users to see their results in no time. However, this approach makes it harder to specify or focus on some deeper characteristics of Ruby. For example, the user is currently unable to set or change runtime flags, but we are planning to provide this feature in the near future.

### D. Environment

Benchmarking was performed in two independent environments.

- Baremetal Ubuntu 14.04, kernel version 3.13.0-36-generic x86_64 GNU/Linux, with Intel(R) Core(TM) i5-2450M CPU @ 2.50GHz, 2x 8GB Samsung SODIMM DDR3 Synchronous 1333 MHz RAM located on an Intel Emerald Lake motherboard.
- A virtual private server provided by CERIT-SC, configured to 16GB RAM and 8 virtual CPU running Debian 3.16.7-ckt4-3 bpo70+1. This virtual machine is hosted on 2x Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz (12 cores) 96 GiB DDR3 1333 MHz. Center CERIT-SC (CERIT Scientific Cloud) offers computing resources and also participates in research and development activities.

Each Ruby implementation was run with its default settings. We did not use any runtime command flags as there is no support for this feature yet. The only exception was made during a parallelism test when JRuby and Rubinius were run with JIT disabled.

This tool was developed as an open-source project and its sources are publicly available on Github[7].

## IV. RESULTS

Using benchmarks from the official Ruby repository enables us to present results in the following categories:

- MRI Ruby Compilers - a comparison of Ruby compilers (Clang, GCC) tested on MRI Ruby Versions 2.2.0.
- Ruby Implementations - the difference between various implementations and between handling single-thread vs. multi-thread tasks.
- Ruby 2.2.0 Garbage collection - the progress of new incremental Garbage collection announced in Ruby 2.2.0.

Each benchmark was successfully run ten times to provide stable and usable results from both environments. During the benchmarking and development process, the most recent versions of selected implementations were: 2.2.0 for MRI, 2.4.1 for Rubinius, and 9.0.0.0.pre1 for JRuby.

### A. MRI Ruby Compilers

As MRI Ruby (also called CRuby) is written in C, the choice of a C compiler and its compilation flags can affect the performance of a Ruby interpreter.

In December 2014, Peter Wilmott published an article [6] about MRI Ruby compilers. All of his tests were run on AWS from an m3.medium EC2 instance and he used Ruby version 2.1. Using benchmarking suite developed by Antonio Cangiano, his results show a great performance increase from GCC 4.8 to GCC 4.9 and he also pointed out that optimization on level 2 works better than on level 3.

Running all of the official Ruby benchmarks on the Ruby version 2.2.0 on both a baremetal and a virtual server provided by CERIT-SC group resulted in our results being different than those obtained by Peter. Both sets provided us with the same result (Fig. 1 and Fig. 2), thus showing that GCC 4.8 on optimization level 3 (the default shipped for Ubuntu 14.04) is still ahead by a small amount (1-2% faster than GCC 4.9 -O3). The results also provided that level 3 optimization is currently faster for Ruby 2.2.0 (an almost 4% speed increase from GCC 4.8 -O2 to GCC 4.8 -O3).

### B. Ruby Implementations

The main difference between MRI Ruby and Rubinius, JRuby, is that MRI Ruby uses GIL (global interpreter lock) which makes single threaded tasks run faster, but does not permit real concurrency. This is why MRI Ruby is significantly better in overall results (official ruby benchmarks that are used for each category are single thread simple tests and often too short to start a JIT compiler on JRuby and Rubinius, thus making their results even worse) as is shown on Fig. 3 and Fig. 4

Although we used the parallelism benchmark provided in the Rubinius repository[8] to determine the differences in handling parallel tasks. In this benchmark we manually edited the tool to pass command line arguments disabling JIT compilers for JRuby and Rubinius (-X-C for JRuby and -Xint for Rubinius ) because this benchmark is built to run multiple times and provide the best results gathered. This would allow the JIT compilers to activate and radically reduce the execution times.

[7] https://github.com/Ryccoo/rubyfy-me-docker-suite

[8] https://github.com/rubinius/rubinius-benchmark/blob/master/parallelism.rb

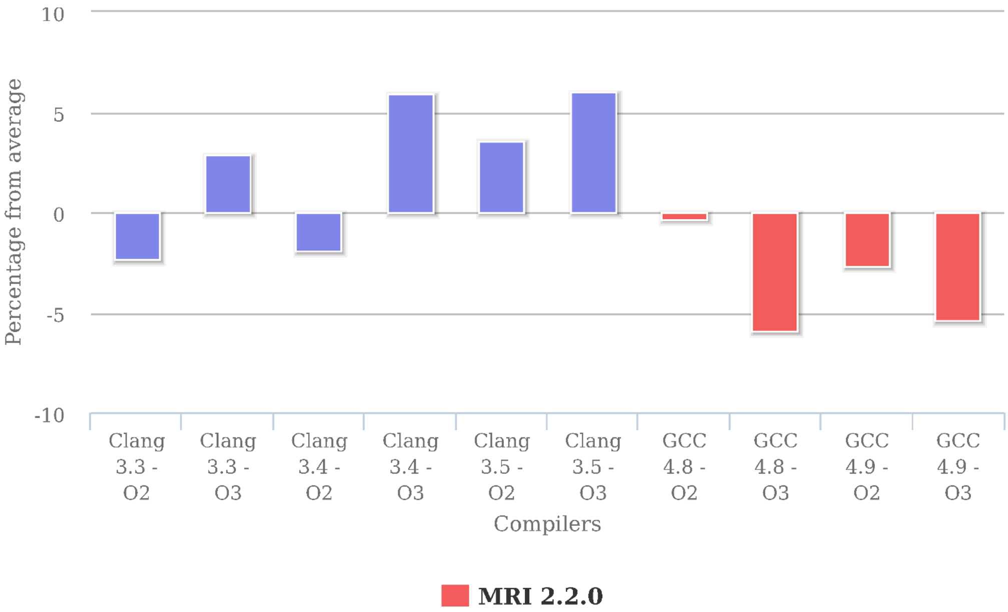


Fig. 1. Overall results for MRI compilers run on CERIT-SC virtual server

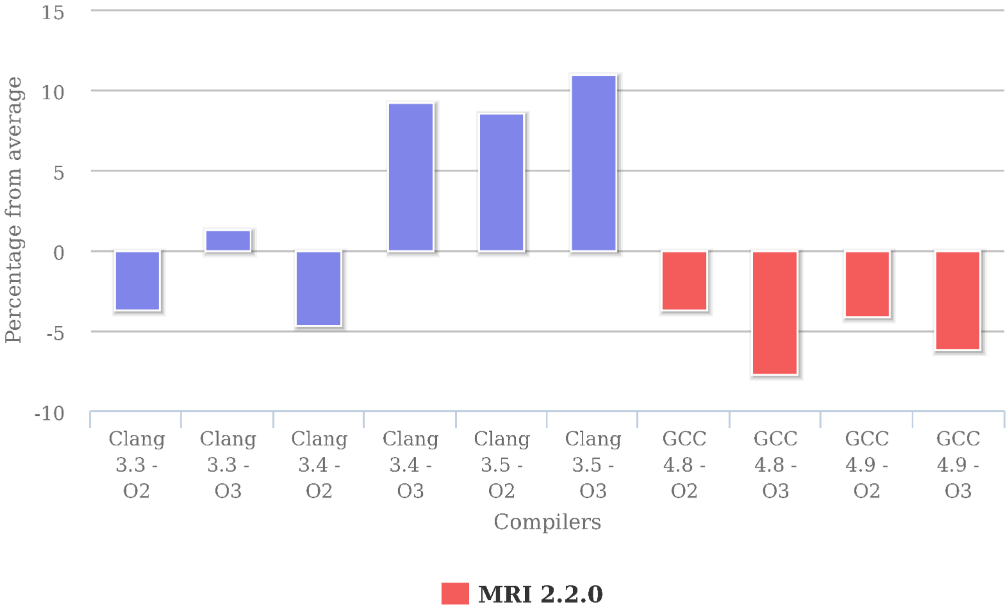


Fig. 2. Overall results for MRI compilers run on a baremetal Ubuntu machine

On Fig. 5 we can see the true power of JRuby parallelism as well as its progression.

The benchmark first computes the amount of work needed to keep the thread busy for two seconds, then runs the same amount of work on each of four threads. Because virtualization (the usage of virtual CPUs) makes it harder to compute and calibrate needed work amount, we present only results from the baremetal machine.

### *C. MRI 2.2.0 Incremental garbage collection*

The MRI Ruby version 2.2.0 has announced new incremental garbage collection. From this version, symbols are also garbage collectable. Symbols are now divided into two categories: mortal and immortal. Immortal symbols are defined inside code while mortal symbols are created dynamically during execution. MRI Ruby 2.2.0 now collects mortal symbols allowing to free up more memory. We were able to watch the decrease of memory usage compared to the previous version as shown on Fig. 6.

These tests were also run separately and aimed at the last

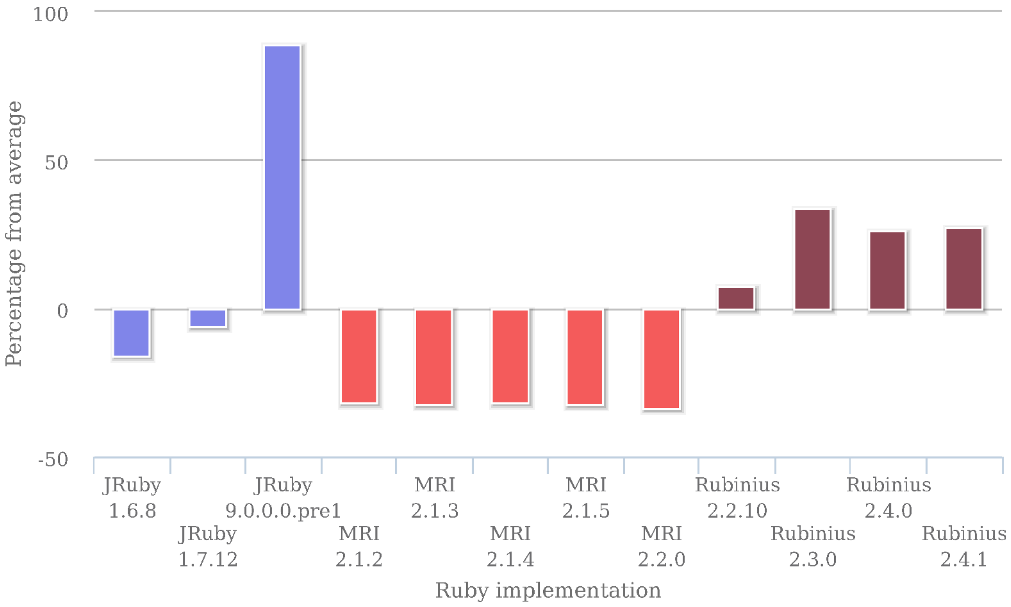


Fig. 3. Overall time performance on different Ruby implementations tested on a baremetal Ubuntu machine

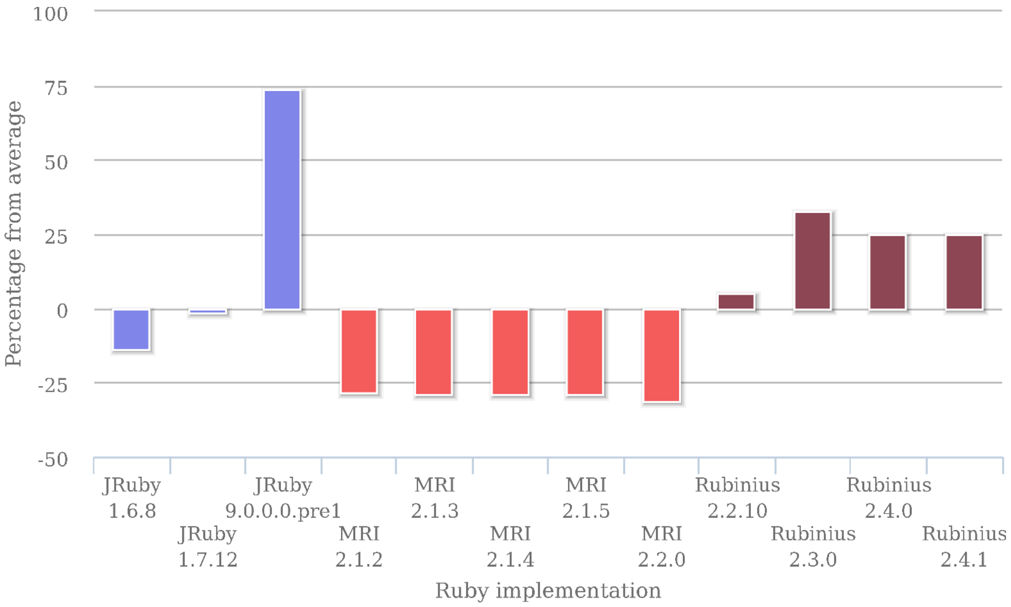


Fig. 4. Overall time performance on different Ruby implementations tested on a CERIT-SC virtual server

Ruby versions which confirmed the gap between ruby 2.1.x and 2.2.0 as shown on Fig. 7.

## V. Conclusion

Our results show that in the case of compiling the newest version of MRI Ruby (currently 2.2.0 during benchmarking) for normal, non-experimental or special use cases, the choice of GCC 4.8 with an O3 flag will provide the best available performance. While this version of a GCC compiler is the default for the predominant group of linux distributions, there is no need to make any changes during Ruby installation.

MRI Ruby is the best choice for computing single threaded simple and short tasks while the GIL (Global Interpreter Lock) is prohibiting to starting real concurrency. This is when the choice of other implementations like JRuby and Rubinius is important. As shown on Fig. 5, JRuby provides us with the best concurrency. This ability increases from version to version. Therefore, JRuby is the best choice for big cloud computations offering multiple virtual CPUs.

It is important to remind that used benchmarks and benchmarking techniques did not allow the warm-up required by JIT compilers, thus results with proper and long enough warming

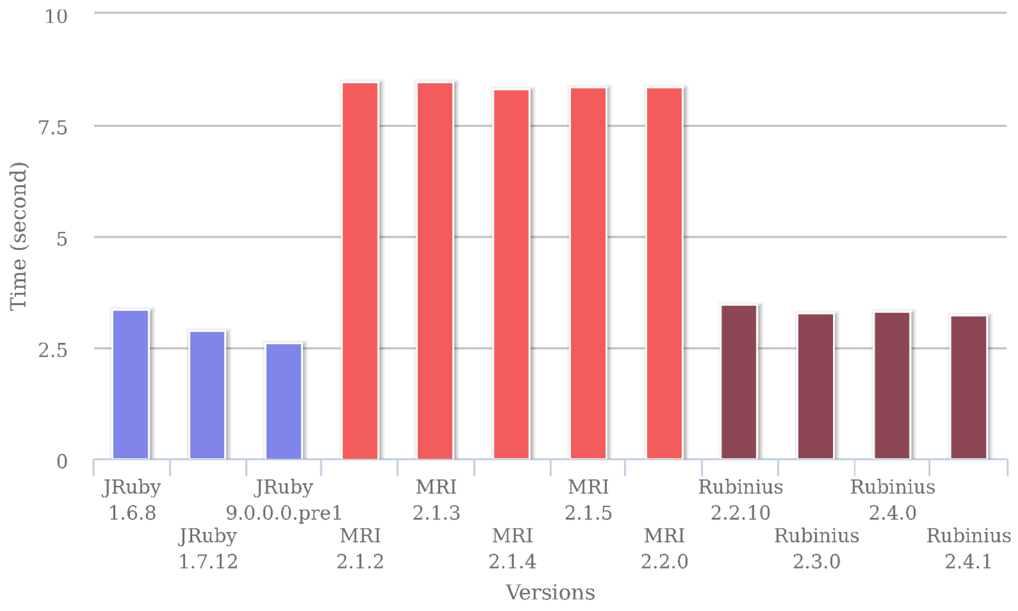


Fig. 5. Rubinius parallelism benchmark - running 4 parallel threads

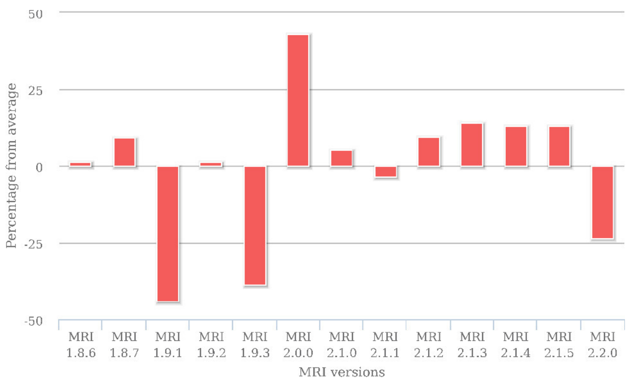


Fig. 6. Memory usage difference from the average

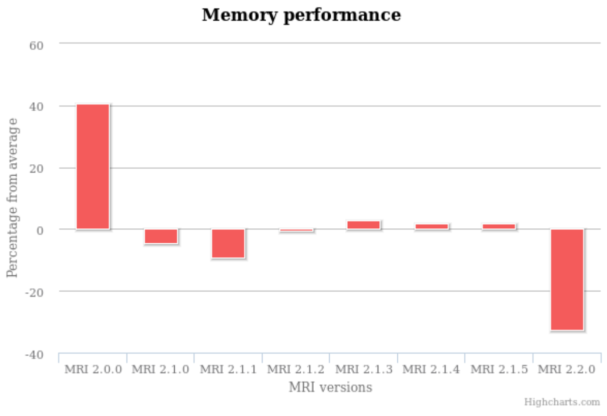


Fig. 7. Difference from average memory usage on recent Ruby versions

could differ on JRuby and Rubinius.

The introduction of new garbage collection in MRI 2.2.0 with the ability to collect mortal symbols can fix a common programmer mistake as a side effect. The problem occurred when the program was converting user inputs to symbols. These symbols were not garbage collectable and the program often drained the entire system memory and ended up freezing. Symbols converted from MRI 2.2.0 are tagged as mortal and garbage collected after being unused, thus not draining the entire memory.

## REFERENCES

[1] Hirschfeld, Robert, and Kim Rose, eds. Self-Sustaining Systems: First Workshop, S3 2008 Potsdam, Germany, May 15-16, 2008, Proceedings. Vol. 5146. Springer Science & Business Media, 2008.

[2] Dirk Merkel. 2014. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014, 239, pages.

[3] Rei Odaira, Jose G. Castanos, and Hisanobu Tomari. 2014. Eliminating global interpreter locks in ruby through hardware transactional memory. In Proceedings of the 19th ACM SIGPLAN symposium on Principles and practice of parallel programming (PPoPP '14). ACM, New York, NY, USA, 131-142.

[4] Call for official long running Ruby benchmark, http://samsaffron.com/archive/2013/12/11/call-to-action-long-running-ruby-benchmark

[5] Antonio Cangiano Ruby Benchmark, https://github.com/acangiano/ruby-benchmark-suite

[6] MRI Ruby Compilers Benchmark, https://www.p8952.info/ruby/2014/12/12/benchmarking-ruby-with-gcc-and-clang.html

# Source Code Annotations as Formal Languages

Milan Nosáľ, Matúš Sulír, and Ján Juhár
Department of Computers and Informatics
Technical University of Košice
Letná 9, 042 00 Košice, Slovakia
Email: milan.nosal@gmail.com, {matus.sulir,jan.juhar}@tuke.sk

***Abstract*—Attribute-oriented programming (source code annotations) is a program level marking technique that enables enrichment of program elements with custom metadata. In this paper we hypothesize that there is a correspondence between source code annotations and conventional formal languages in general. We analyze our observations about source code annotations from three aspects of language description: concrete syntax, abstract syntax, and semantics. The discussion provides evidence of the hypothesized correspondence and we use it as a basis for our definition of an *annotation-based language* (abbreviated: @L). However, the analysis also shows that compared to conventional formal languages, source code annotations have some specificities mainly connected to their binding to host program elements. The presented analysis contributes to the field of attribute-oriented programming by discussing the relationship between annotations and conventional formal languages, and by surveying relational idioms in annotations' usage that can be inspirational for annotations' authors.**

## I. Introduction

ATTRIBUTE-ORIENTED PROGRAMMING (abbreviated: @OP) as a technique of marking source code elements with source code annotations [1], [2] became quite popular during the last decade, as is manifested by multiple frameworks, such as the Spring Framework. The annotations as a metadata format found the same popularity in the academic environment as well. As an example we can mention an annotation-based parser generator YAJCo, which uses annotations with an object-oriented language model for syntax [3] and references definition [4]. We can also recognize a research field dedicated to annotations. Most important examples include a book about attribute-enabled software development by Cepa [5] that summarizes his research in the field, multiple research articles about enforcing annotations' dependencies in source code such as work of Noguera et al. [1], or one of the first articles on the topic of how to use (or how not to use) annotations by Correia et al. [6].

In our previous work [7] we analyzed the correspondence between annotations and XML in the scope of configuration languages. The main manifestation of the correspondence was the discovery of a set of mapping patterns between annotations and XML. Using the discovered mapping patterns we showed that annotations and XML can be considered equivalent in terms of their expressibility. Based on our previous work, in this paper we want to examine the relationship of annotations and formal languages deeper. A formal language is defined by an alphabet and a set of formation rules (grammar[1]). We will try to show that the same can be applied to annotations.

Our work presented in [7] can be considered a case study on correspondence between annotations and a representative of a generic language, XML. It indicates that there is probably a correspondence between annotations and formal languages in general, as well. Therefore in this paper we hypothesize that *there is a correspondence between annotations and conventional formal languages*. We will provide an analysis of our observations supporting this hypothesis and discuss the annotations from three language description aspects, concrete syntax (abbreviated: CS[2]), abstract syntax (abbreviated: AS), and semantics. The evidence presented in this work indicates that we can use the formal language theory when we are working with annotations. In consequence, the application of the approaches from the language theory can provide benefits for annotations' authors and users. In related work in section IX we will discuss several other works that connected annotations with languages, however, none of them considered this correspondence a hypothesis and tried to prove it. We will conclude the paper in section X with a brief discussion of consequences of proving this hypothesis.

The contributions of this work are as follows:

- observations of corresponding characteristics between source code annotations and formal languages (sections III, VI, and VII),
- discussion of discrepancies between annotations and formal languages (and thus identification of the main specificity of annotations in comparison with formal languages, section IV),
- survey (overview) of relational idioms between annotations and between annotations and their host language that can help annotations' authors during designing the annotations (sections III-B, IV-B, and IV-D),
- discovery of reversed code-wise relations between annotations and their target program elements that emphasize the significance of annotations relation to their host language (section IV-C), and

This work was supported by VEGA Grant No. 1/0341/13 "Principles and Methods of Automated Abstraction of Computer Languages and Software Development Based on the Semantic Enrichment Caused by Communication".
This paper shares contents with the first author's dissertation thesis.

[1] Grammar [8] described by concrete syntax (including also the lexical syntax) and abstract syntax. Abstract syntax describes the structural restrictions between language concepts (words). Concrete syntax describes the actual representation of the language sentences in the given alphabet.

[2] Not to be confused with Counter-Strike.

978-83-60810-66-8/$25.00©2015, IEEE

- definition of *annotation-based language* (abbreviated: @L) and its aspects from the viewpoint of formal language theory (throughout the whole paper with a summary in section VIII).

## II. Source Code Annotations

Before going into discussion about the correspondence we will clarify several essential terms connected with source code annotations. An *annotation-enabled language* (abbreviated: @EL) is any formal language that supports attribute-oriented programming. A language supports the attribute-oriented programming if its grammar (and therefore its parser too) allows adding custom declarative tags to annotate standard program elements. These tags have to be structured and therefore parsable by the parser (or by some additional tool, as in case of XDoclet[3] technology). An example of an @EL is Java programming language from version 1.5. In previous versions the attribute-oriented programming was supported by a 3rd party tool XDoclet. This means that if we add these tags to a valid host language sentence, it still *has* to be a completely valid host language sentence without any preprocessing or other manipulations. In other words, the @EL parser has to have an extension point, a grammar rule, that enables these tags. In case of Java annotations, they are a part of the Java language. In case of XDoclet, the XDoclet annotations are part of standard Java comments.

Annotations do not directly change the source code semantics. They only add metadata to source code. Annotations can be queried and processed on demand by frameworks or tools, or the program itself, thus indirectly changing program semantics.

Custom declarative tags supported by @EL are *source code annotations* (abbreviated: annotations). An annotation annotates an annotation-enabled program element[4]. E.g., a program element can be a method, a function, a class, a statement, etc., depending on language's programming paradigm. We will call the program element annotated by an annotation the *target host program element* (abbreviated: target element) of that particular annotation.

Source code annotations can be also dynamic as discussed by Noguera et al. [9] and Cazzola et al. [10]. Dynamic annotations can have properties that are evaluated dynamically when the annotations are processed (conventional static annotations are evaluated during compilation, just as constants). A dynamic annotation can have a property that is bound to some code property, e.g., a property that is bound to count of loop iterations cannot be evaluated during compile time. During runtime, when the loop is evaluated, the annotation will remember the iteration count and it can be queried. However, dynamic annotations cannot directly change program semantics as well.

The relation between an annotation and its target element has to be expressed by in-place binding. In-place binding requires that an annotation is placed next to or directly into its target program element declaration. The placement has strict rules that are dependent on the host @EL grammar. In General-Purpose Languages (abbreviated: GPL) it is usually expressed by prepending an annotation directly before the program element declaration. Our definition of annotations is presented in Definition 1.

**Definition 1.** *Annotations are custom declarative structured tags in host @EL that are bound to host program elements using in-place binding. Annotations have to be parsable by standard @EL parser (or a 3rd party tool) that allows implementing semantics in form of a plug-in.*

What follows is an analysis of the correspondence between annotations and formal languages.

## III. Abstract Syntax Correspondence

The main idea of the correspondence between a formal language and a set of related annotations stems from the fact that we can observe a structure (abstract syntax) in using annotations from the given set. The observed regularities are not accidental, and in all cases they are even enforced by the processing tools.

### A. Structural Correspondence Example

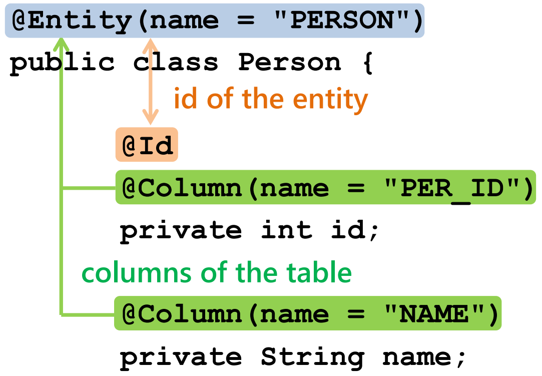
Let us begin with an illustrative example based on Java Persistence API (abbreviated: JPA) annotations. JPA is an object-relational mapping specification for Java. In JPA annotations are used to specify mapping between Java classes and a relational database. In a simple example presented in Figure 1 an `@Entity` annotation is used to specify that the `Person` class is going to be a persisted entity, and `@Column` annotations specify the mapping of fields to table columns. JPA specification requires the user to use the `@Entity` annotation to include the class in the persistence management setup[5]. Without the `@Entity` annotation all the `@Column` and the `@Id` annotations would not be processed by any JPA compliant object-relational mapping (ORM) tool. This relationship is represented by green arrows in Figure 1. Another requirement is that each entity marked with the `@Entity` annotation has to be annotated by the `@Id` annotation to specify which of the fields represent a primary key. This relationship is highlighted in orange in Figure 1.

Considering the example from Figure 1 one can easily notice that the `@Entity` annotation and the `@Column` annotations mimic an abstract syntax tree (abbreviated: AST). Annotations and their properties are nodes, thus modelling a tree. A sketch of such an AST is shown in Figure 2. We also added a simplified AST of the host language to illustrate the binding of annotations to their target elements.

[3] http://xdoclet.sourceforge.net/xdoclet/index.html

[4] Annotation-enabled program element is a program element that can be annotated. In some cases not all program elements can be annotated. E.g., in the Java language the `if` program element cannot be annotated by Java annotations (statements in general).

[5] One can alternatively use `@Embeddable`, or `@MappedSuperclass`, but those have slightly different semantics and are not important for the discussion.



```
@Entity(name = "PERSON")
public class Person {
    @Id
    @Column(name = "PER_ID")
    private int id;
    @Column(name = "NAME")
    private String name;
```

Fig. 1. Mapping of the Person entity class to database using JPA annotations

Now when we look at the given AST, we can easily devise a simple external domain-specific language[6] (abbreviated: DSL) that would to a degree copy the structure of these annotations. Snippet in Listing 1 can be an illustration of an external DSL that expresses the same information as annotations. Here we can see by an example that there is a correspondence between annotations and formal languages.

Listing 1. Person entity definition

```
Entity "PERSON" {
  Id Column   "PER_ID"
  Column      "NAME"
}
```

### *B. Idioms in Structural Relations*

The example shown in section III-A is just an illustrative example. In practice, however, we can find multiple structural stereotypes (idioms) in annotations' usage that support our hypothesis. Each *idiom specifies common structural relationship between annotations* (annotation-based language concepts) and as such defines a part of the annotation-based language's grammar. In this section we will provide a survey of commonly known and recognized structural idioms of annotations' usage. Although for the sake of the discussion about the correspondence a single example would be enough, this overview can be useful for annotations' authors in the process of designing an annotation-based language.

**Vectorial annotation** idiom described by Guerra et al. [12] enables us to add multiple annotations of the same type to the same target program element. An example showing vector of multiple `@Alternative` annotations is presented in Listing 2. Annotation `@Alternatives` has a single parameter of array of `@Alternative` annotations.

Listing 2. Vectorial annotation idiom example

```
@Alternatives(
   {@Alternative(B.class), @Alternative(C.class)})
public class A { ... }
```

**Composite annotation** idiom described by Guerra et al. [12] has annotations as its parameters as well. It allows creation of a tree-like structure of annotations, however, bound to the same target program element. A composite annotation example is shown in Listing 3. In the example there is the `@Author` annotation nesting `@Name` and `@Contact` annotations.

Listing 3. Composite annotation idiom example

```
@Author(
  name=@Name(firstName="Milan", surname="Nosal"),
  contact=@Contact(email="milan.nosal@gmail.com")
)
public class Person { ... }
```

**Inside relation** idiom requires an annotation @A to be inside scope of another annotation @B. E.g., if @A annotates a field of a class then a class (or its outer class or package) has to be annotated by @B. This idiom was described by Noguera et al. in [1] and also by Ruska et al. in [13], where they call it the Parent-child relation. This idiom is a special case of the Required attribute (annotation) idiom described by Cepa et al. [14] specifying that an annotation requires a usage of another annotation to be valid. In the Required attribute (annotation) idiom the scope is arbitrary. We have already seen an example of the Inside relation idiom in Figure 1 and it is reiterated in Listing 4. In it, the `@Id` and `@Column` annotations had to be used to annotate fields of a class annotated by the `@Entity` annotation. Annotated fields are inside the scope of the `@Entity` annotation annotating the enclosing class.

Listing 4. Inside relation idiom example

```
@Entity(name = "PERSON")
public class Person {

  @Id
  @Column(name = "PER_ID")
  private int id;

  @Column(name = "NAME")
  private String name;
  ...
```

**Neighbor** idiom specifies that a valid usage of annotation @A is only in case when its target program element is also annotated with another annotation @B. This does not necessarily mean that usage of @B requires @A (this relation is not commutative). This idiom is also a specialization of the Required attribute (annotation) idiom described by Cepa et al. [14]. Noguera et al. [1] describes this idiom as the Requires idiom. A variation of this idiom is described by Ruska et al. in [13] as the Occurrence of multiple annotations idiom that is commutative. In the following example in Listing 5 the `@Table` annotation requires the `@Entity` annotation annotating the same class in order to be considered for processing by the annotations semantics implementation. In this example only an entity class can be mapped to database.

Listing 5. Neighbor idiom example

```
@Entity
@Table("PERSON_TABLE")
public class Person { ... }
```

[6] A language focused on, and usually also restricted to a particular problem domain [11].

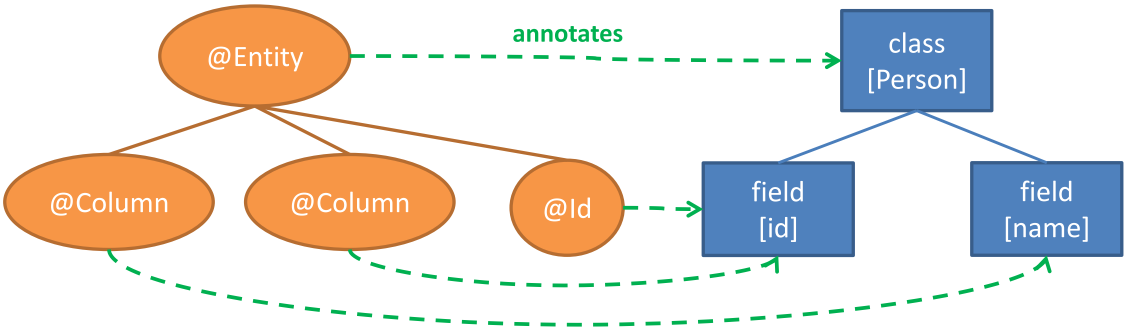


Fig. 2. AST model of the Person entity definition using JPA annotations

**Mutual exclusivity** idiom prohibits usage of an annotation @A if the target program element is annotated with @B (@A and @B are mutually exclusive). The idiom is described by Ruska et al. in [13]. Noguera et al. [1] call the same idiom the Prohibits idiom. This idiom is a specialization of the Disallowed idiom described by Cepa et al. [14] that can specify exclusivity on various scopes, not only on the same target program element. Following code sample from Listing 6 shows an example of invalid use of annotations that are mutually exclusive. In our annotation-based language a field cannot be annotated both by both `@Id` and `@Basic` because a column cannot be both identifier and regular column at the same time.

Listing 6. Mutual exclusivity idiom violation example

```
// incorrect usage of @Id and @Basic
@Id
@Basic
private String name;
```

**Unique annotation occurrence** idiom requires that only one instance of an annotation type can be present in a given scope of the host @EL. The idiom is described by Ruska et al. [13]. A variation of the idiom is mentioned by Noguera et al. [1] as the Unique idiom that can be applied to annotation parameters. This way an annotation-based language designer can specify that the parameter value of the annotation can occur only once in a given scope. In the example from Listing 7 there is a violation of the Unique annotation occurrence for `@Id` in the scope of the class. In this language there cannot be two identifiers for the same entity.

Listing 7. Unique annotation occurrence idiom violation example

```
@Table("PERSON")
public class Person {

  @Id
  private int id;

  // cannot be defined again
  @Id
  private String name;

  ...
```

**Refers to** idiom requires an annotation parameter to be a reference (usually implemented as a simple string) to a parameter value of another annotation. The idiom is described by Noguera et al. [1]. The code snippets in Listing 8 show a Refers to idiom where the `@Author` annotation from class `A` is reused for class `B` by using a reference. `@AuthorRef`'s `value` parameter[7] refers to the `id` parameter of the `@Author` annotation.

Listing 8. Refers to idiom example

```
@Author(
  id="milan.nosal",
  name=@Name(firstName="Milan", surname="Nosal"),
  contact=@Contact(email="milan.nosal@gmail.com")
)
public class A { ... }

@AuthorRef("milan.nosal")
public class B { ... }
```

## IV. Annotations' Abstract Syntax Specificities

We discussed an important observation about correspondence between a set of annotations and a formal language in terms of abstract syntax. However, there are also discrepancies between annotations and a common standalone formal language definition. These discrepancies stem from the very nature of annotations. Annotations are meant to annotate program elements of their host @EL. We have already seen it in the example from Figure 1. The `@Entity` annotation annotated the `Person` class. The `@Id` annotation annotated the `id` field. And so on. In Figure 2 these relations based on annotating are represented by dashed arrows from the AST of annotations to the simplified AST of the host @EL. The use of annotations in context of the host language creates an unusual aspect of abstract syntax that is not present in a standalone language – *relations between annotations and host program elements* that are annotated. In general, grammar defines restrictions for relations between concepts of the formal language. However, in case of an annotation-based language, there are also restrictions on relations between annotations (@L concepts) and host language elements (not @L concepts).

[7] If an annotation type declares a `value` parameter then the name of the parameter can be omitted from the annotation instance as in case of `@AuthorRef`.

These restrictions are characteristic for annotations and in language theory they correspond to language composition.

We distinguish two types of relations between annotations and host language program elements, based on the relation direction:

- *code-wise relations* define annotations' requirements posed on their target elements and the program they are used in, and
- *reversed code-wise relations* specify host language's requirements for the annotations to be used in a context of a given host program element.

In the following sections we will discuss in detail both types.

## A. Code-wise Relations

*Code-wise relations* are relations defining restrictions by annotations to their target program elements. Each annotation may specify some requirements for its target program element. These annotations cannot annotate arbitrary host language element, because the given restrictions have to be kept. Basically, code-wise relations specify on which program elements annotations can be used.

As a standard example of enforcing existing code-wise relations we can mention Java's `@Target` metaannotation (annotation that annotates annotation types). Using the `@Target` metaannotation we can specify what types of Java program elements can a given annotation annotate. E.g., we can use `@Target(ElementType.METHOD)` to restrict annotations to annotate only methods. In C# the support for this type of restrictions is even of a finer grain. However, there are still many useful restrictions that cannot be defined by standard tools. E.g., we cannot restrict an annotation to annotate only an implementation of some interface, such as restricting the `@WebServlet` annotation to annotate only an implementation of the `Servlet` interface.

To illustrate it on a real world example, we can take a look back at the JPA annotations. The code snippet in Listing 9 replays the Person entity mapping with an added field for the person's age. However, JPA does not allow mixing annotating fields and getters/setters. In this case, the last `@Column` annotation will not be processed by the JPA implementation. In case of the `age` field, the JPA `@Column` annotation annotates the setter method, not like in the case of the other fields, which violates JPA specification (we could say it violates the JPA annotations abstract syntax).

Listing 9. Incorrect JPA mapping

```
@Entity(name = "PERSON")
public class Person {
  // an annotated field - OK
  @Id
  @Column(name = "PER_ID")
  private int id;

  // another annotated field - OK
  @Column(name = "NAME")
  private String name;

  private int age;

  // an annotated setter - incorrect
  // (it should annotate the field)
  @Column(name = "FIRST_NAME")
  public void setAge(int age) { ... }
  ...
```

## B. Idioms in Code-wise Relations

There are common structure stereotypes in code-wise relations as well. They indicate that this aspect of annotations usage structure has to be considered a part of annotations abstract syntax. Again, listed idioms are not only an evidence of the discussed aspect of the abstract syntax, but also can be inspirational for annotations' authors. Following is a list of known code-wise relations idioms.

**Target restriction** idiom specifies which types of target program elements can be annotated by the annotations. This idiom is implemented both in Java and C# as a standard validation. Noguera et al. [1] introduces finer-grained validations of the Target restriction idiom on Java platform that they call AvalTarget. Kellens et al. [15] focus on designing a framework for declaring arbitrary requirements on annotations' target program elements. Code sample in Listing 10 shows a standard Target restriction implementation in Java. The `@Override` annotation can annotate only methods, because only methods can override a behaviour. In addition we would want to restrict `@Override` to methods that really override inherited methods. In Java this checking is implemented in standard compiler for the `@Override` annotation, but in general checking like this is left to the annotations author.

Listing 10. Target restriction idiom example

```
// compile-time error
@Override
public class Clazz {

  // valid use (toString overrides inherited method)
  @Override
  public String toString() { ... }
}
```

**Refers to element** idiom specifies that an annotation parameter is a reference to program element from @EL (other than annotations target program element). The idiom is described by Ruska et al. in [13] as Annotation values referencing other elements idiom. A specialized version of the idiom where the annotation refers to a class is described by Guerra et al. [12] as an Associative Annotation idiom. Code sample from Listing 11 shows a usage of the Refers to element idiom to refer to a specific validation implementation class for an annotation type through a parameter of the `@Validator` metaannotation. The `Validation` interface defines validation operation for a given annotation annotated by the `@Validator` metaannotation. The concrete `OverrideValidation` defines specific validation for the `@Override` annotation.

Listing 11. Refers to element idiom example

```
public @interface Validator {
  Class<? extends Validation> value();
}
...
@Validator(OverrideValidation.class)
public @interface Override { }
```

**Type** idiom requires that target program element yields or conforms to a specific runtime type in host @EL. An example might be the type of the field, or the interface of the class. This idiom is described by Noguera et al. [1]. One of the uses for the Type idiom is using annotations as type modifiers (e.g., `@NotNull` annotating only reference types). The following code in Listing 12 shows a valid use of `@WebServlet`. The `@WebServlet` annotation registers a servlet to the web container. Therefore the annotation is supposed to annotate only an implementation of the `Servlet` interface.

Listing 12. Valid Type idiom example

```
@WebServlet(urlMappings={"/MyApp"})
public class MyAppServlet implements Servlet {
  ...
```

### C. Reversed Code-wise Relations

*Reversed code-wise relations* pose restrictions on the host language according to the annotations that are present in the code. E.g., the same way as we can specify the type of the field of a class is the `Person` class, we could also require that the field type (in this case the `Person` class) has to be annotated by a particular annotation (e.g., `@Entity`). Or, we may want an argument of a method to be of a type that is annotated by a particular annotation. Basically, reversed code-wise relations specify how can host language use annotated program elements.

Such requirements we call reversed code-wise relations because they revert the direction that we use to look at annotations and their binding with the host language. Code-wise relations specify in which cases the annotations are not correctly used. Reversed code-wise relations specify in which cases the host language source code is not correctly used with respect to annotations. According to code-wise relations the annotation is invalid if its target program element (or some other program element of the host language) does not exhibit some characteristics. According to reversed code-wise relations the program element is incorrectly used if it or some related program element misses a particular annotation (or has an incorrect one). This kind of relations increases integration of annotations into the host language.

As an example of an error caused by violating this type of requirements is a quite common error in using dependency injection in Enterprise Java. Code sample in Listing 13 shows an invalid use of dependency injection. Both classes are marked as stateless beans and therefore should be managed by a container. `@EJB` annotation marks a server of a mail service implementation to be injected. However, then the `Notifications` class takes control of the `MailService` instantiation by using the `new` keyword. When the objects are created like this, the dependency injection annotations will not take effect and the server will not be injected. This is a case of reversed code-wise relation, because the annotations are used properly, however, the code that uses annotated program elements is not correct. The `MailService` instance should be obtained through the `InitialContext` to be correct.

Listing 13. Using `@EJB` for dependency injection

```
@Stateless
public class MailService {

  @EJB
  private Server server;

  public void sendMail(String mail) {
    server.send(mail);
  }

  ...
}

@Stateless
public class Notifications {
  public void notify() {
    MailService ms = new MailService();
    ms.sendMail(
       "This will throw NullPointerException.");
  }
}
```

### D. Idioms in Reversed Code-wise Relations

We have recognized and identified following two reversed code-wise validations:

**Annotated type** idiom poses a restriction upon entities[8] of the host @EL. It requires that an entity type has to be annotated by an annotation. E.g., a method argument might require an object of a type annotated with a specific annotation. This idiom can enhance the object-oriented dynamic binding of the @EL. Using this idiom a method could require object of any type but annotated with a particular annotation, or with particular annotations in its scope. In the code sample from Listing 14 we use the `serialize` method to serialize an object of the `Person` class. However, using the Annotated type idiom the method serialize requires that the type of the object to be serialized to be annotated by the `@Serializable` annotation (e.g., instead of Serialize marker interface). Therefore the `Person` class has to be annotated by the `@Serializable` annotation for the example to be valid.

Listing 14. Annotated type idiom example

```
public class Serializer {
  // serialize requires argument with
  // type annotated by @Serializable
  public static void serialize(
      @AnnotatedType(Serializable.class)
      Object object) { ... }
}

...

public static void main(String[] args) {
  Person person = new Person("Milan");
  Serializer.serialize(person);
}

...

// for program to be valid, the Person
// class has to be annotated with @Serializable
@Serializable
public class Person { ... }
```

[8] Method argument, variable, etc.

**Annotated program element** idiom requires that the program elements which exhibit some particular characteristics must be annotated by a particular annotation. This idiom is very close to the work of Kellens et al. [15] that uses dependencies like these for co-evolution of annotations with source code. As an example we can use the `@Override` Java annotation in Listing 15. If a method is overriding a method from a superclass it has to be annotated by the `@Override` annotation. If the annotation is not used the program should be incorrect (as in case of the `override` keyword in C#). Code in Listing 15 shows examples of both valid and invalid program elements.

Listing 15. Annotated program element idiom example

```
public class A {
  public void a() {}
  public void b() {}
}

public class B extends A {

  // invalid program element
  public void a() {}

  // valid program element
  @Override
  public void b() {}
}
```

## V. Abstract Syntax Correspondence Summary

According to the presented discussion we can summarize our conclusions as follows. Considering the abstract syntax (AS) of a language defined by source code annotations we have to consider following components of annotations' AS:

- relations between annotations – **structural restrictions**,
- annotations' requirements on @EL concepts – **code-wise restrictions**, and
- @EL concepts' requirements on annotations – **reversed code-wise restrictions**.

While the relations between annotations match the standard abstract syntax of a standalone formal language, the other two are specific for annotations. The code-wise and reverse code-wise restrictions can be in general compared to embedding as a form of language composition, but annotations tend to come with more complex constraints on their usage and in case of an embedded language the embedded portions are usually not dependent on each other (without structural restrictions). And, from the implementation viewpoint, embedded languages usually have own parser.

If we consider a set of related annotations an *annotation-based language*, then based on our discussion we will define the abstract syntax of an @L as in Definition 2. So far there is not a standard formal apparatus for describing the full AS of @L. However, there are several approaches to this problem in the academy that we will review in section IX.

**Definition 2.** *The abstract syntax of an @L is a set of structural, code-wise and reversed code-wise restrictions on @L annotations' usage. These restrictions represent formation rules that specify a valid @L sentence.*

## VI. Concrete Syntax Correspondence

Of course, the correlation between annotations and formal languages does not end with abstract syntax aspect of the formal language description. Clearly, the annotations themselves need to have a concrete way of presentation (and serialization). In this section we will take a look at how the concrete syntax of annotations is defined.

@L author can design the @L in terms of the abstract syntax as we have already discussed. Of course, it is only logical that she should be also able to specify how the annotations of the @L will be presented in the source code of the host language. In most common @EL the apparatus for the concrete syntax definition are *annotations type*s. A specific of annotations in terms of concrete syntax when compared to formal languages is annotations' binding to target program elements. Both of these aspects will be discussed in following sections.

### A. Annotation Types

Annotations are instances of annotation types the same way as objects are instances of classes in object-oriented programming. An annotation type is a definition of a structure of a set of annotations. It defines what parameters can an annotation have, what is the name of the annotation, etc. Using annotation types the @L author can specify what are the @L terminal symbols – annotations' and parameters' names, and their values' types. Just for illustration we can take a look at the code snippet in Listing 16 presenting a simple annotation in C#. This annotation type defines annotations with name "Configuration" and two parameters, the first an integer identified as "paramId", and thesecond a string identified as "paramValue". If the @L consisted only of annotations of this type, we could easily identify its lexical symbols - "Configuration", "paramId", "paramValue", integer value and a string. The Java version of the same annotation type is presented in Listing 17.

Listing 16. C# version of the `Configuration` annotation type

```
public class Configuration : System.Attribute
{
  public int paramId;
  public string paramValue;
}
```

Listing 17. Java version of the `Configuration` annotation type

```
public @interface Configuration
{
  public int paramId();
  public string paramValue();
}
```

Another interesting fact the reader might have noticed is that the annotation types can partially define abstract syntax. If one of the parameters accepted annotations[9], then the annotation would be an instance of the Composite annotation idiom.

However, even in CS there are discrepancies between annotations and formal languages. The CS of annotations is restricted to follow some rules in order to be parsable by the

[9] Currently, annotation nesting is not supported in C#, but there are implementations that allow it, e.g., Java annotations.

standard @EL parser (or the 3rd party tool). In @EL GPLs those rules are defined by the GPL grammar.

Let us take a look at a set of grammar rules[10] of Java 8 that specify the grammar for Java annotations that are listed in Listing 18.

Listing 18. Grammar rules for the Java 8 annotations

```
Annotation ->
   NormalAnnotation | MarkerAnnotation
   | SingleElementAnnotation

NormalAnnotation ->
   ’@’ TypeName ’(’ [ElementValuePairList] ’)’

ElementValuePairList ->
   ElementValuePair {’,’ ElementValuePair}

ElementValuePair -> Identifier ’=’ ElementValue

ElementValue ->
   ConditionalExpression
   | ElementValueArrayInitializer | Annotation

ElementValueArrayInitializer ->
   { [ElementValueList] [’,’] }

ElementValueList ->
   ElementValue {’,’ ElementValue}

MarkerAnnotation -> ’@’ TypeName

SingleElementAnnotation ->
   ’@’ TypeName ’(’ ElementValue ’)’
```

These grammar rules specify CS restrictions on Java annotations. As we can see, annotations have to start with the ’@’ sign, followed by the annotation name. Then there are optional annotation parameters enclosed in parentheses, and so on. Considering we had an Java annotation type for the `Configuration` annotation type presented in Listing reflst:javaAnnType, then a concrete Java annotation would look like the annotation in Listing 19.

Listing 19. Java `@Configuration` annotation example

```
@Configuration(paramId=1, paramValue="new")
public class A {
  ...
```

Of course, the different attribute-oriented programming implementations might have different restrictions. E.g., in C# the annotations are not prefixed by the ’@’ but enclosed with brackets, and have some other minor differences from Java annotations. The same `Configuration` annotation in C# (annotation type in Listing 16) would look like the code in Listing 20.

Listing 20. C# version of the `Configuration` annotation

```
[Configuration(paramId=1, paramValue="new")]
public class A {
  ...
```

In a conventional formal language the language author is usually not restricted by any such rules. In this aspect annotations resemble more generic languages[11], such as XML languages, and alike; that are built around a given syntactic skeleton.

### B. Binding Rules

@L CS has another specific aspect – *binding rules*. Each annotation *marks* (annotates) its target program element. This relationship is expressed by the relative position of annotation to its target element. Again, the binding rules for a specific @OP implementations might differ. These rules have to ensure that for each annotation there will be unambiguous mapping to its target language element and that for each program element there will be unambiguous way of finding its annotations. The most common binding for annotations is using annotations as prefixes[12]. E.g., the Java annotations are considered modifiers and therefore they prefix declarations just as Java modifiers do. The reader has seen already seen multiple examples of Java annotations usages. This can be also illustrated by Java grammar for class modifiers in the Java 8 grammar excerpt[13] in Listing 21. The excerpt shows that class annotations prefix the class declarations.

Listing 21. Grammar excerpt for class modifiers showing in-place binding of Java 8 annotations

```
NormalClassDeclaration ->
   {ClassModifier} ’class’ Identifier ...

ClassModifier ->
   Annotation | ’public’ | ...
```

Since Java 8, annotations are supported on types, too. Listing 22 is a grammar rule[14] for primitive types that shows type annotations allowing to annotate types.

Listing 22. Grammar rule illustrating type annotations

```
PrimitiveType ->
   {Annotation} NumericType
   | {Annotation} ’boolean’
```

There might be additional rules for binding, as for example a restriction in Java requiring that two annotations of the same type cannot annotate the same target program element.

### C. Summary

Each particular @EL defines its own concrete syntax skeleton for adding annotations to its source code. While the restrictions posed by @EL have to be kept, the @L author can use annotation types to define the rest of the concrete syntax. Therefore we will define the concrete syntax aspect of @L by Definition 3.

**Definition 3.** *The concrete syntax of an @L is specified by restrictions posed by the host @EL in combination with the*

[10] Source: http://docs.oracle.com/javase/specs/jls/se8/html/jls-19.html

[11] Mernik [16] calls a generic language Commercial Off-The-Shelf (COTS). We find the *generic language* term by Chodarev et al. in [17] more intuitive and therefore we will keep using this term.

[12] However, in general, annotations do not have to prefix annotated program elements. It is sufficient to have an unambiguous rule of how to relate annotations with their target language elements.

[13] The ellipsis (...) indicates that we have shortened the rules to show just the relevant parts. Source: http://docs.oracle.com/javase/specs/jls/se8/html/jls-19.html

[14] Source: http://docs.oracle.com/javase/specs/jls/se8/html/jls-19.html

*set of concrete annotation types of annotations that belong to the @L.*

## VII. Semantics Correspondence

According to Kleppe [18], a sound language description would not be complete without a semantics description. And of course, the same applies to annotation-based language. There are multiple ways of describing semantics of a language. An annotation-based language is usually described by dynamic semantics that is defined by the tool processing the annotations (the *reference implementation* [18]).

There are two approaches to @L reference implementation:

- **compile time processing** is implemented as a pluggable annotation processor plugged to the host @EL compiler, and
- **runtime processing** is implemented as a reflection mechanism.

*Compile time processing* is implemented as a pluggable annotation processor. The host @EL parser creates an AST with annotations and provides it to all registered annotation processors. The AST with annotations can be used to generate code or other software artefacts, to generate documentation or even to manipulate the AST. E.g., in Java there is a standard implementation of pluggable annotation processing API released under JSR 269 specification[15]. This standard annotation processing API does not support AST modification and can be only used to generate new artefacts. An alternative to JSR 269 is a Spoon API by Pawlak [19] that enables fine grained source code modifications.

*Runtime processing* is implemented as an API that allows some form of runtime reflection for querying annotations. Runtime processing is usually used to read configuration of frameworks and programs. Languages such as Java or C# provide standard Reflection API that can be used to query for annotations on program elements such as classes, method, etc. These can be used to find out whether there is a particular annotation annotating the chosen program element. However, these APIs do not enable searching for annotations in a set of program elements (e.g. finding all the occurrences of specific annotation on all the classes on classpath) and likewise operations. These types of queries common in compile-time annotation processors. The need for the same feature in runtime is reflected by commercial runtime APIs, such as Scannotation[16] or Google Reflections[17].

An interesting hybrid of the compile-time and runtime processing is aspect-oriented programming (AOP) with annotations. Annotations in AOP can be used to bind aspects to program elements. This way we can add, remove, or modify the code. The final weaving of the aspect may happen both during compile-time and runtime, depending on used AOP implementation.

Each annotation-based language defines its semantics using one of the discussed approaches. Thus, we define @L operational semantics by Definition 4.

**Definition 4.** *The @L semantics is described by reference implementation using a pluggable annotation processor or a GPL code using reflection API. Reference implementation may use convenience frameworks, such as Google Reflections.*

## VIII. Annotation-based Language

We presented our observations indicating a close correspondence between source code annotations and formal languages. We proposed definitions of the three main annotation-based language definition components - @L abstract syntax, concrete syntax and semantics. Based on the presented discussion we propose to define a term *annotation-based language* to describe a given set of annotations that are processed by the same reference implementation with the same goal. For example, if we have a set of JPA annotations used to describe mapping of Java classes to relational database that are processed by a JPA implementation (e.g., Hibernate), we can consider them an @L. Our formulation of the @L definition is presented in Definition 5. In it we assume that the same reference implementation implies the same annotations problem domain (e.g., object-relational mapping).

**Definition 5.** *Annotation-based language (@L) is a set of all annotations and their parameters (*alphabet*) processed by the same reference implementation. It is defined by the reference implementation (*semantics*), structural, code-wise and reverse code-wise restrictions (*grammar – abstract syntax*), and their annotation types (*grammar – concrete syntax*).*

For example, if we have a set of JPA annotations used to describe mapping of Java classes to relational database that are processed by a JPA implementation (e.g., Hibernate), we can consider them an @L. This @L is of course defined by all three language aspects of CS, AS and semantics according to definitions we have proposed.

We have also noticed that the main source of the discrepancies between @L and a conventional formal language is the binding of annotations to target elements of the host language. Therefore we will emphasize the importance of the binding in the @L. We can therefore identify two main components of the @L:

- @L **concepts** represented by annotations and their structural relations, and a
- *meaningful* **binding** between concepts from @L and host @EL (represented by code-wise and reverse code-wise relations).

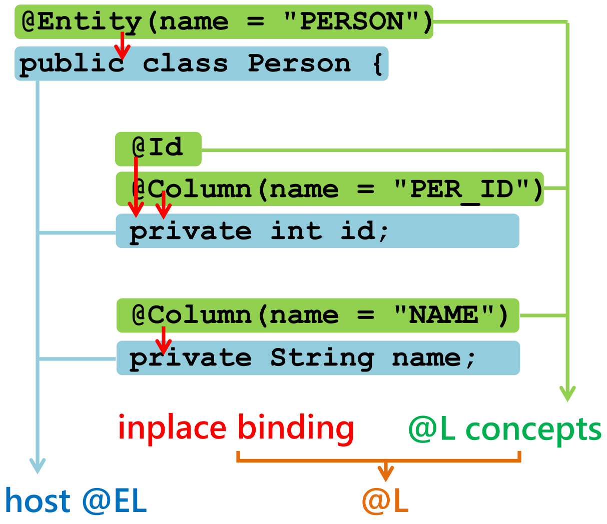
These two @L components are illustrated on the example with JPA mapping of the Person class in Figure 3. @L concepts are the annotations themselves, the binding maps the annotations to host language program elements.

We expect the binding between annotations and their target elements to be meaningful. That means that changing the target element of an annotation should also change the meaning of

[15] https://www.jcp.org/en/jsr/detail?id=269

[16] http://scannotation.sourceforge.net/

[17] https://github.com/ronmamo/reflections



```
@Entity(name = "PERSON")
public class Person {

    @Id
    @Column(name = "PER_ID")
    private int id;

    @Column(name = "NAME")
    private String name;
```

Fig. 3. Illustration of concepts and binding @L components on JPA example

the @L sentence to which the annotation belongs. E.g., if we consider the example of JPA configuration in Listing 4, moving the `@Id` annotation from the `id` field to the `name` field would define a different object-relational mapping, although the @L sentence would consist of the same annotations.

## IX. Related Work

There are several works that aim at annotations' relations definition[18]. They either provide a framework to define and validate more-or-less arbitrary dependencies between annotations and their target elements or they provide implemented common stereotypes in @L structure that can be reused for validation. These idioms show common stereotypes in annotation-based abstract syntax definition. Most of the works were already referred throughout the paper, in this section we will provide a brief summary.

Darwin [21] devised a DSL for dependencies description. His aim, contrary to the other works, was to provide a framework for describing dependencies for third party @Ls. The author of rules was to be a user of @L and not its author. Ruska et al. [13] identify several structural idioms and provide a framework for checking dependencies. In their approach they store the source code model into a database and run SQL queries representing constraints. They designed a Prolog-like DSL for convenient rules writing. Although both of these works do not explicitly mention the term annotation-based language (or any synonym) they admit that annotations have some structured dependencies that we consider a recognition of abstract syntax of @L.

Kellens et al. [15] introduce so called Smart Annotations. Smart Annotations declare their sufficient and necessary requirements. Necessary requirements declare what characteristics the annotated element must exhibit so the annotation usage is valid. Sufficient requirements are dual to necessary and they declare that if there is an entity in the code exhibiting characteristics required by it then it should be annotated by the annotation. Their work is focused on code-wise dependencies and they aim to provide better control of evolution of annotated software. Their sufficient requirements exhibit characteristics of the reversed code-wise dependencies.

Cepa et al. [14] is one of the first works that concerns checking dependencies between annotations and the host @EL. Although their framework was quite limited in supported restrictions[19], in their work they also look at annotations as a form of a language framework. They realized that annotations can be used to enhance the host GPL with domain-specific concepts thus acknowledging @OP as a form for DSL-like language extensions.

Noguera et al. [1] went even further than Cepa et al. and stated that *"a set of annotations dedicated to a given domain-specific concern can be referred to as an Attribute Domain-Specific Language (AttDSL)"*. They therefore consider annotations a language and not only a language extension. They distinguished between structural and code-wise dependencies (we used their naming).

Song et al. [22] introduce metadata invariants that are not exclusive for annotations but work for XML languages, too. They use Metadata Invariant Language (MIL) to write invariants for annotations or XML documents. MIL can be used to check structural and code-wise dependencies. Their work assumes that for an @L there can be a standalone external language that shares part of the @L semantics (multiple concrete syntaxes). We have utilized the same property of a subset of @Ls in our previous work [23]. We designed a tool that is able to create AST for @L and combine it with concepts from a separate XML-based language that is an alternative notation for the @L. The tool takes a mapping between the two formats.

The furthest advance towards @L we consider the work of Noguera and Duchien [24]. Not only they recognize @L, but they also designed a method for creating Annotation Model, a UML model of annotation-based language. They use this model and its binding to code model (AST of the host @EL) to check annotations restrictions.

Cepa in his book about @OP [5] connects the annotations to languages too. He mentions that annotations can be seen as a convenient way of extending the grammar of the language. With the work of Noguera et al. [1] Cepa's ideas were the main inspiration for this work. However, in his work he considers the annotations only as an extension of the grammar and as an alternative for adding domain-specific abstractions to the source code, and only briefly mentions annotations as a restricted form of embedded DSL. He does not further examine the correspondence between annotations and languages.

[18] We focus solely on the source code annotations defined in section II as a language implementation strategy. Therefore we will not discuss works such as Bonenfant et al. [20], which also use the term *annotation language*. However, they use it in different context than us. In their work the annotations are any custom metadata (their *annotation language* is an XML-based language).

[19] E.g., it cannot check dependencies between annotations on program elements in the same scope in program model.

From the viewpoint of existing @Ls, we could find multiple sets of annotations defined for the same domain and processed by the same semantics reference implementation. There are sets of annotations defining GUI from data model [25], annotations used for plugin extension definition [26], @L for design patterns definition [27], @L for model checking [28], and so on.

There are not many works concerning @L design, however, one can find some inspiration in analysis provided by Mancini et al. [29]. They discuss options of using annotations and their design for data validation definition. All the above mentioned works about annotations usage restrictions define some annotations AS idioms that can be used for @L design as well. Guerra et al. [12] discuss solely annotations idioms both in CS and AS. Correia et al. [6] discuss bad smells that can be result of bad annotations design or their usage. They also provide a set of possible solutions to remove them. Correia et al. [6] show that annotations syntax can be as important as their domain usability [30].

## X. Conclusion

This paper presented our observations about correspondence between source code annotations and formal languages. The correspondence was illustrated on three definition aspects of formal languages – abstract syntax, concrete syntax and semantics. For each of these aspects also the discrepancies were discussed. While the correspondence indicates that we should look at annotations as a form of language (and therefore we define the term *annotation-based language* for a set of domain-specific annotations), the discrepancies identify the main specificity of annotations that separates it from conventional formal languages. This specificity is the binding of annotations to their host annotation-enabled language.

Realizing the correspondence between annotations and formal languages we can draw some consequences for future research directions. For example, XML generic language provides several mechanisms to define an abstract and concrete syntaxes of concrete XML languages, e.g., XML schema definition, Document Type Definition, etc. In practice we can notice notorious lack of similar mechanisms for @Ls. Annotation types are usually not sufficient to describe the relations that are common in @L. A common practice is to describe the grammar using natural language documentation. Considering the presented correspondence, then instead of natural language documentation, the formal methods of abstract syntax definition can be applied. Methods such as EBNF are commonly known and therefore their application can make the annotation-based language syntax more comprehensible. In section IX we discussed several academic works that implement frameworks for @L abstract syntax definition, but so far none of them became industrial standard. So a logical consequence of the correspondence is the *need for standard AS definition formalism/mechanism for annotations*. In this matter there was a small step further in Java 8 type annotations.

The standard AS definition formalism/mechanism for annotations cannot be restricted to mere @L AS validation. Another important feature is the ability to create a fully annotated AST. In Figure 2 we have sketched the AST with annotations' relations in mind. However, currently the @EL implementations do not support @L AS and therefore the AST nodes representing annotations have no explicit relations with other annotations (unless annotation types support it, such as in case of Composition annotation idiom in Java). Supporting these relations explicitly could prove beneficial for @L authors and their semantics implementation.

Another observation that is related to our discussion is the lack of unified API for semantics reference implementation in runtime and in compile time. Currently, standard @EL implementations provide two distinct APIs for annotation processing and for runtime reflection. This observation was authored by Cepa [5] quite a long time ago, but to our best knowledge no real advance was made in industry in this matter.

In general we can consider annotations a *generic embedded language*. They provide a syntactic skeleton that on one hand restricts @L author in concrete and abstract syntax, but on the other hand provides standard tools for their parsing (either the host language parser or a third party tool). But in contrast to generic languages they are restricted to embedding into the host language; they have to annotate its program elements. Based on this observation, in our future work we want to analyze options of using annotations for language composition. From the observations presented in section IV we learned that annotations' specificity is their binding to host language. In the future work we want to examine types of language composition that can utilize annotations as an implementation technique.

## References

[1] C. Noguera and R. Pawlak, "AVal: an extensible attribute-oriented programming validator for Java: Research Articles," *Journal of Software Maintenance and Evolution*, vol. 19, no. 4, pp. 253–275, Jul. 2007. http://dx.doi.org/10.1002/smr.349

[2] R. Rouvoy and P. Merle, "Leveraging Component-Oriented Programming with Attribute-Oriented Programming," in *Proceedings of the 11th International ECOOP Workshop on Component-Oriented Programming*, ser. WCOP 2006, 2006.

[3] S. Chodarev, D. Lakatoš, J. Porubän, and J. Kollár, "Abstract syntax driven approach for language composition," *Central European Journal of Computer Science*, vol. 4, no. 3, pp. 107–117, 2014. http://dx.doi.org/10.2478/s13537-014-0211-8

[4] D. Lakatoš, J. Porubän, and M. Bačíková, "Declarative specification of references in DSLs," in *2013 Federated Conference on Computer Science and Information Systems*, ser. FedCSIS 2013, Sept 2013, pp. 1527–1534.

[5] V. Cepa, *Attribute enabled software development: illustrated with mobile software applications*. Saarbrücken, Germany: VDM Verlag, 2007.

[6] D. A. A. Correia, E. M. Guerra, F. F. Silveira, and C. T. Fernandes, "Quality Improvement in Annotated Code," *CLEI Electron. J.*, vol. 13, no. 2, 2010, article ID 7.

[7] M. Nosáľ and J. Porubän, "XML to Annotations Mapping Definition with Patterns," *Computer Science and Information Systems*, vol. 11, no. 4, pp. 1455–1477, 2014. http://dx.doi.org/10.2298/CSIS130920049N

[8] J. Kollár, I. Halupka, S. Chodarev, and E. Pietriková, "pLERO: Language for grammar refactoring patterns," in *2013 Federated Conference on Computer Science and Information Systems*, ser. FedCSIS 2013, Sept 2013, pp. 1503–1510.

[9] C. Noguera, A. Kellens, D. Deridder, and T. D'Hondt, "Tackling Pointcut Fragility with Dynamic Annotations," in *Proceedings of the 7th Workshop on Reflection, AOP and Meta-Data for Software*

*Evolution*, ser. RAM-SE '10. New York, NY, USA: ACM, 2010, pp. 1:1–1:6. http://dx.doi.org/10.1145/1890683.1890684

[10] W. Cazzola and E. Vacchi, "@Java: Bringing a richer annotation model to Java," *Computer Languages, Systems & Structures*, vol. 40, no. 1, pp. 2–18, 2014, special issue on the Programming Languages track at the 28th ACM Symposium on Applied Computing. http://dx.doi.org/10.1016/j.cl.2014.02.002

[11] S. Zawoad, M. Mernik, and R. Hasan, "FAL: A forensics aware language for secure logging," in *2013 Federated Conference on Computer Science and Information Systems*, ser. FedCSIS 2013, Sept 2013, pp. 1579–1586.

[12] E. Guerra, M. Cardoso, J. Silva, and C. Fernandes, "Idioms for Code Annotations in the Java Language," in *Proceedings of the 17th Latin-American Conference on Pattern Languages of Programs*, ser. Sugar-LoafPLoP, 2010, pp. 1–14.

[13] Š. Ruska and J. Porubän, "Defining Annotation Constraints in Attribute Oriented Programming," *Acta Electrotechnica et Informatica*, vol. 10, no. 4, pp. 89–93, 2010.

[14] V. Cepa and M. Mezini, "Declaring and Enforcing Dependencies Between .NET Custom Attributes," in *Generative Programming and Component Engineering*, ser. Lecture Notes in Computer Science, G. Karsai and E. Visser, Eds. Springer Berlin Heidelberg, 2004, vol. 3286, pp. 283–297. http://dx.doi.org/10.1007/978-3-540-30175-2_15

[15] A. Kellens, C. Noguera, K. De Schutter, C. De Roover, and T. D'Hondt, "Co-evolving Annotations and Source Code through Smart Annotations," in *14th European Conference on Software Maintenance and Reengineering*, ser. CSMR 2010, 2010, pp. 117–126. http://dx.doi.org/10.1109/CSMR.2010.20

[16] M. Mernik, "An object-oriented approach to language compositions for software language engineering," *Journal of Systems and Software*, vol. 86, no. 9, pp. 2451–2464, 2013. http://dx.doi.org/10.1016/j.jss.2013.04.087

[17] S. Chodarev and J. Kollár, "Language Development Based on the Extensible Host Language," in *Proceedings of CSE 2012 International Scientific Conference on Computer Science and Engineering*. EQUILIBRIA, s.r.o., 2012, pp. 55–62.

[18] A. Kleppe, "A Language Description is More than a Metamodel," in *4th International Workshop on Language Engineering*, ser. ATEM 2007, 2007.

[19] R. Pawlak, "Spoon: Compile-time Annotation Processing for Middleware," *IEEE Distributed Systems Online*, vol. 7, no. 11, pp. 1–, Nov. 2006. http://dx.doi.org/10.1109/MDSO.2006.67

[20] A. Bonenfant, H. Cassé, M. de Michiel, J. Knoop, L. Kovács, and J. Zwirchmayr, "FFX: A Portable WCET Annotation Language," in *Proceedings of the 20th International Conference on Real-Time and Network Systems*, ser. RTNS '12. New York, NY, USA: ACM, 2012, pp. 91–100. http://dx.doi.org/10.1145/2392987.2392999

[21] I. Darwin, "AnnaBot: A Static Verifier for Java Annotation Usage," *Advances in Software Engineering*, vol. 2010, p. 7, 2010, article ID 540547. http://dx.doi.org/10.1155/2010/540547

[22] M. Song and E. Tilevich, "Metadata invariants: checking and inferring metadata coding conventions," in *Proceedings of the 2012 International Conference on Software Engineering*, ser. ICSE 2012. Piscataway, NJ, USA: IEEE Press, 2012, pp. 694–704. http://dx.doi.org/10.1109/ICSE.2012.6227148

[23] M. Nosáľ and J. Porubän, "Supporting multiple configuration sources using abstraction," *Central European Journal of Computer Science*, vol. 2, no. 3, pp. 283–299, Oct. 2012. http://dx.doi.org/10.2478/s13537-012-0015-7

[24] C. Noguera and L. Duchien, "Annotation Framework Validation Using Domain Models," in *Proceedings of the 4th European Conference on Model Driven Architecture: Foundations and Applications*, ser. ECMDA-FA '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 48–62. http://dx.doi.org/10.1007/978-3-540-69100-6_4

[25] M. Monteiro, P. Oliveira, and R. Goncalves, "GUI generation based on language extensions: a model to generate GUI, based on source code with custom attributes," in *Proceedings of the 10th International Conference on Enterprise Information Systems*, ser. ICEIS 2008, Jun. 2008, pp. 449–452. http://dx.doi.org/10400.8/147

[26] R. Wolfinger, M. Löberbauer, M. Jahn, and H. Mössenböck, "Adding genericity to a plug-in framework," *SIGPLAN Not.*, vol. 46, no. 2, pp. 93–102, Oct. 2010. http://dx.doi.org/10.1145/1942788.1868308

[27] P. Kajsa and P. Návrat, "Design Pattern Support Based on the Source Code Annotations and Feature Models," in *SOFSEM 2012: Theory and Practice of Computer Science*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7147, pp. 467–478. http://dx.doi.org/10.1007/978-3-642-27660-6_38

[28] G. Ferreira, E. Loureiro, and E. Oliveira, "A Java Code Annotation Approach for Model Checking Software Systems," in *Proceedings of the 2007 ACM Symposium on Applied Computing*, ser. SAC '07. New York, NY, USA: ACM, 2007, pp. 1536–1537. http://dx.doi.org/10.1145/1244002.1244330

[29] F. Mancini, D. Hovland, and K. Mughal, "Investigating the Limitations of Java Annotations for Input Validation," in *Proceedings of International Conference on Availability, Reliability, and Security, 2010*, ser. ARES '10, Feb 2010, pp. 513–518. http://dx.doi.org/10.1109/ARES.2010.29

[30] M. Bačíková and J. Porubän, "Domain usability, user's perception," in *Human-Computer Systems Interaction: Backgrounds and Applications 3*, ser. Advances in Intelligent Systems and Computing. Springer International Publishing, 2014, vol. 300, pp. 15–26. http://dx.doi.org/10.1007/978-3-319-08491-6_2

DOI: 10.15439/2015F179
ACSIS, Vol. 5

# Unified Compile-Time and Runtime Java Annotation Processing

Peter Pigula and Milan Nosáľ
Technical University of Košice
Letná 9, 042 00 Košice, Slovakia
Email: peter.pigula@student.tuke.sk, milan.nosal@gmail.com

*Abstract*—**Java provides two different options for processing source code annotations. One of them is the annotation processing API used in compile time, and the other is the Reflection API used in runtime. Both options provide different API for accessing program metamodel. In this paper, we examine the differences between those representations and we discuss options on how to unify these models along with advantages and disadvantages of this approach. Based on this proposal, we design a unified Java language model and present a prototype tool which can populate a unified model during both compilation and runtime. The paper includes the designed API of this unified language model. To verify our approach, we have performed experiments to show the usability of the unified metamodel.**

## I. Introduction

SINCE the year 2004, when annotations were first introduced to Java, this programming technique is on the rise. More and more programs use annotations during compile time for code generation [1], [2], and in runtime for configuration and reflection [3]. In spite of the fact that usage of annotations in both instances is often very similar, means to access annotations and program model, as well as their representations, are different. If a developer needs to complete the same task in both compile time and runtime, she has to be familiar with both of them, and would have to create two different versions of the source codes. That is because code using annotation processing API cannot be used during runtime, and vice versa.

We came across this problem during our previous work. We designed a tool (*Bridge to Equalia* [4]) that provided an abstraction layer to two most common configuration formats in Java – XML and source code annotations. *Bridge to Equalia* (abbreviated: *BTE*) uses XML to annotations mapping patterns [5] to shorten the configuration interface code in cases when both annotations and XML should be accepted interchangeably. To query the annotations, *BTE* uses Java Reflection API in combination with the Scannotation [6] library. Therefore, it works only in runtime (Reflection API requires compiled classes). As we wanted to use *BTE* also in projects that work during compile time, we faced the problem of working with different APIs for annotation processing during in runtime and in compile time.

**Algorithm 1** Checking whether a field is public during runtime

```
boolean isFieldPublic (Field field) {
  if (Modifier.isPublic(field.getModifiers())) {
    return true;
  }
  return false;
}
```

**Algorithm 2** Checking whether a field is public during compile time

```
boolean isFieldPublic (Element field) {
  for (Modifier modifier : field.getModifiers()) {
    if (modifier == Modifier.PUBLIC) {
      return true;
    }
  }
  return false;
}
```

*BTE* is a medium size project with a complex implementation, therefore we will not use its source code as a case study for the illustration of the problem. Just as an example of the APIs' diversity we can take a look at the representation of the accessibility modifiers of Java program elements. During runtime, modifiers are encoded to a single integer value. Every bit of this integer represents one modifier. On the other hand, during compile time modifiers are represented as a set of enumeration types `Set<Modifier>`. Each element of this set represents one modifier. This means that for example a method, which checks whether a field is public or not, would be different in both runtime and compile time. Implementation of such a method during runtime is presented in the algorithm 1. Implementation of the same method in compile time is presented in the algorithm 2. We have to note that although both listings use the `Modifier` type, they are both different and belong to different packages. Of course, the *BTE* implementation works not just with modifiers, but many other types of metadata (program elements' names, annotations, annotations' parameters, code tree structures, etc.).

In this work we want to *share our experience with designing and implementing a tool that provides a unified API to Java source code model during runtime and compile time*. We will discuss challenges in the design that we faced, and we

This work was supported by VEGA Grant No. 1/0341/13 "Principles and Methods of Automated Abstraction of Computer Languages and Software Development Based on the Semantic Enrichment Caused by Communication".

978-83-60810-66-8/$25.00©2015, IEEE

will present the designed API (that was also implemented and can be found at https://github.com/mallynth/UModel). We conclude with a simple experiment that evaluates the tool's performance against standard APIs.

Since there is currently no unified representation in Java, developers who come across this problem are left with only a single option, which is to create two versions of the source code dealing with the same problem in different representations. In small tasks, this option is not that time consuming. But with increasing complexity of the processed annotations it may lead to code difficult to manage, maintain and evolve. In cases like this, unified representation could reduce the needed work by half, and apply the Single point of responsibility principle.

Having two versions of the code accessing annotations goes against *Single point of responsibility* principle. This leads to potential problems during evolution and maintenance (as we faced with the *BTE* tool). Suppose we have two programs that are used for the same purpose, but one of them is used during compile time, and the other during runtime. Degree of difference between these programs would not be very high, because they were developed using the same algorithm, and they just use different APIs. Now let us suppose that we need to implement a change of the base algorithm. That means we need to change both versions of the code separately from one another in order to implement the desired change. This increases the risk of a bug, e.g., when a developer forgets to make a change in one of the versions, or she misses an important difference in APIs. This leads to overhead in testing, because again, both versions need to be tested separately.

Cepa [7], [8] deals with similar problem about the representation of metamodel in programming languages. He recognizes the need for a generalized representation of the program by using a graph of metadata nodes that could be used in compile time, loading time and runtime. He then proposes a Generalized and Attributed Abstract Syntax Tree (GAAST), which is a syntax tree of the language enriched by annotations. The recognition for the need of GAAST support stemmed from his work on his framework MobCon [9], which is a framework used to generate mobile applications in Java. His work in combination with our previous work with *BTE* was the main inspiration for this work.

Contributions of this paper are following:

- we analyse options by which unification of both program model representations in Java can be achieved,
- we discuss and explain the pitfalls that need to be considered during the design and implementation of the unified model, and
- we present the designed API for projecting standard source code models into a unified Java program model during both compile time and runtime.

## II. Existing Program Models

Annotations in Java are used as a source code decoration mechanism. It means that they do not directly influence the control or data flow of the program that is annotated. A study [10] performed on the curated collection of programs Qualitas Corpus [11] showed that more than 60% of analyzed programs are using annotations. That includes systems developed before annotations were introduced to Java. This study provides evidence that annotations are an important part of Java.

**Algorithm 3** Declaration of annotation with source code retention

```
@Retention(RetentionPolicy.SOURCE)
public @interface SourceAnnotation {}
```

**Algorithm 4** Declaration of annotation with runtime retention

```
@Retention(RetentionPolicy.RUNTIME)
public @interface RuntimeAnnotation {}
```

Annotations can be used to express metadata of a program. They can be accessed by the developer in two different ways:

- during **compile time** using annotation processing API, and
- during **runtime** using Reflection API.

Both of these ways offer metadata to the developer in a different representation. Both representations are basically *program models* that include all metadata available at the given time. Program model available during compile time is represented by classes in a package `javax.lang.model.element` [12] and the model available during runtime is represented by classes in a package `java.lang.reflect` [13]. These models are accessed and queried using different APIs. When their difference is not important for the sake of discussion, we will refer to them simply as **basic models**.

Probably, the main reason of their difference is that they are used in a different phase of a program life cycle and different metadata are available during those phases. However, they are still similar since they represent the same program structure.

To illustrate the difference, we can mention annotations marked by the `@Retention` metaannotation. In the code fragment 5, the `SomeClass` class is annotated with two annotations `SourceAnnotation` and `RuntimeAnnotation` which are declared in code fragments 3, and 4 respectively. In this example, if we would access the metamodel during compilation, we would get both annotations. However, if we would access the metamodel during runtime, we would find out that the `SomeClass` class is annotated only with the `RuntimeAnnotation` annotation. That is because `SourceAnnotation` is marked with meta-annotation `@Retention(RetentionPolicy.SOURCE)` to indicate that it should be discarded after annotation processor finishes, and therefore it will not be available during runtime.

Another difference is the option for running methods of the code. In reflection, the developer can invoke methods unknown during compile time using reflection API. However, in compile time the API does not support invoking of processed code, since in that time the classes have not yet been compiled.

**Algorithm 5** Usage of annotations

```
@SourceAnnotation
@RuntimeAnnotation
public class SomeClass {
  ...
}
```

The differences do not end with different models. It is very common that the developer only needs some metadata that are present in both models, e.g., names of classes that are marked with a specific annotation. The problem is that the API for acquiring those metadata are different in both models, in spite of the fact that they are the same metadata. In some cases, in addition to the method of acquiring metadata being different, the representation of the metadata is also different (as in case of modifiers discussed in the introduction).

Examples of similar elements in both models can include names of classes, their methods or hierarchy of classes.

### A. Annotation Processing Program Model

Annotation processor is executed before compilation of the program and the execution itself is divided into separate rounds. Annotation processor can acquire metadata that are present in source code as well as metadata that were created in previous rounds of this processor execution.

Main class, that is used to represent entities in the program model for this API is the `Element` class. Instances of this class can represent multiple different entities, such as classes, interfaces, methods or fields.

*1) Accessing Metadata in Annotation Processor:* The main method that every annotation processor must have is the `process` method. As its name suggests, this method is doing the annotation processing. One of the parameters of this method is an environment of current round represented by the `RoundEnvironment` class. From this environment, the developer can acquire metadata that she needs.

One of the methods for acquiring metadata from round environment is by using the method `getRootElements`. This method returns list of all root elements of the program that are represented by `Element` class. The `Element` class has a field `kind` which defines what this element represents (class, interface, method etc.). By querying these elements, it is possible to find metadata that are needed.

Different way would be to use the `getElementsAnnotatedWith` method which returns a list of elements that are marked with a specific annotation. When using this method, there is a useful annotation `@SupportedAnnotationTypes`, which can be used to mark the annotation processor class. It specifies annotation types which annotations are supposed to be retrieved from the source codes during annotation processing.

### B. Reflection Program Model

Reflection can be used during runtime and it provides a way for accessing the metadata of the running program.

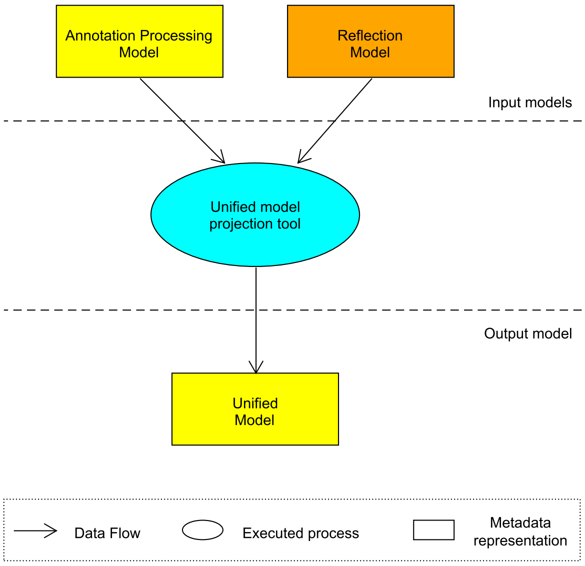


Fig. 1. Concept of the unified API idea

*1) Accessing Metadata in Reflection:* To gain access to metadata provided by reflection in a specific class, one can use `ClassName.class` parameter of every class which returns a `Class` class that represents a class with the name "ClassName". This class can be then used to retrieve metadata about the class and elements defined inside the class, such as methods, constructors, fields or annotations, or even inner classes. It is possible to retrieve lists of those elements or search for specific ones. For example retrieving the list of annotations by which the class is marked can be done by calling a method `getDeclaredAnnotations`.

By calling methods provided by classes of the reflection model, developer can acquire any metadata existing during runtime. One of the big differences in usability of the two basic models is that in reflection, methods can be executed, which is impossible in annotation processing.

## III. Program Model Unification

Due to the similarities in these models, we expected that it should be possible to design a tool that would be able to provide a unified API to data from both basic models. Simple overview of the idea is illustrated in Fig. 1. It is enough to note that input data used in the tool will always come from only one of the input models at the time. Which model it is depends on the time when the tool is used.

### A. Unified Model Projection Types

In order to create a unified model, first we need to determine how the unification tool will project basic models into the unified API.

We will illustrate the problem of unified model types on program trees. A node in the tree represents a program element (class, method, etc.) and annotations. The structure models the

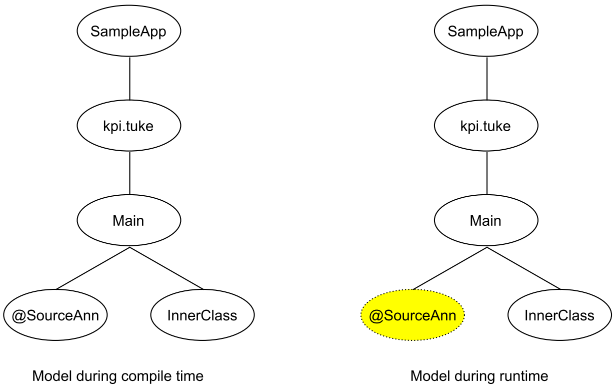


Fig. 2. Model comparison illustration

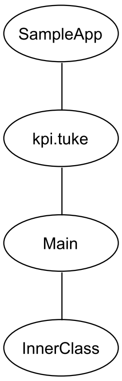


Fig. 3. Equality-based projection

encapsulation relationships of program elements. The tree's root represents Java project (library, application), its children are Java packages, the packages contain classes, or might be annotated by annotations, the nodes representing classes have children nodes representing variables and methods, and again annotations that belong to the classes, and so on. E.g., we might have a root node 'sampleApp' with a single child node 'kpi.tuke' representing package. The 'kpi.tuke' node could have a child node representing the 'Main' class, and this 'Main' node could have two child nodes: a node for the 'InnerClass' class, and a node for the '@SourceAnn' annotation, etc. Program elements and annotations are source code entities that the unified API should expose.

An illustration of both basic models for the discussed example is presented in Fig. 2. Nodes that are **equal** in both models are represented by ellipses with white color and solid border. Equal in this context means that both basic models expose the same metadata about the element in both models.

Node '@SourceAnn', which is represented by an ellipsis with yellow color and dotted border in a model during runtime shows a situation when element **does not exist** in runtime model, but exists in the compile time model. This is a result of using the `@Retention(RetentionPolicy.SOURCE)` meta-annotation that marks the `@SourceAnn` annotation to be discarded by the compiler during the compilation.

Projection to the unified model can be designed in different ways depending on which metadata would be included in it. There are three main approaches to design of a unified API:

1) **equality-based** projection,
2) **combined** model projection, and
3) **problem-specific** model projection.

First, the **equality-based** projection for the discussed example is shown in Fig. 3. In this type, the unified model includes only those elements and metadata, that are available in both basic models. That means node '@SourceAnn', which does not exist in runtime model, will not be projected to the unified model either (regardless whether created during runtime or compile time).

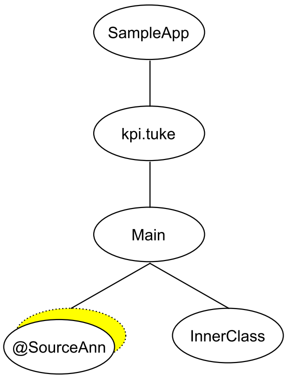


Fig. 4. Combined model projection

Second, the **combined** model projection is shown in Fig. 4. In this type the unified model exposes all the elements and their metadata that were included in the basic model. For example, if the unified model was created during compile time then the unified model would expose all the metadata that are available during compile time, including those that are not available during runtime. This projection type deals only with different APIs, and does not regard the differences of information exposed by basic models (the equality-based projection hides the differences in exposed information about the source code).

Third, the **problem-specific** projection covers cases when the projection is optimized for a given problem (something like a DSL [14] in languages). Basically it represents ad-hoc solutions for a family of problems. E.g., it might be a case of a simplified model, restricted in the tree depth to classes (therefore the only child nodes of the class nodes would be nodes for their annotations). We will not discuss this type of projections further, since its applicability is limited to a restricted set of problems.

### B. Conclusion

Both equality-based and combined model projection types can be used in our solution. Both of these types have their specific uses, but we decided to go with a model that is based on a **combined model** projection type. There is also an option of implementing both types and letting the user decide which model he needs at the time, but in current implementation we support only combined model. The reasoning behind the decision to use combined model over equality-based model is that it gives the user more information when using our tool, since the equality-based model hides metadata that are not available in both basic models.

## IV. Designing Unified Representation

When designing a unified representation, we faced several challenges. In this section we discuss some of the most important of those questions and problems.

### A. Ways to Access Metadata

One of the most important questions is how to access metadata in basic models in order to use them in a unified model. There are two possibilities how to achieve this:

1) **Direct access** to metadata with the use of an API that will unify the access methods in both basic models, and
2) **Model creation** and accessing metadata from the created unified model.

In the following sections, both of these ways are described in detail.

*1) Direct Access:* The most important thing when using **direct access** approach of accessing metadata is the interface that will be able to unify access to both basic models. In object oriented programming, this interface can be implemented using a adapter design pattern [15] [16].

In terms of accessing metadata, this design pattern can be expressed in such a way that class that needs metadata can access them by requesting them from the adapter, which handles acquiring metadata from basic model and transforms them into the unified model. Simplified example of this can be seen in Fig. 5, where the class `Client` calls a method `getMetadata` on a class `Adapter`. This method collects the requested metadata from one of the basic models which is available at the time, transforms them into the unified model that is expected and returns them to `Client`. The `Model` in this figure can represent either annotation processing model or reflection model. Which of these is currently represented depends on what part of the program life cycle it is currently in. Adapter therefore must be able to tell whether it is called during compile time or during runtime and will then choose how to retrieve metadata accordingly.

Using this method means that the ***basic model is accessed during every single call from the client***, because it has to retrieve the required metadata.

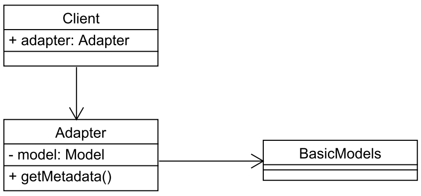


Fig. 5. Adapter Pattern

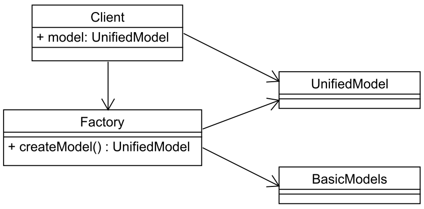


Fig. 6. Model Factory

*2) Model Creation:* The most important step when using **model creation** approach is the creation of the unified model itself. Unified model can be provided by a method similar to a factory method pattern [15].

In terms of model creation, this method will not be exactly the same as the factory method, because we know which type of class this method will return, but we do not know (and do not care) which basic model was used to create the class. Simplified example can be seen in Fig. 6, where once again the `Client` class needs access to metadata. Unlike the adapter patter, where the `Client` called a method which provided metadata, it now calls a method `createModel` on the `Factory` class. This method creates the unified model from one of the basic models and returns it to the `Client`. Same as with direct access, model that is currently represented by the `Model` class depends on what part of the program life cycle it is currently in. Therefore, just as Adapter had to be able to tell whether it was called during compile time or runtime, the `Factory` has to also be able to tell too. This method can also be compared to reverse engineering [17].

Unlike the direct access method, ***it is not needed to access basic model during every single call from the client***. Basic model needs to be only accessed once, when creating the unified model. Accessing metadata after the creation of unified model is handled by the unified model itself.

### B. Additional Costs of Accessing Metadata

In basic models, which are already part of Java, costs to access specific metadata can be easily analysed. During each request to access metadata it is required to search the basic model and find the metadata that were requested. After they are found, they simply are returned in the same format as the basic model defines. If we disregard implementation details of basic models, this process is the same for reflection model as well as annotation processing model.

*1) Direct Access:* Additional costs when using direct access method is created mainly because during every request to access metadata from basic model, it is needed to search this model, find metadata that are needed and then transform them into the specified unified metadata representation that is expected as an output of the request. In comparison to acquiring metadata from basic model, it has one additional step, which is transforming metadata to the unified model. However, this one additional step is executed during every request, and it increases the time needed to retrieve metadata.

*2) Model Creation:* Additional cost when using model creation method is the creation of the unified model itself, but this process is only executed once. After the unified model is created, all requests are handled through the unified model itself and results do not have to be transformed into the representation that is expected as an output, because they already are in form of the unified model representation.

Comparison of retrieving metadata from unified model and from basic models depends on the exact implementation of the unified model. In ideal circumstances, the access to metadata through unified model can be faster than accessing them through basic models (in case the unified model is better suited for specific requests). It is more likely, however, that accessing metadata through unified model will be slower than through basic models.

### C. Minimizing Additional Costs

In spite of the fact that reducing additional time costs to access metadata when using unified model is nearly impossible and it is not the purpose of this paper, we still have the possibility to employ some methods in order to minimize the performance impact of using unified model.

To help minimize additional costs we can use caching of results in both direct access and model creation methods.

Main idea of this in direct access method is that the results will be stored for use in another request during every request to access metadata. In every request after that, those metadata will already be at hand and it will not be needed to search the basic model for them again. Building on that method, we can make it so that it is very close to model creation method. That can be achieved by storing the cached data in such a way, that they would resemble the unified model created by model creation method. Downside of this is that we introduce a new step into the process of retrieving metadata: a step to check whether the requested metadata are already stored or not.

In model creation method, data would not be stored during requests, because even before the first request, the unified model was already built and put into memory. In this case, data would be prepared for most common requests during the model creation itself.

Because of the additional steps, caching of results runs into problems with small amount of requests and is getting better and better with more requests.

Different approach to minimizing additional costs is **selective model creation**. This method is based on the idea that user knows which metadata he or she will use and he or she can specify which parts of the model are required. User would be able to define which parts of the model will be created and which parts will be ignored. The disadvantage is that the user has to learn how to configure the tool in such a way, that it will provide desired results. If the configuration is too complicated then this disadvantage can be big enough to overshadow the biggest advantage of using a unified model, which is to make it easier for the user to gain access to metadata.

### D. Query Methods

As it was mentioned before, both of the basic models have different way of accessing the same metadata. The metadata can be queried via methods that help the user easily navigate through the model. Both basic models and their APIs provide query methods. They are primarily used to conveniently query metadata. Both API query methods and model accessor methods (getters and setter) differ in Reflection API and Annotation Processor API. Example of this difference can be the way how to retrieve list of annotations that class `A` is marked with. To retrieve metadata about these annotations in annotation processing model, developer has to call a method `getAnnotationMirrors` on the instance of the class `Elements` that represents class `A`. On the other hand, in reflection model, developer has to call a method `getDeclaredAnnotations` on the instance of the class `Class` that also represents class `A`. If both of these methods were called in the same program on the same class, their results can be different (disregarding the fact that they are represented in a different way), because some of them can be marked with meta-annotation `@RetentionPolicy.SOURCE` or `@RetentionPolicy.CLASS`, but **they represent the same metadata**, which are annotations of class `A`.

The simplest way how to design query methods is to support both names of the same methods from basic models. For the aforementioned example it would mean that there would be two query methods, one called `getAnnotationMirrors` and second one called `getDeclaredAnnotations`. These methods would always return the same results.

Better way, which is in the spirit of unification, would be to use only one method that would not have to be named exactly the same as the ones in the basic models, but it would be named name clearly enough that there would be no confusion what the purpose of that method is.

There is also a way which mimics the principles of XPath [18] for XML files. It is based on the fact that the unified model is represented in a tree-like structure that can be traversed much like XML[1]. This way is best used as an addition to the method of using one unified method, because it would provide the developer with better control over the search queries.

### E. Conclusion

For accessing metadata from basic models we decided to use the **model creation** approach. Main reason for this is that

[1] XML grammar [19] is commonly defined using XML schema [20].

in the direct access approach, searching through basic model during every request could be slow. Using the model creation approach pushes the additional time costs of using the unified API to a single point of the tool initialization.

On the other hand, the method of model creation can be designed in such a way, that when it comes to retrieving metadata from unified model, most common requests could be optimized so that they will be as fast as possible.

## V. Prototype Implementation

Based on previous analysis, in this section we present a prototype tool *UModel*[2] that can be used to create a unified model. The name *UModel* comes from the term Unified Model. Because this tool has to be able to create unified model from both basic models, it has to be able to recognize which of the models is available during that time. This is easily done, because both of the basic models have different context in which they are created and these context can be easily recognized from one another.

### A. Unified Model Creation

The abstract concept of unifying tool was illustrated in Fig. 1. In that figure, the process of creating unified model was not specified. Fig. 7 illustrates phases that are needed to create a unified model.

Model creation during runtime as well as compile time is divided into four main phases:

1) **Initialization phase**,
2) **Model creation phase**,
3) **Reference resolving phase**, and
4) **Finalisation phase**.

**Initialization phase** represents the phase in which the tool itself is initialized. That includes the initialization of structures that will be used during model creation as well as initialization of structures needed for result caching. Besides initialization of these structures, this phase also creates the basic structure of the model itself, which will be filled in during the next phase.

In the **model creation phase**, the unified model itself is created. Both runtime and compile time models are created by advancing through the provided basic model from top to bottom. Since metamodel can be represented as a tree structure, the tool starts with the root (topmost) element and gradually advances through its descendants to the leaves (bottommost elements). For every node in the basic model that the algorithm passes through, one node is created in the unified model. If the algorithm finds a reference to another element (e.g., method return type is a class included in the model), then this reference is only saved as a plain text. These references are resolved after the second phase finishes creating the whole model.

The model creation phase also saves the results for the queries encapsulated by the query methods presented in section V-D. These common queries, such as finding all the program elements annotated with a given annotation type, are this way cached and therefore should result in better time efficiency.

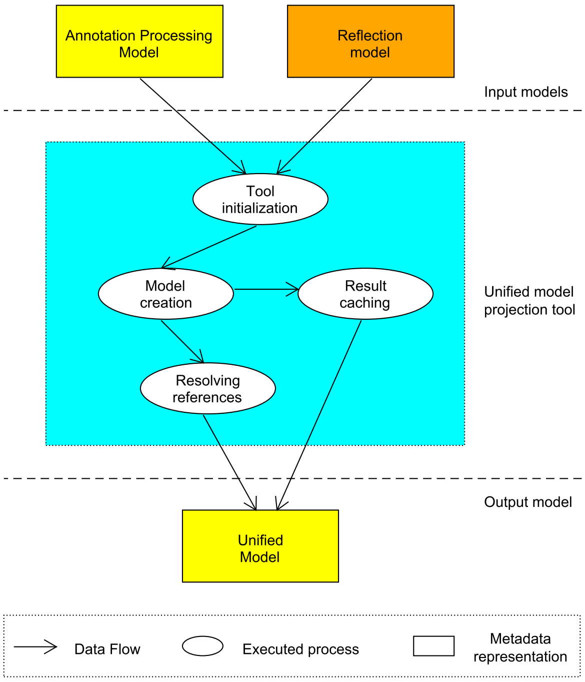


Fig. 7. Tool structure

After the second phase finishes, the **reference resolving phase** begins. During this phase, the tool resolves all of the references between the types that can be resolved (e.g., one class having a field of the type defined by another class). Reference that can be resolved is a reference to an element that is a part of the unified model. This kind of references can be used to easily traverse through the unified model. References that refer an element that is not a part of the unified model are not resolved and are left as a plain text. During this phase, the tool does not use the basic model, it works on the unified model created by phase two.

During the **finalisation phase**, the cache and the unified model (with resolved references from phase three) are joined together. After they are joined, the unified model is complete and can be passed to the tool's client.

### B. Unified Model API

This section contains description of the set of most fundamental classes representing the unified model API. All classes start with the letter U which represents their unification from both models.

The `UModelFactory` class is a factory class in a singleton design pattern which is used to instantiate unified model. It only has one factory method, the `createModel` method that returns an object of the `UModel` class.

[2] https://github.com/mallynth/UModel

*1) UModel:* This class represents the unified model that contains specific packages and classes for a given project (the 'sampleApp' node from Fig. 2). Main purpose of this class is to wrap the components of the unified model into a single root node object and to provide query methods that provide API with common queries. These query methods will be described in section V-D.

*2) UPackage:* The `UPackage` class represents a package in the source code.

Selected fields:

- `name` - full name of the package.
- `containingClasses` - set of classes implemented in this package.
- `containingEnums` - set of enumeration types implemented in this package.

Packages in the unified model *are not represented with a tree structure*. Instead of using tree structure that organizes packages into a tree with one root (as the file system does), we use a simple structure where every package is represented on its own (the standard Java model).

*3) UClass:* The `UClass` class represents a class, interface or declaration of new annotation type.

Selected fields:

- `classType` - enumeration type that defines which language element is represented by this class. Possible values are `CLASS` for classes, `INTERFACE` for interfaces and `ANNOTATION_TYPE` for declaration of new annotation type.
- `name` - name of the element.
- `enclosingPackage` - reference to an enclosing package.
- `modifiers` - modifiers represented by Integer.
- `parent` - reference to a parent element.
- `interfaces` - set of interfaces that are implemented by this class. If the `UClass` object represents an interface, then it represents which interfaces are extended and if the `UClass` represents declaration of new annotation type then this set will be empty.
- `annotations` - set of annotations represented by the `UClassAnnotation` class.
- `enums` - set of declared enumeration types.
- `constructors` - set of constructors represented by the `UConstructor` class.
- `fields` - set of fields represented by the `UField` class.
- `methods` - set of methods represented by the `UMethod` class. If `UClass` represents a declaration of new annotation type then this set contains parameters of declared annotation.

*4) UField:* The `UField` class represents a field (variable, attribute) of the class.

Selected fields:

- `name` - field name.
- `enclosingClass` - reference to a class where this field is declared in.
- `modifiers` - modifiers encoded to `Integer`.
- `annotation` - set of declared annotations represented by the `UFieldAnnotation` class.
- `type` - type of the field. It can be represented either by a string or by a reference.

*5) UMethod:* The `UMethod` class represents one method of a class. If it is included in annotation type declaration then it represents a parameter of the new annotation.

Fields:

- `name` - method name.
- `enclosingClass` - reference to a class where this method is declared in.
- `modifiers` - modifiers represented by Integer.
- `annotation` - set of declared annotations represented by the `UMethodAnnotation` class.
- `parameterTypes` - ordered list of arguments. They are represented either by a string or a reference.
- `returnType` - type of the return value. It is represented either by a string or a reference.

The `UConstructor` class is very similar to the `UMethod` class. The only difference is that it does not include a name.

*6) UAnnotation:* The `UAnnotation` class is an abstract class that represents an annotation of any language element. Classes that extend the `UAnnotation` are:

- `UClassAnnotation` which represents a class annotation,
- `UConstructorAnnotation` which represents a constructor annotation,
- `UFieldAnnotation` which represents a field annotation,
- `UMethodAnnotation` which represents a method annotation.

Selected fields:

- `annotationClass` - type of the annotation. It is represented by a full name of the annotation type. If this annotation type is represented in the unified model, then it is also represented by a reference to that annotation type.
- `annotatedElement` - represents an element which is annotated by this annotation. Type of this element depends on the element that is annotated. For example, in `UClassAnnotation` this element has a type of `UClass`.
- `parameters` - set of parameters of this annotation. Each parameter is represented by the `AnnotationParameter` class which includes three `String` fields: its `name`, `type` and `value`.

### C. Representation of specific elements

The fact that unified model is created from two existing model means that the representation of some language elements can be problematic, especially when it comes to elements that are represented differently in both basic models. One of them is the problem of representing modifiers, which were explained in the introduction of this paper. Other problem we had to face was how to represent references.

*1) Representation of References:* References, that are created during the second phase of unified model creation, are represented by the `AbstractMap.SimpleEntry` class. This class represents a key-value pair, which can for example be used in hash tables such as `HashMap`.

Key of this key-value pair is a full name of the element that is referenced. Key of every reference is filled in during the model creation phase. If possible, the value of this key-value pair is filled in during the reference resolving phase. The value is filled in only when the reference references an element included in the unified model. If that is the case, then the value of the key-value pair will be a reference to the element itself. If the reference references element that is not included in the unified model, then the value will be left as `null`. That means that every reference has a key, which represents the full name of the referenced class, but not every reference has a valid value.

*2) Representation of Modifiers:* As we mentioned before, one of the differences between annotation processing model and reflection model is the representation of modifiers. The difference is that modifiers in annotation processing are represented as a set of enumeration types `Set<Modifier>` and in reflection, they are represented as one integer.

In the unified model, we decided to use the representation identical to the one in reflection, which means all modifiers in unified model are represented by bits of an integer. When the unified model is created from annotation processing model, then the set of modifiers is converted into one integer by using bitwise OR.

### *D. Query Methods*

These methods provide API for convenient metadata queries. Most of these methods are related to annotations, since annotations are the main point of this model.

Query methods included in the class `UModel`:

- `getClasses` - provides all classes that are included in the unified model. Returns a set of `UClass` objects.
- `findClassByFullName` - provides a class specified by its name. Return type is a `UClass` class.
- `findClass` - provides a class specified by a class `Class`. Return type is a `UClass` class.
- `getPackages` - provides all packages included in the unified model. Returns a set of `UPackage` objects.
- `findPackageByName` - provides a package specified by its name. Return type is the `UPackage` class.
- `getEnums` - provides all enumeration types that are declared in the unified model. Return type is a set of `UEnum` classes.
- `getAnnotatedElements` - provides all elements that are annotated by at least one annotation. Return type is a set of classes that implement the interface `AnnotableElement`. This interface is implemented by all elements that can be annotated.
- `getElementsAnnotatedWith` - provides all elements annotated by a specific annotation. Return type is a set of classes that implement the interface `AnnotableElement`.
- `getMethodsAnnotatedWith` - provides all methods annotated by a specific annotation. Return type is a set of `UMetod` classes.
- `getFieldsAnnotatedWith` - provides all fields annotated by a specific annotation. Return type is a set of `UField` classes.
- `getConstructorsAnnotatedWith` - provides all constructors annotated by a specific annotation. Return type is a set of `UConstructor` classes.
- `getClassesAnnotatedWith` - provides all classes annotated by a specific annotation. Return type is a set of `UClass` classes.

Last five methods, which provide elements annotated by a specific annotation, have two alternatives. In one of them, the annotation is specified by a class that extends the `UAnnotation` class. In the other, the annotation is specified by the `Class` object representing the given annotation type. This is a convenience alternative that allows using `Class` objects in order to enjoy type checking.

Beside query methods that are included in the `UModel` class, there are query methods that check whether an element is annotated by an annotation of a specific type. This annotation can either be specified by a class that extends the `UAnnotation` class or by the `Class` class of that annotation.

## VI. Experimental Evaluation of Prototype

Although this paper does not deal with creating a faster way to access metadata when compared to basic models, it is still important that the time needed to access metadata by using unified model is not too demanding in comparison to basic models, otherwise the unified model would be unusable.

For the purposes of determining whether the created experimental tool is fast enough, we have conducted two experiments which are explained in this section.

### *A. Time Needed for Unified Model Creation*

In the first experiment, we tested how much time is needed to create the unified model when using the created tool. This metric is very important, since this time can be considered as **additional time** when compared to basic approach.

For the purposes of this experiment we chose two programs, one of them consisted of 122 classes and the other consisted of 1745 classes. Both of these programs have on average approximately 100 lines of code per one class.

We only tested how much time it would take to create the unified model. That means the difference between time of initialization of the tool and the time that unified model was returned.

*1) Results:* Times that are presented in the table I are average times after 100 runs. It is important to note there were quite big deviations, especially during the compile time.

From the times presented in the table, we can see that the time it takes to create unified model from a small program

TABLE I
TIME NEEDED TO CREATE UNIFIED MODEL

| | 122 classes | 1745 classes |
|---|---|---|
| **Compile time** | 42 ms | 670 ms |
| **Runtime** | 387 ms | 2043 ms |

is not that high, especially during compile time. Time needed to create unified model from a bigger project was naturally higher, because more metadata needed to be processed.

One important thing to note from this experiment is how the tool is scaling when it comes to bigger programs. Second program is approximately 14 times bigger then the first one which means that if the scaling of the tool was linear, it would take 14 times more time to create unified model from bigger program than from smaller program. Creation of unified model from bigger program in compile time took little over 15 times more time than creation of unified model from smaller program. Similarly, creation of unified model from bigger program in runtime took approximately 5 times more time than creation of unified model of smaller program. From these data we can see that scaling of the tool during runtime is much better than scaling during compile time. The fact that the scaling during compile time is worse than linear suggests it may be a good idea to try to optimize implemented tool for the use in compile time.

### *B. Overall Speed of Prototype*

The second experiment was designed to test the speed of the tool when performing a task. For purposes of this experiment, we used the same two programs as we used for the first experiment.

During this experiment we measured four ways in which a given task could be implemented:

- **Annotation processor model** - using metamodel available during compile time.
- **Reflection model** - using metamodel available during runtime. Helper libraries Scannotation [6] and Google Reflection API [21] were used for this measurement.
- **Unified model during compile time** - using the tool to create unified model during compile time and using this unified model to complete the task.
- **Unified model during runtime** - using the tool to create unified model during runtime and using this unified model to complete the task.

The task consisted of two parts (to make the test a little less trivial):

1) Find the names and types of all elements that are marked with the test annotation.
2) Retrieve all fields of a class that is specified by its full name.

For the first part of this experiment a simple test annotation was created and then added to source codes of the programs. In the program that consisted of 122 classes, 8 elements were marked with the test annotation and in the program that consisted of 1745 classes, 137 elements were marked with the test annotation.

TABLE II
OVERALL SPEED OF TASK COMPLETION

| | 122 classes | 1745 classes |
|---|---|---|
| **Annotation processing model** | 12 ms | 484 ms |
| **Reflection model** | 1154 ms | 2822 ms |
| **Unified model during compile time** | 47 ms | 949 ms |
| **Unified model during runtime** | 472 ms | 2245 ms |

*1) Results:* Results of the experimented are presented in table II.

Times that are presented in this table are the average times after 100 runs. Same as in the first experiment, some of the times that were measured when using the unified model were often very different when compared to the average, especially during the tests with the bigger program. During compile time, the lowest measured value was 477 ms and the highest measured value was 1189 ms. During runtime, the lowest measured value was 1823 ms and the highest measured value was 2480 ms. These fluctuations could have been caused by the fact that we used unordered sets in the unified model and the position of requested metadata is always different.

From the data we can see that the time needed to complete the task in unified model during compile time was more then double when compared to the time needed to complete the same task during compile time without the use of the unified model. This fact reinforces the idea of trying to optimize unified model creation during compile time.

We can also see that during runtime, the tool was actually faster in completing the given task. Main reason for this is that helper libraries Scannotation and Google Reflection API were used to make the task significantly easier in reflection model. These libraries made the task easier at the cost of additional time during execution, which is also the main reason behind the unified model.

## VII. RELATED WORK

As it was mentioned before, main inspiration of our solution to the problem we faced with the *BTE* [4] tool was the idea of Generalized and Attributed Abstract Syntax Trees (GAAST) proposed by Cepa [7], [8], [22]. In his work [7] he states that object oriented languages should support an explicit representation of the program as a graph of meta-objects that can be then accessed through a well defined API. He also shows, that usage of meta-model in a GAAST structure helps support model-driven development.

Noguera has a similar approach to Cepa in his tool AVal [23], [24], which is used to validate frameworks that are using annotations. Purpose of this validation is checking, whether the framework uses specific annotations in a correct way. Specification of how annotation is used correctly is designed through meta-annotations, which define constraints on the usage of those annotations. He uses two models for this validation. One of them is *annotation model*, which is a model derived from annotation types and second one is *code model*. Code model is very similar to an abstract syntax tree.

We used a similar approach to the unification of two models which was discussed in this paper in the tool Bridge To Equalia [4]. This tool is able to create a unified model of annotation-based and XML-based configurations using by extensible metamodel.

There are libraries that help querying metadata during runtime, such as Scannotation [6] and Google Reflections [21]. These libraries provide convenient APIs for annotations processing, however, they do not address the problem of runtime and compile time model unification.

## VIII. Conclusion

In this paper we explored differences between compile time and runtime models, explained the problem of their diversity and analyzed the approaches how to unify these models by providing their advantages and disadvantages. The discussion should provide a basis for consideration of providing a standard unified API for accessing metadata in programming languages. We presented the API we designed and implemented in our tool for unified model creation, which was used to support the analysis. The tool was evaluated by experiments to verify the usability of the proposed tool. They and showed that both the creation of unified model and its querying are reasonably fast. The paper can be considered a technical report that should be helpful for developers dealing with the same problem, or developers that are interested in either of the two existing approaches to annotation processing (reflection and annotation processing API).

## References

[1] S. Chodarev, D. Lakatoš, J. Porubän, and J. Kollár, "Abstract syntax driven approach for language composition," *Central European Journal of Computer Science*, vol. 4, no. 3, pp. 107–117, 2014. http://dx.doi.org/10.2478/s13537-014-0211-8

[2] D. Lakatoš, J. Porubän, and M. Bačíková, "Declarative specification of references in DSLs," in *2013 Federated Conference on Computer Science and Information Systems*, ser. FedCSIS 2013, Sept 2013, pp. 1527–1534.

[3] Z. Havlice, "Auto-Reflexive Software Architecture with Layer of Knowledge Based on UML Models," *International Review on Computers & Software*, vol. 8, no. 8, 2013.

[4] M. Nosáľ and J. Porubän, "Supporting multiple configuration sources using abstraction," *Central European Journal of Computer Science*, vol. 2, no. 3, pp. 283–299, Oct. 2012. http://dx.doi.org/10.2478/s13537-012-0015-7

[5] M. Nosáľ and J. Porubän, "XML to Annotations Mapping Definition with Patterns," *Computer Science and Information Systems*, vol. 11, no. 4, pp. 1455–1477, 2014. http://dx.doi.org/10.2298/CSIS130920049N

[6] Scannotation, "Scannotation project homepage," 2015. [Online]. Available: http://scannotation.sourceforge.net/

[7] V. Cepa and M. Mezini, "Language support for model-driven software development," *Science of Computer Programming*, vol. 73, no. 1, pp. 13–25, 2008, special Issue on Foundations and Applications of Model Driven Architecture (MDA). http://dx.doi.org/10.1016/j.scico.2008.05.003

[8] V. Cepa, *Attribute enabled software development: illustrated with mobile software applications*. Saarbrücken, Germany: VDM Verlag, 2007.

[9] V. Cepa and M. Mezini, "Mobcon: A generative middleware framework for java mobile applications," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005. HICSS '05.*, Jan 2005, pp. 283b–283b. http://dx.doi.org/10.1109/HICSS.2005.431

[10] H. Rocha and M. T. Valente, "How Annotations are Used in Java: An Empirical Study," in *SEKE*, 2011, pp. 426–431.

[11] E. Tempero, C. Anslow, J. Dietrich, T. Han, J. Li, M. Lumpe, H. Melton, and J. Noble, "The Qualitas Corpus: A Curated Collection of Java Code for Empirical Studies," in *Proceedings of the 2010 Asia Pacific Software Engineering Conference*, ser. APSEC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 336–345. http://dx.doi.org/10.1109/APSEC.2010.46

[12] Oracle, "javax.lang.model.element documentation," 2015. [Online]. Available: http://docs.oracle.com/javase/7/docs/api/javax/lang/model/element/package-summary.html

[13] ——, "java.lang.reflect documentation," 2015. [Online]. Available: http://docs.oracle.com/javase/7/docs/api/java/lang/reflect/package-summary.html

[14] S. Zawoad, M. Mernik, and R. Hasan, "FAL: A forensics aware language for secure logging," in *2013 Federated Conference on Computer Science and Information Systems*, ser. FedCSIS 2013, Sept 2013, pp. 1579–1586.

[15] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-oriented Software*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1995.

[16] B. Benatallah, M. Dumas, M.-C. Fauvet, F. A. Rabhi, and Q. Z. Sheng, "Overview of Some Patterns for Architecting and Managing Composite Web Services," *SIGecom Exch.*, vol. 3, no. 3, pp. 9–16, Jun. 2002. http://dx.doi.org/10.1145/844339.844346

[17] E. J. Chikofsky and J. H. Cross II, "Reverse Engineering and Design Recovery: A Taxonomy," *IEEE Software*, vol. 7, no. 1, pp. 13–17, Jan. 1990. http://dx.doi.org/10.1109/52.43044

[18] T. Grigalis and A. Čenys, "Using XPaths of inbound links to cluster template-generated web pages," *Computer Science and Information Systems*, vol. 11, no. 1, pp. 111–131, 2014. http://dx.doi.org/10.2298/CSIS130416020G

[19] J. Kollár, I. Halupka, S. Chodarev, and E. Pietriková, "pLERO: Language for grammar refactoring patterns," in *2013 Federated Conference on Computer Science and Information Systems*, ser. FedCSIS 2013, Sept 2013, pp. 1503–1510.

[20] M. Pušnik, M. Heričko, Z. Budimac, and B. Šumak, "XML Schema metrics for quality evaluation," *Computer Science and Information Systems*, vol. 11, no. 4, pp. 1271–1289, 2014. http://dx.doi.org/10.2298/CSIS140815077P

[21] Google, "Google reflection api project homepage," 2015. [Online]. Available: https://github.com/ronmamo/reflections

[22] V. Cepa and M. Mezini, "Declaring and Enforcing Dependencies Between .NET Custom Attributes," in *Generative Programming and Component Engineering*, ser. Lecture Notes in Computer Science, G. Karsai and E. Visser, Eds. Springer Berlin Heidelberg, 2004, vol. 3286, pp. 283–297. http://dx.doi.org/10.1007/978-3-540-30175-2_15

[23] C. Noguera and L. Duchien, "Annotation Framework Validation Using Domain Models," in *Proceedings of the 4th European Conference on Model Driven Architecture: Foundations and Applications*, ser. ECMDA-FA '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 48–62. http://dx.doi.org/10.1007/978-3-540-69100-6_4

[24] C. Noguera and R. Pawlak, "AVal: an extensible attribute-oriented programming validator for Java: Research Articles," *Journal of Software Maintenance and Evolution*, vol. 19, no. 4, pp. 253–275, Jul. 2007. http://dx.doi.org/10.1002/smr.349

DOI: 10.15439/2015F224
ACSIS, Vol. 5

# A Model-driven Approach to Data Structure Conceptualization

Sonja Ristić
University of Novi Sad,
Faculty of Technical Sciences,
Trg D. Obradovića 6,
21000 Novi Sad, Serbia
Email: sdristic@uns.ac.rs

Slavica Kordić, Milan Čeliković
University of Novi Sad,
Faculty of Technical Sciences,
Trg D. Obradovića 6,
21000 Novi Sad, Serbia
Email: {slavica, milancel}@uns.ac.rs}

Vladimir Dimitrieski, Ivan Luković
University of Novi Sad,
Faculty of Technical Sciences,
Trg D. Obradovića 6,
21000 Novi Sad, Serbia
Email: {dimitrieski, ivan@uns.ac.rs}

*Abstract*— **Reengineering of an existing information system can be carried out: to improve its maintainability, to migrate to a new technology, to improve quality or to prepare for functional enhancement. An important phase of a data-oriented software system reengineering is a database reengineering process and, in particular, its subprocess – a database reverse engineering process. The reverse engineering process contains two main phases: data structure extraction and data structure conceptualization. In the paper we present a blueprint of a model-driven approach to database reengineering process that is one of the results of our research project on model-driven intelligent systems for software system development, maintenance and evolution. Within that process hereinafter we focus on the data structure conceptualization process and propose a model-driven approach to data structure conceptualization. Proposed process is based on model-to-model transformations implemented by means of Atlas Transformation Language.**

## I. Introduction

AN information system (IS) implemented to fulfill organizational information requirements would adapt to emerging business models and technology changes and innovations. Organizations are facing the problem of maintenance, evolution or even replacing of a part or whole legacy information/software system. A new system can be redeveloped from scratch, but in that case the knowledge captured in the legacy system is lost. Legacy system replacement or reengineering can be done with significantly reduced amount of effort and cost if the conceptual models are reconstructed from them. Summarizing the definitions of reengineering process presented in [1]–[5] it can be concluded that reengineering entails: (i) the reverse engineering activities aimed at creation of a more abstract view of the system; (ii) the restructuring of the abstract view; and (iii) the implementation of the system in a new form by means of forward engineering activities. A majority of software systems may be characterized as data-oriented systems. They are centered around a persistent data structure, like set of files or a database. The reengineering of a data-oriented software system makes up the code reengineering and the database reengineering processes that could be carried out almost independently one of another. In the paper we deal with a phase of database reengineering process.

The persistent structures, primarily databases, are at the core of most company information systems. The knowledge captured in them can serve as an important resource in a legacy information system modernization project and they are a common source of reverse engineering processes. The database reverse engineering (DBRE) is, according to Hainaut *et al.* [6], the process of recovering the conceptual schema of a database.

The emergence of Model-driven Software Engineering (MDSE) enables the building of more effective reverse engineering solutions. Model-driven Reverse Engineering (MDRE) is the application of MDSE principles, methods and tools to the reverse engineering process, relying on: meta-models, models and model transformations. The Model-driven Architecture (MDA) specified by the Object Management Group (OMG) [7] currently is the most mature formulation of MDSE paradigm.

The OMG has noticed a necessity to understand and evolve existing software assets and launched the Architecture-driven Modernization (ADM) initiative [8]. Within our research project on model-driven intelligent systems for software system development, maintenance and evolution, spanning throughout last past years we have developed a model-driven approach to database reengineering process, adhering ADM framework. In this paper we present just a blueprint of that approach using ADM horseshoe as a reference model. After that we focus on the database reverse engineering (DBRE) process within the database reengineering process.

Hainaut *et al.* in [9] have presented the general architecture of the DBRE process divided in two main phases: data structure extraction and data structure conceptualization. During the data structure extraction phase it is hard to fully apply model-driven principles. The main reason is the fact that some activities need user interaction and the process cannot be fully automated. Therefore, in the

Research presented in this paper was supported by Ministry of Education, Science and Technological Development of Republic of Serbia, Grant III-44010, Title: Intelligent Systems for Software Product Development and Business Support based on Models.

paper we propose a model-driven approach to data structure conceptualization phase of database reverse engineering process. It is implemented within the IIS*Studio development environment aimed at evolutive and incremental IS development and reengineering [10]–[14]. Our approach is mainly based on MDSE and Domain Specific Language (DSL [15], [16]) paradigms. The approach is purely platform independent and strictly differentiates between the specification of a system and its implementation on a particular platform. IIS*Case tool, as a part of IIS*Studio is aimed at IS forward engineering, while IIS*REE tool is aimed at IS reverse engineering. Model transformations implemented in these tools are based on several database meta-models and their classification will be presented at the very beginning of the paper.

Apart from Introduction and Conclusion the paper has 5 sections. In Section 2 the classification of IIS*Studio meta-models is presented. A sketch of a model-driven approach to database reverse engineering is proposed in Section 3. Detailed description of a data conceptualization process is given in Section 4. A case study is presented in Section 5, and related work is treated in Section 6.

## II. IIS*Studio meta-models classification

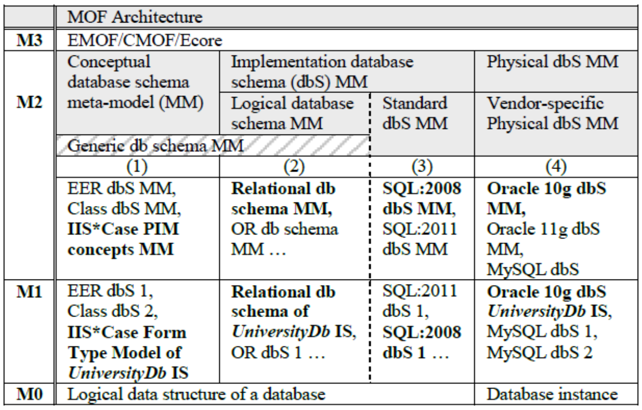
Software development process produces numerous models of complex application artifacts. In the paper we focus on models relating to databases. For these models we use the generic name **database models**. Numerous data models are proposed and they can be classified according to the types of concepts they use to describe the database structure, as follows: i) high-level or **conceptual data models**; ii) representational or **implementation data models**, and iii) low-level or **physical data models**. In the context of MDSE data models may be seen as meta-models. A database schema is expressed by means of the concepts of a selected meta-model. According to the MDSE terminology, such database schema is called database model and it should conform to the appropriate meta-model. An EER database model conforms to the EER meta-model, while a relational database model conforms to the relational meta-model.

Common examples of conceptual data models are (Extended) Entity-Relational (ER, EER) data model and Object-oriented (OO) (also called Class) data model. Relational and Object-relational (OR) data models are used predominantly for logical (implementation) database schema design and database implementation.

IIS*Studio uses form-type (FT) data model for conceptual IS and database design. Like EER and OO data models, it is conceptual data model. A *form type* is the main modeling concept in FT (data) meta-model. It generalizes document types, i.e. screen forms that users utilize to communicate with an information system. Using the form type concept, a designer specifies screen or report forms of transaction programs and, indirectly, specifies (i) an initial set of attributes, constraints and future database schema, (ii) basic functionalities of future transaction programs and (iii) components of their user interface. A form type concept, as well as related concepts of a domain and attribute, is platform independent. A form type is a named tree structure, whose nodes are called component types. Each *component type* is identified by its name in the scope of the form type, and has nonempty sets of attributes and keys, and a set of unique constraints that may be empty. Besides, to each component type must be associated a set of allowed database operations. A part of FT meta-model can be seen in the third row of the table presented in Fig. 4. A detail description of FT data model and FT meta-model may be found in [17].

In this paper, we use the MDSE terminology based on the four-layered architecture proposed by OMG common meta-meta-model Meta-Object Facility (MOF) standard [18]. System under study (SUS) is at the M0 level. The concept of a model is specialized depending on the level, in which a model is located: a model at M1 level, a meta-model at M2 level and a meta-meta-model at M3 level. An SUS is represented by a model at M1 level, which conforms to a meta-model at M2 level that is conformant with a meta-meta-model at M3 level. In [19] we have proposed a classification of database meta-models and distribution of database meta-models and database models across the MOF level stack. Adapted version of the classification is presented in Fig. 1 where can be seen different kinds of database meta-models that describe database models at certain abstraction level:

1. generic database schema meta-models:
   - generic conceptual database schema meta-models,
   - generic logical database schema meta-models;
2. standard physical database schema meta-models; and
3. vendor-specific physical database schema meta-models.



| | MOF Architecture | | | |
|---|---|---|---|---|
| **M3** | EMOF/CMOF/Ecore | | | |
| **M2** | Conceptual database schema meta-model (MM) | Implementation database schema (dbS) MM | | Physical dbS MM |
| | | Logical database schema MM | Standard dbS MM | Vendor-specific Physical dbS MM |
| | Generic db schema MM | | | |
| | (1) | (2) | (3) | (4) |
| | EER dbS MM,<br>Class dbS MM,<br>**IIS*Case PIM concepts MM** | **Relational db schema MM,**<br>OR db schema MM … | **SQL:2008 dbS MM,**<br>SQL:2011 dbS MM | **Oracle 10g dbS MM,**<br>Oracle 11g dbS MM,<br>MySQL dbS |
| **M1** | EER dbS 1,<br>Class dbS 2,<br>**IIS*Case Form Type Model of *UniversityDb* IS** | **Relational db schema of *UniversityDb* IS,**<br>OR dbS 1 … | SQL:2011 dbS 1,<br>**SQL:2008 dbS 1** … | **Oracle 10g dbS *UniversityDb* IS,**<br>MySQL dbS 1,<br>MySQL dbS 2 |
| **M0** | Logical data structure of a database | | | Database instance |

Fig. 1 Classification of database meta-models

In each field of the table in Fig. 1, that is on the intersection of column (**i**), where **i** ∈ {1, 2, 3, 4}, and row M**j**, where **j** ∈ {0, 1, 2, 3}, a list of exemplar models is given. In the column (1) at the M2 level there are several conceptual database schema meta-models: EER, OO (Class) and FT meta-model (so-called IIS*Case PIM concepts MM, where PIM stands for Platform Independent Model). Throughout the paper the names FT MM and IIS*Case PIM concepts MM will be used interchangeably. In the column (1) at the M1 level there are several conceptual database schemas (database models) that are conformant with an appropriate meta-model from the field in column (1) and level M2. For instance: IIS*Case Form Type Model of *UniversityDB* IS is a conceptual database schema of *UniversityDB* information system and it conforms to the IIS*Case PIM concepts MM (FT MM).

It could be noticed that models (meta-models) placed in the same field of the table are at the same abstraction level. In a way, EER database schema meta-model is at the same abstraction level as class database schema meta-model and IIS*Case PIM concepts meta-model. The abstraction level of the models (meta-models) throughout the columns of the same row is decreasing with increasing of the column number **i.** The models (meta-models) placed in column 1 are at the highest abstraction level, while the models (meta-models) placed in column 4 are at the lowest abstraction level. A relational database schema of *UniversityDb* IS is at higher abstraction level than its implementation under the Oracle10g database management system, represented as vendor-specific Oracle 10g database schema of *UniversityDb* IS.

Presented classification is very important for our DBRE approach and for understanding model transformations that are based on classified meta-models. Detailed description of meta-models that we use in this paper may be found in our papers [14], [17] and [19]–[20].

## III. A Model-driven Approach to Database Reengineering Process

Reengineering is one of the key concepts in software maintenance and evolution. It generally includes some form of reverse engineering followed by restructuring (optional) and some form of forward engineering. Previous researches on information system (IS) reengineering have made great achievements concerning: the forward engineering, the identifying of system's components and their interrelationships and the creation of representations of the system in another form. But, they have not yet been successful enough in creating representations of the system at a higher level of abstraction. A reverse engineering process (and consequently a whole reengineering process, too) can benefit of integrating the meta-modeling and meta-models in the process. In this section we present a blueprint of a model-driven approach to database reengineering implemented in IIS*Studio. In Fig. 2 we use ADM horseshoe model to illustrate main steps of proposed approach.

Database reengineering process begins with a legacy database. Relational databases are at the core of most company information systems and that is why, here in Fig. 2, we assume that legacy database is a relational database. Starting from a physical database schema, that is recorded in the relational database schema data repository, the conceptual database schema or logical database schema may be extracted. All of these database schemas represent models at different levels of abstraction.

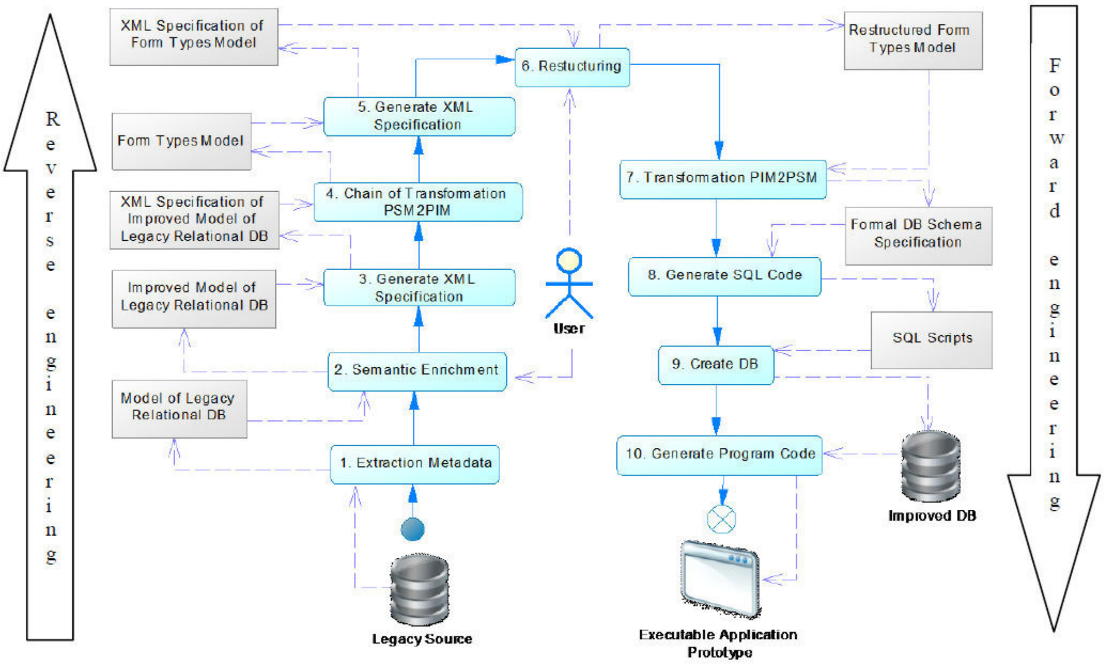

XML Specification of Form Types Model
Restructured Form Types Model
6. Restucturing
5. Generate XML Specification
Form Types Model
7. Transformation PIM2PSM
XML Specification of Improved Model of Legacy Relational DB
4. Chain of Transformation PSM2PIM
Formal DB Schema Specification
Improved Model of Legacy Relational DB
3. Generate XML Specification
8. Generate SQL Code
User
SQL Scripts
2. Semantic Enrichment
9. Create DB
Model of Legacy Relational DB
1. Extraction Metadata
10. Generate Program Code
Improved DB
Legacy Source
Executable Application Prototype
Reverse engineering
Forward engineering

Fig. 2 Horseshoe model of model-driven approach to database reengineering implemented in IIS*Studio

Reverse engineering phase covers steps 1–5, where steps 1–3 belong to the data extraction phase, and steps 4 and 5 make the data conceptualization phase. Step 6 is optional and represents restructuring phase. Forward engineering phase covers steps 7–9. Step 10 is not the step of database reengineering process and it is an activity of code reengineering process.

In the first step information about supported data types, tables, columns, check constraints, primary key and unique key constraints accompanied with foreign key constraints are captured from a legacy database repository and are placed in IIS*REE repository.

Additional information about the inverse referential constraints and homonym inconsistencies are discovered in the step 2 (Semantic Enrichment step). User interaction is required to control data extraction process at this stage. This interaction prevents full automation of data extraction phase and that is the reason why this phase is realized by means of traditional Java programming in OracleJ Developer environment. The similar problem arises again in step 6 (Restructuring step) of the reengineering process because this step requires user interaction, too.

Once it is captured, the physical database schema (model) may be transformed into desired conceptual database model. A designer may choose target conceptual meta-model among EER, Class and FT meta-models. In Fig. 2 FT meta-model is selected. Depending on selected target meta-model a model transformation or a chain of model transformations will be executed. These transformations are based on appropriate meta-models that can be at different levels of abstraction. The distribution of the model transformations and supporting meta-models across MOF and abstraction levels is given and explained in Section IV.

In the purpose of specifying and managing meta-models we have used the Eclipse Modeling Framework (EMF), Eclipse Juno 4.2.1. and OCL 3.2.1. Model transformations used in step 4 are implemented in ATL IDE (ATLAS Transformation Language Integrated Development Environment), version 3.3.1 [21]. Hereinbefore is mentioned that steps 1 and 2 are implemented in one technological space (OracleJ Developer), and step 4 in another one (EMF). Step 3 is used to bridge the technological gap and to export captured data by means of XML document conformant with XML meta-model. Step 6 is implemented in the same technological space as steps 1 and 2, due to the fact that it requires user interaction, too. Step 5, like the step 3, is used to bridge the technological gap, but in the opposite direction. Therefore, strictly speaking, it is not part of the data conceptualization phase.

In a forward engineering process, designers start with a high abstraction level model, abstracting from all kinds of platform issues. Through a chain of model-to-model (M2M) transformations, ending up with a model-to-text (M2T) transformation, the initial platform independent model transforms iteratively to a series of models with less degree of platform independency, introducing more and more platform specific extensions. Details of forward engineering process with activities in steps 7–10 are presented in some of ours already published papers, like [10]–[12], and we omit these details here.

In the next section we focus on the data structure conceptualization phase.

## IV. Data structure conceptualization phase

A complex system may consist of many interrelated models organized through different levels of abstraction. Each of them is conformant to a meta-model. In a forward engineering process a chain of M2M transformations should be completed starting from an initial model at the highest level of abstraction (Platform Independent Model, PIM), through the less abstract models, with different levels of platform specificity (Platform Specific Models, PSMs), and resulting in an executable program code that represents a model at the lowest level of abstraction (fully PSM). These M2M transformations transform a model conformant to a meta-model into another one conformant to a different meta-model. Conversely, in a reverse engineering process, the abstraction level of models and degree of platform independency are increasing throughout the chain of transformations.

In this section we present the activities of data structure conceptualization phase. Model transformations implemented in the conceptualization phase are presented in a rectangular coordinate system in Fig. 3. The abscissa represents MOF level of a model, and the ordinate represents abstraction level of a model in the context of its platform specificity. The input of the process is XML specification of captured physical database model. This XML specification conforms to XML meta-model. The data conceptualization phase is realized as a chain of three M2M transformations: 1. XML2RDBMS, 2. RDBMS2RM, and 3. RM2IISCase. The fourth M2M transformation presented in Fig. 3 is IISCase2XML transformation and it corresponds to the step 5 in Fig. 2, and, as it was explained in Section III, it is not part of data conceptualization phase.

The first transformation transforms a model conformant with XML meta-model into a model conformant with an SQL standard meta-model. It could be SQL: 2008 dbS meta-model, e.g., listed in bold style in column (3) and row M2 in Fig. 1.

The transformation RDBMS2RM transforms a model conformant with an SQL standard meta-model into a model conformant with generic relational db meta-model (bolded text in column (2), row M2 in Fig. 1). It is not possible to transform a model conformant with an SQL standard directly into a model conformant with FT (IIS*Case PIM concepts) MM. The reason lies in the fact that FT approach to database design is based on the Universal relation schema assumption (URS assumption). Physical database meta-models and database meta-models based on SQL standard do not support URS, while generic relational db meta-model does.

Both of aforementioned transformations are PSM2PSM transformations. The third one, RM2IISCase is a PSM2PIM transformation. It transforms a model conformant with generic relational db meta-model into a model conformant with FT meta-model (so-called IIS*Case PIM concepts MM). It is placed in column (1) and row M2 in Fig. 1. In a way the data conceptualization phase is finished and conceptual database model is generated from a legacy database. The obtained model is transformed in an XML document by means of the forth transformation IISCase2XML.

It is important to emphasize that model transformations in Fig. 3 are modeled as conformant with ATL meta-model. Each of them is based on two meta-models: source and target meta-model. The source meta-model of a model transformation is at lower abstraction level than the target meta-model of that transformation in a reverse engineering process. At M1 MOF level each transformation execution transforms an input database schema (model) into an output database schema (model). Although having the same name in Fig. 3, relational database schemas that are input models for transformations 1, 2 and 3, respectively, are not the same models and they are conformant with different meta-models. Their platform independency is ascending throughout the chain of model transformations.

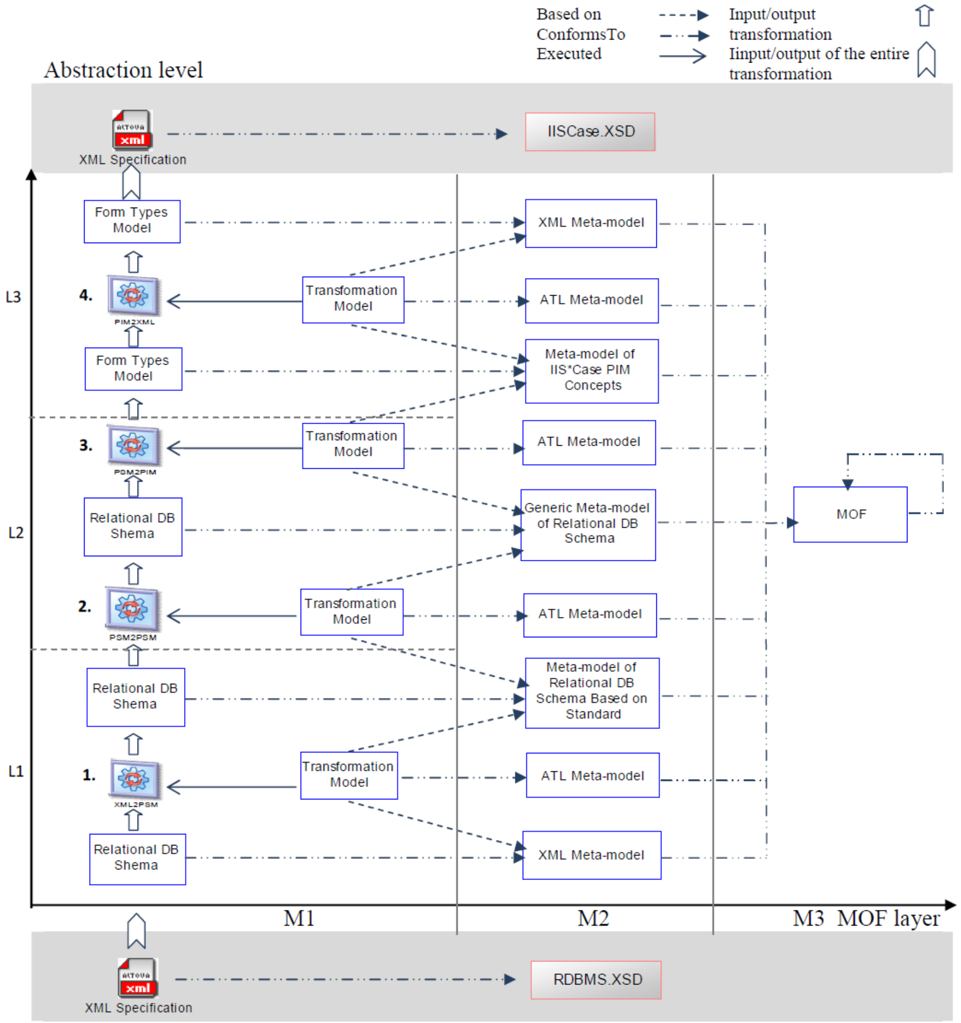
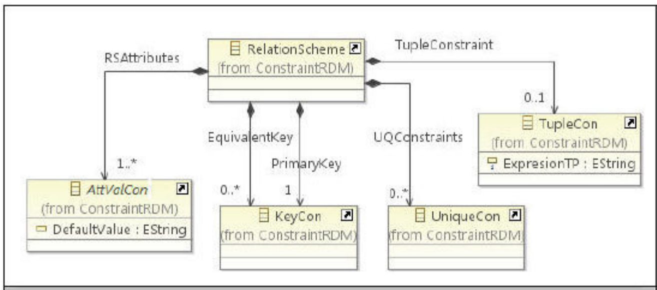


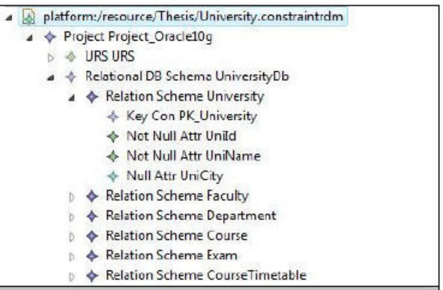
Fig. 3 Data structure conceptualization through a chain of M2M transformations

## V. Case study

In this section an example of a chain of model transformations that implements data conceptualization phase of database reverse engineering process is presented. *UnivesityDb* database instance is SUS consisting of two tables: *University* with columns *UniId*, *UniName* and *UniCity*; and *Faculty* with columns *UniId*, *FacId*, *FacShortName*, *FacName* and *Dean*. It is implemented under the Oracle database management system. In the data extraction phase vendor-specific physical database schema is extracted from legacy database repository and afterwards semantically enriched with the information collected from legacy database and through interaction with designers. The specification of vendor-specific physical database schema is written in an XML document that conforms to XML meta-model. At the very beginning of the process the transformation XML2RDBMS is executed and a SQL standard database schema is generated, conformant, in the particular case, with the SQL: 2008 dbS meta-model. By means of RDBMS2RM transformation the SQL standard database schema is transformed into a generic relational database schema. These two database schemas are not the same and they conform to different meta-models. In Fig. 4 a transformation RM2IISCase, from a generic relational database schema into a form type data model is presented in more details.



A part of generic relational dbS MM

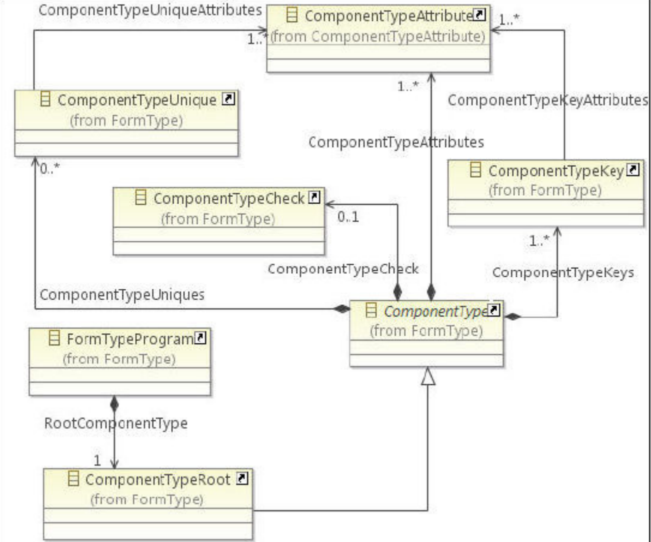


- platform:/resource/Thesis/University.constraintrdm
  - Project Project_Oracle10g
    - URS URS
    - Relational DB Schema UniversityDb
      - Relation Scheme University
        - Key Con PK_University
        - Not Null Attr UniId
        - Not Null Attr UniName
        - Null Attr UniCity
      - Relation Scheme Faculty
      - Relation Scheme Department
      - Relation Scheme Course
      - Relation Scheme Exam
      - Relation Scheme CourseTimetable

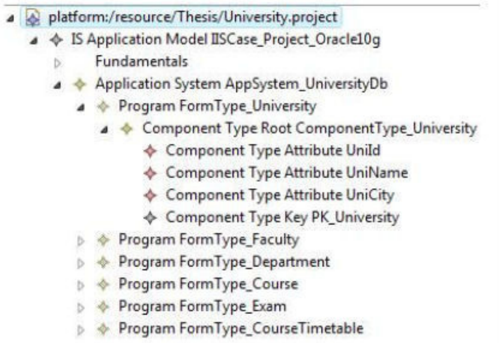
A database model conformant with generic relational dbS MM

```
rule RelationScheme2ComponentType{
from
   rs:RM!RelationScheme
to
   ft: IISCase!FormTypeProgram(
      Name <- 'FormType_' + rs.Name,
      Title <- 'FormType_' + rs.Name,
      ConsideredINDBSchDesign <- true,
      Frequency <- 1,
      ResponseTime <- 1,
      RootComponentType <- ct
   ),
   ctr: IISCase!ComponentTypeRoot(
      Name <- 'ComponentType_' + rs.Name,
      Title <- 'ComponentType_' + rs.Name,
      ComponentTypeAttributes <- rs.RSAttributes->collect(e|
         if (e.oclIsTypeOf(RM!NotNullAttr)) then
            thisModule.RSAttributes2NotNullCompTypeAttribute(e) else
            thisModule.RSAttributes2NullCompTypeAttribute(e) endif),
      ComponentTypeKeys <- (rs.EquivalentKey->collect(e|
         thisModule.EquivalentKey2CompTypeKey(e))).append(
         thisModule.EquivalentKey2CompTypeKey(rs.PrimaryKey)),
      ComponentTypeUniques <- rs.UQConstraints->collect(e|
         thisModule.UniqueCon2ComponentTypeUnique(e,rs)),
      ComponentTypeCheck <-
         thisModule.TupleConstraints2ComponentTypeCheckCon(rs.TupleConstraint),
      Query <- true,
      Delete <- false,
      Insert <- false,
      Update <- false
   )
   do{
      thisModule.attributes <- Sequence{};
   }
}
```

ATL rule for RelationScheme2FormType mapping



A part of IIS*Case PIM (FT) meta-model



- platform:/resource/Thesis/University.project
  - IS Application Model IISCase_Project_Oracle10g
    - Fundamentals
    - Application System AppSystem_UniversityDb
      - Program FormType_University
        - Component Type Root ComponentType_University
          - Component Type Attribute UniId
          - Component Type Attribute UniName
          - Component Type Attribute UniCity
          - Component Type Key PK_University
      - Program FormType_Faculty
      - Program FormType_Department
      - Program FormType_Course
      - Program FormType_Exam
      - Program FormType_CourseTimetable

A database model conformant with IIS*Case PIM (FT) MM

Fig. 4 An example of RM2IISCase transformation

In the first row a part of the generic relational dbS meta-model and a database model conformant with it are presented. In the third row the target IIS*Case PIM concepts meta-model and a conformant form-type database model are presented. In the middle of the table, the second row contains an ATL rule for mapping concepts from the generic relational dbS meta-model to the concepts of IIS*Case PIM concepts meta-model. The model transformation gets the relational database model from the first row as input and transforms it into the form-type database model presented in the third row.

ATL rules presented in Fig. 4 are just the small excerpt from all ATL rules used to specify model transformations in IIS*REE. Presented rules are aimed at mapping concepts of source and target meta-models at the highest level of details. Mappings at the lower level of details alongside with belonging helpers are omitted here.

## VI. Related work

Hainaut *et al.* [6], [9], [22] describe main steps of database reverse engineering and several authors based their research on these proposals. Perez *et al*. [23] and Boronat *et al*. [24] create object-oriented (class) conceptual database schema from the relational data dictionary. Gogolla *et al.* [25] have sketched how syntax and semantics of the ER and relational data model and their transformation can be understood as platform independent and platform specific models. Beggar *et al*. [26] propose a reverse engineering process based on MDSE that presents a solution to provide a normalized relational database which includes the integrity constraints extracted from legacy data. In [27] a process is proposed to automatically generate Web Services from relational databases. SQL-92 meta-model has been used to represent the conceptual database model. Polo *et al.* [28] present the technical and functional descriptions of a tool specifically designed for database reengineering based on simplified relational and object-oriented meta-model.

Aforementioned approaches to database conceptualization are mostly based just on two database meta-models. Vendor-specific physical or standard relational meta-model mainly are found on the source side of M2M transformation. On the other side, EER, class or standard/vendor-specific relational meta-models occur on the target side of M2M transformation, depending on selected conceptual schema. Our approach supports transformations between vendor-specific relational, standard relational, generic relational, FT, EER, and class meta-models. Our classification and distribution of database models across the MOF level stack (Fig. 1) enables systematic approach for mapping specification between different models/meta-models and development of appropriate M2M transformations. To the best of our knowledge, there is no similar systematical overview of database meta-models.

Some authors obtain relational database schema as the final result of data structure conceptualization process. According to Hainaut *et al.* [22] relational database schema cannot be seen as a pure conceptual database schema. It is not fully platform independent due to the fact that it is compliant with the data model of a family of database management systems. In our approach FT database schema is obtained as the result of the data structure conceptualization process. FT specification is based on business forms, users are familiar with, and in that manner it models system as-is in a platform independent way. At the same time, the specification is platform independent prescription model of future screen and report forms and input for series of M2M transformations that ends up with M2T transformation generating application prototype.

Research work on database model transformations can be broadly classified as unidirectional model transformation algorithms and multi-model (multi-language) transformation algorithms. The examples of unidirectional model transformation algorithms, proposed by Lano and Kolahdouz-Rahimi and Beggar *et al.* can be found in [29] and [26], respectively. Multi-model transformation algorithms can be ranged from linking each model to a graph [30], or description logic language (like those presented in [31]–[33]) to transformations mediated by a dictionary of common terms [34]. These results we consider as important in the context of our future research on unifying of different data model meta-models.

## VII. Conclusion

When a legacy system become too costly to maintain, or when new technologies need to be incorporated, system need to be replaced or somehow reengineered. An important phase of a data-oriented software system reengineering is a database reengineering process and, in particular, its subprocess – a database reverse engineering process. In the paper we focus on the data structure conceptualization process and propose a model-driven approach to data structure conceptualization that is based on M2M transformations implemented by means of ATL.

The proposed framework offers a wide range of M2M transformations. IIS*Studio supports transformations between models conformant with EER, FT, relational and class meta-models. In the paper we present database conceptualization into a FT database schema. There are several reasons why we decide to support such kind of conceptualization. Firstly, the end users would participate in the restructuring of a conceptual schema during the database and IS reengineering process. The FT concept is closer to the end-users' perception of data, than the concepts of relational data model, that makes it easier to involve them in the restructuring process. Secondly, extracted FT specifications can be enriched with the specifications of the transaction programs and business applications and they can be used not only to support generation of a database schema of an IS, but to support the generation of an IS program code and appropriate graphical user interface (GUI). In that context, our future research can be directed towards improvement in the data extraction phase. Information captured from application GUI could contribute to the semantic richness of reverse engineered form-type model. Results presented in [35]–[38] support our idea. According to the future research in connection with data conceptualization phase we plan to specify meta-models of different data models and to unify different meta-models as the basis for platform independent modeling of M2M transformations between database meta-models conformant with different data models meta-models, bearing in mind results presented in [39].

## References

[1] E.J. Chikofsky, and J.H. Cross, "Reverse engineering and design recovery: A taxonomy," *IEEE Software*, vol. 7(1), pp. 13–17, Jan. 1990.

[2] R. Kazman, S. Woods and J. Carrière, "Requirements for Integrating Software Architecture and Reengineering Models: Corum ii," in *Proc. of Working Conference on Reverse Engineering* (*WCRE*), Washington, 1998, pp. 154–163.

[3] R. S. Arnold, "A road map guide to software reengineering technology," in *Software Reengineering*, R. S. Arnold, Ed. Los Alamitos CA: IEEE Computer Society Press, 1993.

[4] S. Tilley and D. Smith, "Perspectives on Legacy System Reengineering", SEI White Paper, 1995.

[5] I. Jacobson and F. Lindström, "Re-engineering of Old Systems to an Object-oriented Architecture," in *Proc. of the ACM Conference on Object Oriented Programming Systems Languages and Applications*, New York, Oct. 1991, pp. 340–350.

[6] J. Hainaut, J. Henrard, D. Roland, V. Englebert, and J. Hick, "Structure Eliction in Database Reverse Engineering," in *Proc. of the 3th Working Conference on Reverse Engineering*, IEEE Computer Society Press, Los Alamitos, CA, 1996, pp. 131–139.

[7] J. Mukerji, and J. Miller, "MDA Guide Version 1.0.1, document omg/03-06-01 (MDA Guide V1.0.1)," <http://www.omg.org/>, (retrieved May, 2015)

[8] OMG Architecture Driven Modernization (ADM), <http://www.adm.omg.org >, (retrieved May, 2015)

[9] J. Hainaut, J. Henrard, D. Roland, V. Englebert, and J. Hick, "Knowledge Transfer in Database Reverse Engineering: A Supporting Case Study," in *Proc. of the 4th Working Conference on Reverse Engineering*, IEEE Computer Society Press, Los Alamitos, CA, 1997, pp. 194–203.

[10] I. Luković, P. Mogin, J. Pavićević, and S. Ristić, "An approach to developing complex database schemas using form types," *Software: Practice and Experience*, vol. 37 (15), pp. 1621–1656, 2007. doi: 10.1002/spe.820

[11] S. Aleksić, I. Luković, P. Mogin, and M. Govedarica, "A generator of SQL schema specifications," *Computer Science and Information Systems*, vol. 4(2), pp. 81–100, 2007.

[12] I. Luković, A. Popović, J. Mostić, and S. Ristić, "A tool for modeling form type check constraints and complex functionalities of business applications," *Computer Science and Information Systems*, vol: 7(2), pp. 359–385, 2010. DOI 10.2298/CSIS1002359L

[13] S. Aleksić, S. Ristić, I. Luković, and M. Čeliković, "A Design Specification and a Server Implementation of the Inverse Referential Integrity Constraints," *Computer Science and Information Systems*, vol. 10(1), pp. 283–320, 2013. DOI: 10.2298/CSIS111102003A

[14] S. Ristić, S. Aleksić, M. Čeliković, I. Luković, "Generic and Standard Database Constraint Meta-Models," *Computer Science and Information Systems*, vol. 11(2), pp: 679–696, 2014. DOI: 10.2298/CSIS140216037R

[15] M. Mernik, J. Heering, and A. M. Sloane, "When and how to develop domain-specific languages," *ACM computing surveys (CSUR)*, vol. 37(4), pp. 316–344, 2005.

[16] T. Kosar, N. Oliveira, M. Mernik, V. J. M. Pereira, M. Črepinšek, C. D. Da, and R. P. Henriques, "Comparing general-purpose and domain-specific languages: An empirical study," *Computer Science and Information Systems,* vol. 7 (2), pp. 247–264, 2010. DOI: 10.2298/CSIS1002247K

[17] M. Čeliković, I. Luković, S. Aleksić, V. Ivančević, "A MOF based meta-model and a concrete DSL syntax of IIS*case PIM concepts," *Computer Science and Information Systems*, vol. 9 (3) pp. 1075–1103, 2012. DOI: 10.2298/CSIS120203034C

[18] OMG Meta Object Facility (MOF), <http://www.omg.org/mof>, (retrieved May, 2015)

[19] S. Ristić, S. Aleksić, M. Čeliković, V. Dimitrieski, and I. Luković, "Database reverse engineering based on meta-models," *Central European Journal on Computer Science*, vol. 4(3), pp: 150–159, 2014. DOI: 10.2478/s13537-014-0218-1

[20] V. Dimitrieski, M. Čeliković, S. Aleksić, S. Ristić, I. Luković, "Extended entity-relationship approach in a multi-paradigm information system modeling tool, " in *Proceedings of the 2014 FEDCSIS*, Warsaw, Poland, ACSIS, Vol. 2, pp. 1611–1620, 2014. DOI: 10.15439/2014F239

[21] Eclipse Foundation, "ATL (ATLAS Transformation Language) Project, ATL/User Guide", available at: http://wiki.eclipse.org/ATL/User_Guide, 2010. (retrieved May, 2015)

[22] J-L. Hainaut, J. Henrard, V. Englebert, D. Roland, and J-M. Hick "Database Reverse Engineering," In *Encyclopedia of Database Systems*, L. Liu, and Özsu, T., Ed., Springer-Verlag, 2009.

[23] J. Perez, I. Ramos, and V. Anaya, "Data reverse engineering of legacy databases to object oriented conceptual schemas," *Electronic Notes in Theoretical Computer Science*, vol. 74(4), pp. 1–13, 2002.

[24] A. Boronat, J. Perez, J. A. Cars, and I. Ramos., "Two Experiences in Software Dynamics," *Journal of Universal Computer Science*, vol. 10(4), pp. 428–453, 2004.

[25] M. Gogolla, A. Lindow, M. Richters, and P. Ziemann, "Meta-model transformation of data models," Position paper. WISME at the UML 2002.

[26] Beggar O. E., Bousetta B., Gadi T., Getting Relational Database from Legacy Data-MDRE Approach, Computer Engineering and Intelligent Systems www.iiste.org ISSN 2222-1719 ISSN 2222-2863 (Online) Vol 4, No.4, 2013.

[27] R.P. del Castillo, I. García-Rodríguez, and I. Caballero, "PRECISO: a reengineering process and a tool for database modernization through web services", In: *Jacobson Jr.*, M.J., Rijmen, V., Safavi-Naini, R. Ed., SAC 2009. LNCS, vol. 5867, Springer, Heidelberg, pp. 2126–2133, 2009.

[28] M. Polo, I. Garcia-Rodriguez, and M. Piattini, "An MDA-based approach for database re-engineering," *J. Softw. Maint. Evol*., vol. 19 (6), pp. 383–417, 2007.

[29] K. Lano, S. and Kolahdouz-Rahimi, "Constraint-based specification of model transformations," *Journal of Systems and Software*, vol. 86(2), pp. 412–436, 2013. DOI: 10.1016/j.jss.2012.09.006

[30] M. Boyd, and P. McBrien, "Comparing and transforming between data models via an intermediate hypergraph data model," *Journal on Data Semantics*, vol. IV, pp. 69–109, 2005.

[31] R. P. Fillottrani, and C. M. Keet, "Conceptual Model Interoperability: A Metamodel-driven Approach," *Rules on the Web. From Theory to Applications, Lecture Notes in Computer Science*, pp 52–66, 2014. DOI: 10.1007/978-3-319-09870-8_4

[32] C. M. Keet,and P. R. Fillottrani, "Structural entities of an ontology-driven unifying metamodel for UML, EER, and ORM2," In: *Proc. of MEDI'13. LNCS*, vol. 8216, Amantea, Calabria, Italy, Springer, pp. 188–199, 2013. DOI: 10.1007/978-3-642-41366-7_16

[33] C. M. Keet, and R. P. Fillottrani, "Toward an ontology-driven unifying metamodel for UML Class Diagrams, EER, and ORM2," *Conceptual Modeling , Lecture Notes in Computer Science,* vol. 8217, pp 313–326, 2013. DOI: 10.1007/978-3-642-41924-9_26

[34] P. Atzeni, G. Gianforme, and P. Cappellari, "Data model descriptions and translation signatures in a multi-model framework," *AMAI Mathematics and Artficial Intelligence*, vol. 63, pp. 1–29, 2011. DOI: 10.1007/s10472-012-9277-y

[35] Kreutzová, Michaela, Jaroslav Porubän, and Peter Vaclavik. "First Step for GUI Domain Analysis: Formalization," *Journal of Computer Science and Control Systems*, vol. 4(1), pp. 65–69, 2011.

[36] M. Bačiková, and J. Porubän, "DSL-driven generation of GUI" *Central European Journal of Computer Science*, vol. 4(4), pp. 204–211, 2014. DOI 10.2478/s13537-014-0210-9

[37] M. Malki, A. Flory, and M. K. Rahmouni, "Extraction of Object-oriented Schemas from Existing Relational Databases: a Form-driven Approach," *INFORMATICA*, vol. 13(1), pp. 47–72, 2002.

[38] S. M. Benslimane, M. Malki, M. K. Rahmouni, and DJ. Beslimane, "Extracting Personalised Ontology from Data-Intensive Web Application: an HTML Forms-Based Reverse Engineering Approach," *INFORMATICA*, Vol. 18 (4), pp. 511–534, 2007.

[39] H. Kern, "Study of Interoperability between Meta-Modeling Tools" in *Proceedings of the 2014 FEDCSIS*, Warsaw, Poland, ACSIS, Vol. 2, pp. 1629–1637, 2014. DOI: 10.15439/2014F255

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 985–990

DOI: 10.15439/2015F210
ACSIS, Vol. 5

# Towards Machine Mind Evolution

Ján Kollár, Michal Sičák and Milan Spišiak
Department of Computers and Informatics,
Technical University of Košice
Letná 9, 042 00, Košice, Slovakia
Email: {jan.kollar, michal.sicak, milan.spisiak}@tuke.sk

*Abstract*—**We introduce the principles of symbolization and conceptualization of perceived reality. We show them as processes that are applicable inside of a computer mind. We derive those processes from principles of a human mind. We present process of higher order regular abstraction and state automata acceptance as possible machine reality realization mechanism. We present the algorithm for evolution of machine mind language. Verification of this algorithm is presented in functional language Haskell. On the output, we have obtained fine-grained parallel non-redundant structure. It is in super-combinator form and represents elements of a machine mind language. Count of elements applications highly exceeds the count of elements themselves. Instead of accumulating acquired information, they are dissolved in the language of a mind. Vice versa, the information can be reconstructed from this language. We show that non-redundancy of a machine language is the decisive criterion for restructuring the machine mind language in each moment of communication. It is like that, so we can achieve more powerful and abstract communication between humans and computers or between computers themselves.**

## I. Introduction

ABSTRACTION of the human-computer communication interfaces [1] can be increased by DSL development [2], coming out from the abstract syntax [3]. Grammatical and/or genetical language evolution [4], [5], [6] is more automated approach, but still oriented more to syntactic than to semantic facet of languages.

On the other hand, the development of thinking machine is a great challenge, which can be found in many works of historians, philosophers, psychologists and linguists. For example, Renfrew [7] characterizes humans as symbolic beings that language was evolved in a consequence of communication. Gardenfors [8] refers to the structured substance of concepts. Structured conceptualization for databases can be found in [9]. Chomsky's [10] hypothesis on the universal language of thought, innate to every human being is the subject of many discussions.

In this paper, we present the algorithm in which reality recognized by humans can be conceptualized and represented in the machine mind.

Our approach comes out from the empirical cognition of reality by humans and its reflection in the mind: our imaginations are structural, and our concepts that reflect the meaning of the reality are languages, since they are meaningful. In our opinion, no information acquired by human is being forgotten (it is just suppressed in our mind) and no information is stored multiplicity.

We deal with symbolization just marginally in this paper, more information can be found in [11]. We discuss conceptualization as a process of abstracting the language concepts in more detail, since differently abstracted concepts on input of the algorithm affect the evolved language of the machine mind.

We present the algorithm for the machine mind language evolution in eight subsequent steps. In this paper, we also introduce the principle of applicative deterministic automata, able to recognize reality incrementally.

Also another aspect of symbolized input processing is explored. We present possibility of higher order regular expressions creation.

Our results tend to distributed automata structures [12] and they conform with the principle of the determinism of systems [13]. The proposed solution allows us to separate process forming of language of the machine mind from the process of machine thinking reasoning on concepts. As we hope, this provides a perspective for cognitive and more abstract HC and CC communication.

## II. Symbols and perception

Human is a symbolic creature, i.e. the human communication is symbol based. He creates immaterial concepts inside of his mind that flow right from the symbols that have material substance. Concepts exist only inside of a human mind, i.e. his imagination. Symbols posses their informal meaning and at the same time they might be structured with high degree of complexity during the communication.

Symbolization is a symbol based reality representation process. Simple components of reality, such as the letter $A$ or an image of a square can be symbolized with simple symbols like $a$ or $b$. Therefore we are able to symbolize perceived reality, like the sequence (1), into the symbol sequence $aabababa$.

$$AA\Box A\Box A\Box A\Box A \tag{1}$$

The new structured symbol has been created. The relation between the reality and its symbolized form is bidirectional. It is not enough to symbolize reality at the input in order to communicate, as the reality needs to be reconstructible at the output from the previously obtained symbol structure.

This work was supported by project VEGA 1/0341/13 Principles and methods of automated abstraction of computer languages and software development based on the semantic enrichment caused by communication

978-83-60810-66-8/$25.00©2015, IEEE

The symbolization of the sequence (1) is expressed with the rule (2):

$$\frac{A \leftrightarrow a \qquad \Box \leftrightarrow b}{AA\Box A\Box A\Box A\Box A \leftrightarrow aabababa} \tag{2}$$

Section above line in the rule (2) tells us, that if we have already identified reality of simple facts $A$ and $\Box$ and symbolized them with symbols $a$ and $b$, then new structured reality (1) is going to be symbolized with the use of a new symbol. This is the structured symbol $aabababa$. The structure doesn't necessarily have to be a sequence as it is in our example, but may be any structure. Symbols are just abstractions of reality in our view, recorded and produced from inside of a memory during communication. In order for communication to be faultless, such records of reality are stored inside the memory. They represent reality inside a human or computer mind. Since a human does not think in symbols, we may need to imagine its relation to machine mind abstract symbols as a connective bidirectional links to records of reality. Since we do not posses an architecture of a machine that corresponds to human mind, we find no other way than to consider them as alphabet symbols - terminal symbols of a language, and to represent them in discrete way, like whole non-negative numbers.

The purpose of symbolization is not the same as recognition problem. A machine is able to register and reproduce reality. Cognition is tied with the language evolution and it is inherent property of a mind. The language evolution works together with symbol abstraction. It is based on conceptualization, whose substance is the concept creation inside of an internal machine language.

## III. Conceptualization

A mechanism of conceptualization is a higher form of language abstraction than symbolization is.

String (3) is a concept - language, identical with the symbol sequence perceived on input. An abstraction was not performed upon it.

$$aabababa \tag{3}$$

Sequence (1) is a concept or a language that has meaning if we consider process of symbolization defined in (2) being applied.

Consider a concept in a form of a regular expression (4), where $(r)^*$ is transitive closure of an expression $r$. This concept is created by performed abstraction that searches for repeating patterns, in our case the sequence $ab$.

$$a(ab)^*a \tag{4}$$

This concept is more abstract than concept (3) for there is a relation (5) between the languages (3) and (4).

$$aabababa \subset a(ab)^*a \tag{5}$$

A machine operating with the language (3) cannot effectively communicate with a machine that recognizes the language (4), since its language is less abstract. In automaton theory it means that an automaton derived from the regular expression (3) accepts only terminal symbol string $aabababa$. It can be generated from the regular expression (4) however. On the other hand, the human or the machine (3) can learn from its counterpart (4), since it is capable of producing the symbol set $aa$, $aaba$, $aababa$, etc. It can achieve level (4) by this process. We can say, that language (3) will mutate to the language (4). The premise for machine learning is a massive stream of structured symbols at the input. Language (3) mutation into more abstract form is based on such a stream.

A sequence of symbols at the input shown at (6) leads to conceptualization - the process of concepts creation shown in (7).

$$aa; aaba; aabababa \tag{6}$$

$$\begin{aligned} &1.\quad (\{\}, aa) \Rightarrow \{aa\} \\ &2.\quad (\{aa\}, aaba) \Rightarrow \{a(ab)^{\{0,1\}}a\} \\ &3.\quad (\{a(ab)^{\{0,1\}}a\}, aabababa) \Rightarrow \{a(ab)^{\{0,1,3\}}a\} \end{aligned} \tag{7}$$

In the first step, the mind is enriched by language $aa$ as there is the symbol $aa$ at the input and the mind is empty. This language is stored inside of the machine mind. Next, the symbol $aaba$ arrives at the input. The result of conceptualization is mutation of the language $aa$ inside the mind of a machine into the language $a(ab)^{\{0,1\}}a$ during the second step. Another symbol transforms the language into $a(ab)^{\{0,1,3\}}a$ during the third step. This language enables the machine to communicate in form of the symbols $aa$, $aaba$ and $aabababa$, since (8) holds.

$$\begin{aligned} &a(ab)^0a \Rightarrow aa \\ &a(ab)^1a \Rightarrow aaba \\ &a(ab)^3a \Rightarrow aabababa \end{aligned} \tag{8}$$

The conceptualization example (7) shows us, that symbols on the right side of (8) can be generated and then represented as records of reality from concepts on the left side. Those concepts are languages, since they are sub-languages of most general language (9). This language is a concept only when all symbols generated from it have meaning i.e. represent the reality.

$$a(ab)^{\{0,1\}}a \subset a(ab)^*a \tag{9}$$

Following facts about conceptualization are to be noted:

1) Conceptualization leads to language ambiguity, it expands exponentially, therefore a selection condition is necessary. For example, language $(a)^{\{2\}}$ could have been created in the first step of (7) instead of the language $aa$.
2) It is possible to represent single symbol with multiple languages. For example, if language $a(ab)^{\{0,1,3\}}a$ is created inside of a mind of one machine and language $aa(ba)^{\{0,1,3\}}$ inside of another, they will be equal in terms of inter-machine communication. Machines will "understand" each other.

We can conclude that the machine mind is non-redundant i.e. does not contain any two identical language components. The mind hasn't got an ability to forget and it is deterministic. Language structure of the mind has to be highly parallel, associative, and amount of interconnections between language components highly exceeds the amount of components themselves.

## IV. Structured symbols as regular expressions

We have shown, that reality can be symbolized as a structured symbol describable by a regular grammar. In this paper, we use regular expressions to describe concepts that were obtained by previously symbolized reality. Those expressions themselves provide interesting ways to describe concepts, as we will show in this section.

Regular expressions are usually an in-line way to describe regular grammars. In practice, we use them for string matching. Therefore, fundamental data type (the type of symbol that regular grammar itself is matching to patterns) for accepting is a character. Our expressions are composed of operators sequence, closure and sum. First two have been already described earlier. Sum operator is $n$-ary and designated as $|$ symbol.

Regular expressions represented as a string itself are not well executable. One of better executable solutions is to use a tree form that can be considered meta-executable, as described in [4]. All our experiments were performed in functional language Haskell. It allows easy creation of complex data structures, such as a tree. Definition is based on work [11] and is depicted below:

```
data REXp a = TERM a | SEQ [REXp a] |
    SUM [REXp a] | CLS (REXp a)
```

Data structure `REXp` contains four constructors. Constructor `TERM` represents fundamental type, `SEQ` is for the sequence, `SUM` represents the sum operation and `CLS` is for the closure. It is notable, that we are not using type `Char` as the fundamental type for regular expression. Current definition allows us to match expressions composed of an arbitrary data type. We will return to this proposition later on. Sum and sequence operators are $n$-ary.

Consider following regular expression:

$$a(b|c)^* \tag{10}$$

where $a, b, c \in \Sigma$, where $\Sigma$ is terminal symbol set. This expression matches string that begins with symbol $a$ and after that an arbitrary long string consisting of $b$s and $c$s is matched. Transformed regular expression into the tree form is shown bellow. Transformation is based on scheme described in [11], so we won't present it here for brevity.

```
SEQ [TERM 'a',
      CLS (SUM [TERM 'b',TERM 'c'])]
```

As mentioned earlier, fundamental type of regular expression does not have to be of type `Char`. Definition allows us to use arbitrary data type.

TABLE I
Meaning of regular expression elements

| symbol | meaning |
|---|---|
| $a$ | $d(e)^*$ |
| $b$ | $ef$ |
| $c$ | $(d\|f)$ |

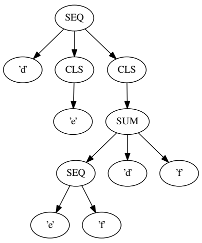


Fig. 1. Less abstract regular expression

Let us consider our regular expression data type as a form of abstract interpretation of another regular expression. Therefore elements $a$, $b$ and $c$ represent other regular expressions, and the resulting type will be `REXp (REXp Char)`. This can be considered a higher order regular expression.

Term "higher order" is used similarly as in functions, where higher order functions take a function as a parameter or return one. Since our regular expression takes another as a parameter, it can be considered higher order regular expression.

Definitions of symbols the $a$, $b$ and $c$ meaning is depicted in Table I. We use different symbols, $d$, $e$ and $f$, in this abstraction to make it clear on what level of we currently are. By substitution of the elements $a$, $b$ and $c$ we get less abstract regular expression:

$$d(e)^*(ef|(d|f))^* \tag{11}$$

Tree form of expression (11) is shown in Fig 1.

Fundamental data type can be arbitrary, even recursive. Question arises, what if we would be able to create data type that is going to be recursive and will contain `REXp a` type. Data type like that would need a terminating and recursive part. In our case, we will use type shown bellow. Regular expression itself would be of type `REXp RexChar`.

```
data RexChar =
      Ch Char | R (REXp RexChar)
```

Every element of regular expression can be either a character (the use of `Char` is not obligatory, we have chosen this type for example purposes) or regular expression of same type as the original expression. Purpose of recursive regular expressions is that they can contain themselves as their fundamental elements entering form of recursion.

In Haskell we can create those expressions statically as a source code or we might obtain them dynamically during execution by parsing input. Here is an example of a higher order regular expression in static Haskell code:

```
a,b :: REXp RexChar
```

```
a = SEQ [TERM (Ch 'a'),
      SUM [TERM (R b), TERM (Ch 'c')]]
b = SUM [SEQ [TERM (Ch 'b'),
      TERM (R b)], TERM (CH 'b')]
```

We can rewrite them as:

$$\begin{aligned} A &\to a(B \mid c) \\ B &\to bB \mid b \end{aligned} \tag{12}$$

If we consider the rules of CFG from Chomsky hierarchy, we can see that our recursive regular expressions (12) are written in format of a BNF rule. We can see, that our recursive regular expressions are corresponding with CFG rules and BNF grammars.

By definition of hierarchy of languages, for every regular language a finite state automaton can be constructed. Automata are efficient way for accepting or even generating phrases of regular languages. Next section discuses the idea of higher order state automata, that may not necessarily be finite, yet for all practical purposes, they will be usable.

## V. State automata derived from higher order expressions

As machine mind processes symbolized input and transforms it into concepts, it needs mechanism for structured symbols acceptance. This can be done with state automata. Deterministic automata can only have one transition for each symbol from one state to another and cannot contain empty transitions. Every nondeterministic automaton can be converted into deterministic. For practical purposes, it is always better to have control over choosing which step to take next.

To obtain state automata we need to have an expression first. We will use following example:

$$A \to aAb|ab \tag{13}$$

Expression (13) can be interpreted as a grammar of context free language that contains strings that have exact number of $a$s and $b$s in succession. This is typical illustration of a non-regular context free language, uninterpretable as a regular expression. However with our recursive expression it is possible. Resulting tree created from (13) will be infinite, depicted on Fig 2. We can see that second argument of sum operator, $ab$, is terminator part. We can conclude that if we consider our regular trees as executable, they are able to accept a context free language.

In work [11] is presented an algorithm that transforms regular expressions directly into minimum state automaton. Same can be applied to higher order regular expressions. Resulting automaton of expression (13) is depicted in Fig 3. This is an automaton corresponding to regular expression labeled as $A$.

There is a transition within body of an automaton in Fig 3 that accepts not a symbol but entire automaton. In this case it accepts itself. We can perceive this fact as a recursive jump to the higher level. For an input string "$aaabbb$", our automaton would have jumped twice into higher levels. We could see this as an automaton with infinite levels, as depicted in Fig 4. We see, that those automata are equivalent to recursive transition networks.

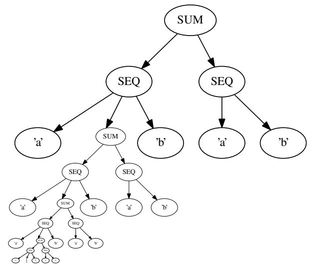


Fig. 2. Tree form of expression (13).

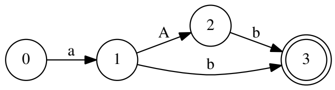


Fig. 3. Context free automaton

Final state on higher level does not always mean possible string acceptance, it can just represent return from a recursion, as depicted in Fig 4, where states 3 of all higher levels have transition into state 2 of lower level. Acceptance of string is achieved only in final states on base level. This does not mean, that higher level final states are not allowed to have transitions into final states of lower levels. In case like that, whole string of symbols can be accepted at higher level and then sequence of empty transitions into base final state will occur.

Consider following higher order expression written according to our rules, labeled as $B$:

$$B \to ba(Bb|(b)^1) \tag{14}$$

Closure in last symbol is parametrized and it designates zero or one repetition. Automaton for expression (14) is depicted in Fig 5. In this particular case see an example of higher order nondeterminism.

Automaton itself is deterministic, if we look at it as residing on one level, where $B$ transition is just another symbol transition. It is unfortunately not our case and $B$ means

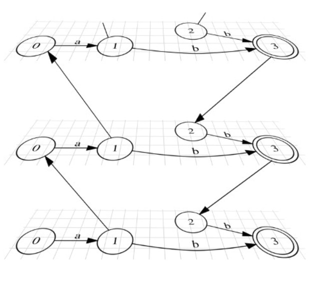


Fig. 4. Recursive infinite state automaton

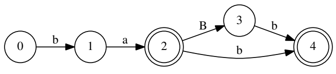


Fig. 5. Nondeterministic higher order state automaton

recursion, where automaton accepts itself. Consider execution where we are on non-zero level and are in the state 2. It is the final state, so execution on current level can terminate and we can return back. There is also a transition into the higher level and a normal transition with the symbol $b$.

Therefore exactly three transitions are possible at this moment. And all three are accepting one symbol, $b$. Jumping into the higher level causes jump into state 0 and here only $b$ is accepted. Returning one level down will get us into the state 3 and here $b$ is only acceptable symbol as well.

Application of automata inside machine mind is a possible solution for perceived reality acceptance, and as we have shown, symbolized structures do not need necessary be simply structured. The accepting part is but a small portion of a whole concept of the machine mind. The more important part is to deal with symbols and conceptualize them into the memory. We will continue to use only simple regular expressions further on for the sake of simplicity. Our focus is to get closer to human comprehension of concepts as possible. In the next section we will present the algorithm of abstract machine mind evolution that satisfies conditions of symbolization and conceptualization laid in previous sections.

## VI. Abstract language evolution algorithm

Let the language $a(ab)^*a$ be a result of conceptualization, that is also a concept, since symbolization process was bound to reality. This concept is composed of infinite amount of sub-concepts, where each of them has its own meaning. The most simple concept is $aa$ that symbolizes meaning of two uppercase letters $AA$. We can abstract from language semantics, since it is defined by separate process of symbolization. Therefore we can see them as simple regular expressions rather than complex languages.

An algorithm for abstract language evolution is composed of those eight steps:

1) Transformation of regular expression to metaexecutable form of (meta)syntactic tree, considering priority and associativity of metaoperations.
2) Selection of arguments  terminal symbols and transformation of each (sub)expression to the form of its application to selected arguments.
3) Abstraction of each expression  deriving the arity of lambda abstraction from number of arguments, selected in step 2. Elementary languages arise being still applied in the tree.
4) Decomposition of elementary languages (elementary language concepts). Elementary languages are dissolved in associative memory represented by a list and applied by the MATCH operation.
5) Abstraction of all applications (APP) to terminal symbols, i.e. replacing them by abstract applications (APPA) to lambda variables. The set of mutually applied abstract languages is formed.
6) Compression - removal of multiple occurrences of the same abstract languages, i.e. removal of redundancy of elementary languages.
7) Filtration - removal of sets of pointers to redundant languages. Manipulation phase.
8) The machine mind language extraction.

The sequential separation of steps is useful for the methodological purposes and also for verification of evolution on a sequential computer. But it can be noticed that sequential steps are executable potentially in parallel as pipelined processes.

## VII. Algorithm implementation using Haskell

In the Haskell implementation, meta-operations of sequence and transitive closure used in abstract concept $a(ab)^*a$ as well as sum (|) meta-operation were used. This allows us to apply the algorithm to the abstract language concept $a(b)^*|c$ as an input sample.

Each elementary language is in supercombinator form, i.e. lambda abstraction (LAM $n$ $e$), where n represents lambda variables $x_1, \ldots, x_n$ and $e$ is a meta-expression. Meta-expression $e$ in the form VAR 1 represents lambda variable $x_1$, otherwise it is the application of binary sequence metaoperation (SEQ), unary transitive closure metaoperation (CLS), and n-nary sum metaoperation (SUM), all being applied to the applications of elementary languages. These applications (APPA (MATCH k) $[x_1, \ldots, x_m]$) apply $k$-th elementary language of arity $m$ to bound lambda variables. By other words, k-th elementary language is a supercombinator, which can be accessed to be applied in a highly distributed associative memory of mind by operation (MATCH $k$) .

## VIII. Algorithm verification

We obtain the abstract language concept $a(b)^*|c$ represented in the machine mind by five elementary languages $L$ in supercombinator form (15).

$$\begin{aligned} L &= \{L_0, L_1, L_2, L_3, L_4\} \\ L_0 &= \lambda x_1.x_1 \\ L_1 &= \lambda x_1.L_0\ x_1 \\ L_2 &= \lambda x_1.(L_0\ x_1)^* \\ L_3 &= \lambda x_1.\lambda x_2.L_1\ x_1 + L_2\ x_2 \\ L_4 &= \lambda x_1.\lambda x_2.\lambda x_3.(L_3\ x_1\ x_2|L_0\ x_3) \end{aligned} \tag{15}$$

Each of elementary languages is applied to admissible abstract symbols, that sets have been derived in the algorithmic evolution. All possible applications (16) represent all meaningful sub-concepts acquired by the most complex input concept.

$$\begin{aligned} &L_0\ a,\ L_0\ b,\ L_0\ c \\ &L_1\ a \\ &L_2\ b \\ &L_3\ a\ b \\ &L_4\ a\ b\ c \end{aligned} \tag{16}$$

For example, we can verify that the applications of identity supercombinator $L_0$ yields concepts $a$, $b$, and $c$, that represent facts, and the application of supercombinator $L_4$, see (17), yields the most complex concept $a(b)^*|c$.

$$L_4\ a\ b\ c = a(b)^*|c \tag{17}$$

Abstraction in the process of conceptualization significantly decreases the amount of elementary supercombinators of the machine mind. For example, abstract concept $a(ab)^*a$, i.e. $a(ab)^ka$ would be represented by 6 supercombinators, for each $k \geq 0$. But if the algorithm is applied to non-abstracted concept (such as $aabababa$), the amount of supercombinators grows linearly ($2k+3$ for $k \geq 0$, i.e. 9 supercombinators for $aabababa$).

Summarization of structurally identical concepts yields no change of supercombinator form of languages $L$, just that amount of their arguments is increased. For example, (18) holds.

$$L(a(ab)^*a|c(cd)^*c) = L(a(ab)^*a|c(cd)^*c|e(ef)^*e) \quad (18)$$

Considering well abstracted conceptualization of large amount of symbols on input, we can await that the amount of application bindings exceeds the amount of elementary languages rapidly, preserving non-redundancy of the set of elementary languages.

## IX. Recognition by applicative automata

As can be seen, derived elementary languages are defined by regular expressions in that a single meta-operation is applied to elementary languages applications, instead of terminal symbols. Let us remind that:

$$L_4\ a\ b\ c = a(ab)^*c$$

In the first step of meta-computation we get:

$$L_4\ a\ b\ c \Rightarrow L_3\ a\ b \,|\, L_0\ c = A \,|\, B$$

Then it is not necessary to recognize concrete sub-concept belonging to $a(ab)^*a$ by generation of DFA for $a(ab)^*a$, but it is sufficient to derive the applicative DFA for regular expression $A|B$, where terminal symbol $A$ represents the set of final states of applicative DFA for ($L_3\ a\ b$) and terminal symbol $B$ represents the set of final states of applicative DFA for application ($L_0\ c$).

Considering (15), the number of applications commonly highly exceeds the number of elementary languages. That is why we can formulate the hypothesis on saturation of the machine mind, which can be realized as slowing-down expansion of set the $L$ with each new concept on input. Multiple occurrences of terminal symbols in arguments in (16) are of less importance, since they physically do not exist, because they represent just interconnections to reality records.

## X. Conclusion

The main advantage of internal machines supercombinator form is addition of automatic conceptualization criterion. It is the minimal count of elementary languages, which themselves are supercombinators as well. Internal language of a mind does not expand with ongoing stream of symbols on input. It is common knowledge that almost everything is new for a little child but an adult hardly finds something new that he hasn't heard or seen. In order for learnable machine to communicate with human, it needs to work on similar thought principles. This thinking is language oriented. If this principle is present, it can then absorb reality, symbolize it and based on abstraction it can create concepts in form of internal languages of a mind.

Our position isn't that supercombinator form is the only form able to represent internal language of mind. It is but an algorithmic realization of a possible way not to store anything multiplicity inside mind but to remember everything. Our supercombinator form has future applications possibilities despite the fact that each of cognition steps could look differently, if someone else performed it. Experiment with higher order expressions shows us that our future work may be involved in other forms of grammars.

We do not think the machine mind evolution is based just on regular expressions, as presented in this paper. But using this simplest case we were able to introduce the principle of applicative automata for output communication. Clearly, machines should be active in communication, i.e. able to formulate the questions.

Derived language of the machine mind consists of elementary languages, such that each of them can be fetched associatively by matching. An interesting structural similarity between the machine mind and neural brain network can be noticed. Both mind meta-operations and brain neurons are the nodes and both mind applications and brain synapses are arcs of a graph. However, in our opinion, no bidirectional relation between brain and mind can be deduced from mentioned structural similarity.

## References

[1] M. Bačíková, J. Porubän, and D. Lakatoš, "Defining domain language of graphical user interfaces." in *SLATE*, 2013, pp. 187–202.

[2] J. Poruban, M. Bacikova, S. Chodarev, and M. Nosal, "Pragmatic model-driven software development from the viewpoint of a programmer: Teaching experience," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on.* IEEE, 2014, pp. 1647–1656.

[3] J. Porubän, M. Forgáč, M. Sabo, and M. Běhálek, "Annotation based parser generator," *Computer Science and Information Systems*, vol. 7, no. 2, pp. 291–307, 2010.

[4] J. Kollár and E. Pietriková, "Genetic evolution of programs," *Central European Journal of Computer Science*, vol. 4, no. 3, pp. 160–170, 2014.

[5] F. Javed, M. Mernik, B. R. Bryant, and A. Sprague, "An unsupervised incremental learning algorithm for domain-specific language development," *Applied Artificial Intelligence*, vol. 22, no. 7-8, pp. 707–729, 2008.

[6] M. O'neill, C. Ryan, M. Keijzer, and M. Cattolico, "Crossover in grammatical evolution," *Genetic programming and evolvable machines*, vol. 4, no. 1, pp. 67–93, 2003.

[7] C. Renfrew, *Prehistory: the making of the human mind.* Random House Digital, Inc., 2009, vol. 30.

[8] P. Gärdenfors, "Symbolic, conceptual and subconceptual representations," in *Human and Machine Perception*. Springer, 1997, pp. 255–270.

[9] M. Pater and D. E. Popescu, "Multi-level database mining using afopt data structure and adaptive support constrains." *International Journal of Computers, Communications & Control*, vol. 3, no. 3, 2008.

[10] N. Chomsky, *Syntactic structures.* Walter de Gruyter, 2002.

[11] J. Kollár, "Formal processing of informal meaning by abstract interpretation," *Smart Digital Futures 2014*, vol. 262, p. 122, 2014.

[12] R. Smith, C. Estan, S. Jha, and S. Kong, "Deflating the big bang: fast and scalable deep packet inspection with extended finite automata," in *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4. ACM, 2008, pp. 207–218.

[13] J. L. Peterson, *Petri net theory and the modeling of systems.* Prentice-hall Englewood Cliffs (NJ), 1981, vol. 132.

DOI: 10.15439/2015F247
ACSIS, Vol. 5

# Generating Database Access Code From Domain Models

Nassima Yamouni Khelifi*†, Michał Śmiałek*, Rachida Mekki†
*Warsaw University of Technology, Poland
Email:{nassima, smialek}@iem.pw.edu.pl
†University of Sciences and Technologies of Oran-Mohamed Boudiaf-, Algeria
Email:{nassima.yamounikhelifi, rachida.mekki}@univ-usto.dz

***Abstract*—Automatic processing of requirements (e.g. to generate code) remains a challenge in contemporary software development. Requirements are still treated as secondary artifacts by software developers, as they are written in natural languages which causes ambiguity. In this paper, we present an approach to generate working code from requirements through applying precisely formulated domain models. As the source, we use the Requirements Specification Language (RSL) which is a precise constrained language, based on a central domain model composed of domain notions. These notions are linked from use case scenarios and create a form of a 'wiki'. Notions are graphically visualized in RSL, and resemble UML classes with attributes. Notions can be used in phrases that can represent various operations used within use case scenarios. In our approach we introduce model transformation algorithms that allow to generate database access code associated with operations to persist (store, retrieve) data in a database system. To focus our work, we present code generated for Hibernate which is an object relational mapping framework.**

***Index Terms*—model-driven requirements engineering, model transformations, database access, metamodelling**

## I. Introduction and Background

TYPICAL requirements specifications in contemporary software projects use natural language, possibly with some elements of modelling. This poses a significant challenge for approaches to automate the process of turning requirements into working code. A prominent field of research that aims at changing this situation is Model-Driven Requirements Engineering (MDRE) [1]. In MDRE, requirements are expressed as models, often by using the Unified Modelling Language (UML) [2]. Such models are intended to be comprehensible to both software developers and end-users. This comprehension is most often assured by introducing a comprehensive vocabulary of the problem domain in the form of a domain model. This allows to use the techniques of domain engineering [3]. Most often, UML class diagrams are used. Classes represent domain objects (noun phrases) with associated atomic attributes (also nouns) and operations (verb phrases). What is more, class models can define relationships between domain objects (notions), thus allowing to build a certain semantic network of related notions.

Still, UML does not offer any means to associate domain models with the remaining models and it has no precise syntax for textual elements like scenarios. Thus, in this paper we will use a language dedicated to formulating precise requirements models, called the the Requirements Specification Language (RSL) [4], [5]. The syntax of RSL is defined through a meta-model using the Meta-Object Facility (MOF) meta-language [6]. An important feature of RSL is that it introduces precise (hyper-)linking of domain models within textually expressed requirements units.

To link textual requirements with domain models, we need to represent requirements by using notions from the domain model in a consistent way [7]. This means – for instance – that use case [8] scenarios should be composed of links to domain model elements – noun phrases and verb phrases, contained in a central domain model. In RSL, this is done very consistently: all the scenario sentences are in fact links to phrases contained in the domain model. This makes the whole RSL-based specification resemble a 'wiki' system (see Fig. 8 for an illustration) with consistent use of hyperlinks to specific vocabulary notions.

This consistency of RSL allowed Śmiałek et al. [9], [10], [11] to formulate formal translational rules for generating code from requirements models (use cases and their scenarios) down to UML design models and Java code. The resulting code follows the Model-View-Presenter (MVP) [12] architectural pattern. The rules focus on generating full code for the View and the Presenter layers, and method stubs for the Model layer. They also permit to generate Data Transfer Objects, that facilitate control flow between the three layers. Unlike for other approaches in MDRE, these rules allow for a fully automatic translation from high-level requirements models (use cases, scenarios, domain vocabularies) down to fully operational code.

The above approach with RSL still lacks rules for generating code associated with persistence operations at the Model layer. Thus, in the current work, we concentrate on defining and implementing rules for generating database access code that is responsible for processing and persisting data. We want this code to be consistent with that for the View and Presenter layers as introduced in the previous paragraph. For this, we will use the Hibernate framework [13] which is an Object Relational Mapping (ORM) that allows for mapping Java Classes (DTOs) to database tables, and for managing data with the Data Access Object (DAO) design pattern [14]. We give rules to generate general Create/Read/Update/Delete (CRUD) operations for persisting Java objects in database tables.

In the following sections we introduce our approach which

978-83-60810-66-8/$25.00©2015, IEEE

is consistent with typical model transformation approaches of Model-Driven Software Development [15], [16]. In Section II we briefly present the domain vocabulary part of RSL which is the source language for our transformations. In Section III we present the transformation itself: selected rules and algorithms that implement them. The rules define the translation from RSL to UML with inserted operational Java code. The algorithms are expressed in a graphical transformation language called MOLA (MOdel Transformation LAnguage) [17], [18]. The last section presents conclusions stemming from implementing the presented algorithms within the RSL environment called ReDSeeDS [19] and using a UML tool (Enterprise Architect - EA, from Sparx Systems, sparxsystems.eu) to visualise the generated UML models and to generate the final Java code.

## II. Requirements Specification Language: Overview of the Domain Notions Part

Requirements Specification Language (RSL) is a semi-formal language for specifying precise requirements [5], [20]. The fundamental idea behind RSL is separation of concerns: separating description of the system's application behavior and the description of the system's problem domain. The behavior of the system is described with ***use cases*** and their ***textual scenarios*** written in constrained natural language. The domain specification is defined using ***notions*** (words). Phrases in scenario sentences constitute the **application logic**, and are linked to notions that constitute the **domain logic**. Notions can be composed of both "nouns" and "verbs".

The RSL's grammar is defined formally through a metamodel written in MOF which is standardized by the Object Management Group (OMG, www.omg.org) [6]. The primary goal of MOF is to allow metamodels to be defined using basic class model syntax (classes with attributes and relationships). The full description of RSL consists of the abstract syntax (i.e. the metamodel), the concrete syntax (definitions of visual language elements), and informally specified semantics (natural language descriptions similar to those in the UML specification [21]). It can be found in a comprehensive report from the ReDSeeDS project [4] which uses the 'Complete' (CMOF) dialect of MOF. A variant that uses the 'Essential' dialect of MOF (EMOF) is presented in the book by Śmiałek and Nowakowski [11]. The book also presents a comprehensive approach to defining RSL's semantics in a formal, translational way.

Figure 1 presents a small excerpt from the RSL metamodel pertaining to notions and their relationships. As we can notice, in contrast to UML, **Notions** contain *notionAttribute*s that are also **Notions**. Attributes can be distinguished from regular notions by their possession of an **AttributeDataType**. The possible data types include:

- "text": string-like textual description;
- "number": an integer number;
- "floating number": a number with a possible decimal;
- "truefalse": a boolean value;
- "date": a value containing date or time;

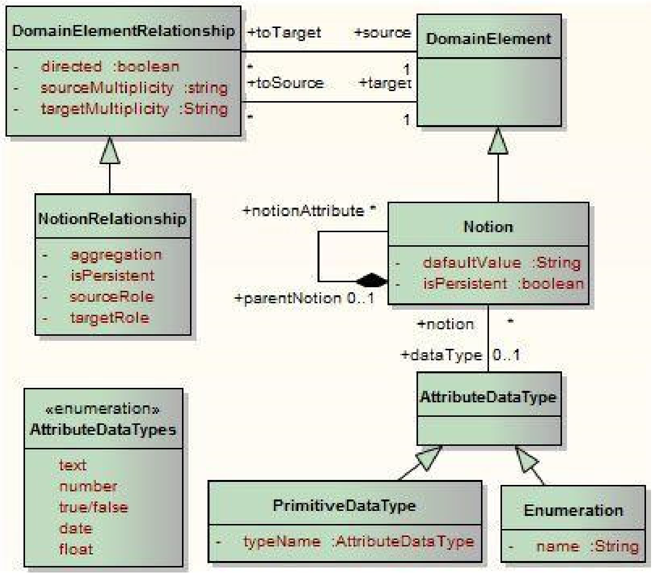


Fig. 1. Part of the RSL's metamodel for notions and their relationships

It can be noted that to represent requirements for software we need to define domain elements in two areas: 1) the problem (business) domain, and 2) the application domain. The problem domain is the actual reality that the software supports. It is stable and quite independent from the application to be built and changes when the reality changes. The application domain changes when the application changes and contains various parameters and user interface elements.

Here we will concentrate on the the problem domain that consists of domain notions (e.g. "book", "publisher", "author"...etc), and their attributes (e.g. "title", "name", "address" ..etc.). The concrete notation for domain notions in RSL is similar to that found in UML class models. An example is shown in Figure 2. The basic type of **Notion** in RSL is the *"Concept"*. Graphically, it resembles a UML class. The second type of **Notion** is the *"Attribute"*. We can notice that unlike in UML, attributes are not contained graphically in the *"Concepts"*. This is a feature of RSL that facilitates sharing attributes between various notions, especially of the 'view' type (see below).

*Concepts* and *attributes* are presented as rectangles adorned with appropriate tags. *Attribute* elements additionally contain information about the data type included in brackets. We should note that the data types are not limited and can be easily extended in the metamodel, depending on the problem domain (e.g. with sound, graphics, etc.), but should be defined in advance prior to developing a transformation.

Other notion types in RSL are called "**Data views**" and are divided into two kinds: "Simple data view", and "List data view". Data views point to sets of attributes. "Simple data views" serve to present single instances of combined attributes. "List data views", are used to present lists, containing many instances. "Data views" and other RSL elements are not treated in detail in this paper, for more details please refer to the book by Śmiałek an Nowakowski [11].

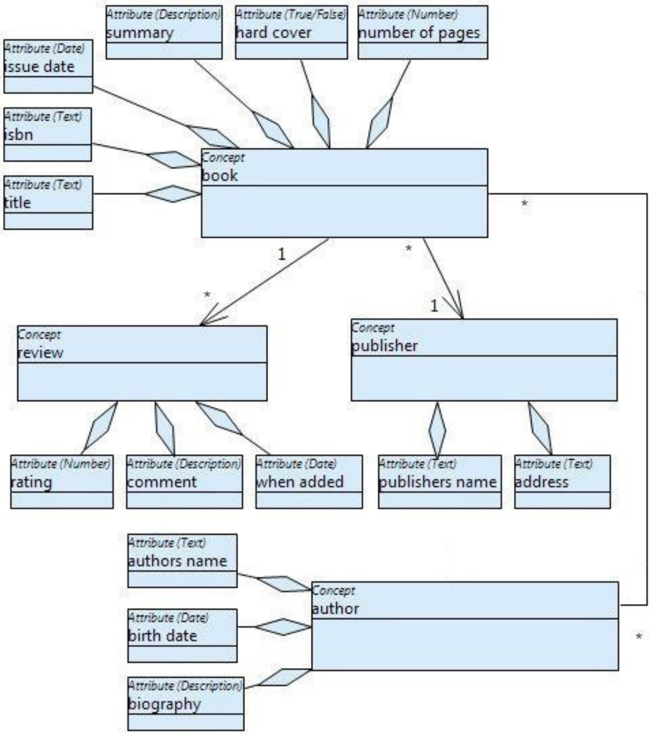


Fig. 2. Example of notions and their relationships in RSL's concrete notation

The different types of domain elements are connected by "relationships". The first kind of relationship is denoted similarly to associations in UML. It relates two concepts, and can have multiplicities. The second type of relationship is containment of attributes within concepts, where the diamond is placed on the concept side. The notation of containment relationship is also taken from UML and resembles aggregations.

In RSL, we can define all kinds of problem domains (e.g. Physics, Aeronautic, Finance, etc.). Figure 2 illustrates an example of RSL model for the "Library Management" problem domain. In this domain we have four elementary **concepts**: book, author, publisher and reviewers. Each concept can hold a number of attributes (shown via the containment relationship). For example, the "book" concept has attributes like: ISBN, title, number of pages, issue date, etc. The four concepts are connected via associations, that have appropriate multiplicities (one-to-many, many-to-one, many-to-many, ... etc.).

## III. Generating Database Access Code From RSL

This section consists of two parts. In the first part we introduce selected transformation rules that constitute translational semantics for RSL's domain vocabulary constructs, where the target languages are UML and Java. In the second part we present algorithms expressed in MOLA that implement these transformation rules.

### A. *Transformation Rules*

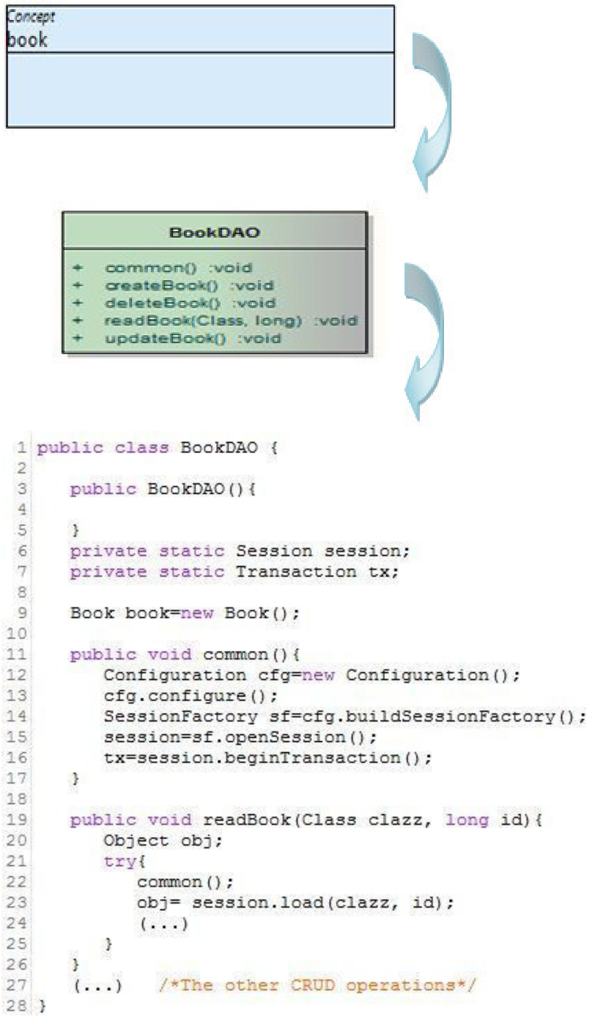
In order to generate database access code from RSL we need to explain its semantics concerning this aspect. For this, we will use a pragmatic translational approach [22], which is based on translating a "Source language" to a "Target language" which has already well-defined semantics [23], [24]. In our case, the source language is "RSL", and the target languages are "UML" and "Java". The reason behind choosing Java and UML as target languages, is that they are widely used and understood by a large community of software developers.

We will define a set of translation rules to generate data base access code from RSL domain models. For this, we will use the Data Access Object (DAO) design pattern [14], that implements the access mechanism required to work with the data source (e.g. Relational Database Management Systems like: MySQL, Oracle, PostgreSQL, etc.). We will apply this pattern in the context of the Hibernate framework, which is an open source Object-Relational-Mapping (ORM) that allows for persisting and storing data in a database [13], via CRUD (Create/Read/Update/Delete) operations within the DAO classes.

Hibernate maps Java persistent classes to database tables, and from Java data types to SQL data types, and provides data query and retrieval facilities. The Java persistent classes are Data Transfer Objects (DTOs) [25] or POJOs (Plain Old Java Objects). Hibernate uses XML files for mapping Java classes into tables, which are: Hibernate Configuration file, and Hibernate Mapping files. The Hibernate Configuration file contains all the required information related to the database, and other related parameters. The Hibernate Mapping files should be generated for each DTO class, and should contain information related to associations between database tables.

In the following, we will present two translation rules that defines semantics of RSL domain models in terms of DAO classes, and Hibernate classes. Other rules, that allow to generate Data Transfer objects and configuration files are out of scope of this work.

- **Rule R1:** Every "Concept" in the RSL domain model is translated into a DAO class. The name of the class is derived from the Concept's name and concatened with the "DAO" string. Each class contains four operations: "Create", "Read", "Update", and "Delete", plus the "common" operation. The operations' names are derived from the name of one the CRUD operations, concatenated with the name of the "Concept". Figure 3 provides a description of Rule R1. In this example we can see that the "book" concept in RSL, is first translated into a UML class named "BookDAO" with four CRUD operations (createBook, readBook, updateBook, and deleteBook), plus the "common" operation. The UML class is then translated into a "BookDAO" Java class. The "common" operation (lines 11-17) initialises various variables required by Hibernate. In lines 19-26, we show a fragment of code for the "readBook" operation, which takes two parameters as input: the Class (i.e. the DTO class) and the identifier (ID), and reads the proper object using the "load" method of the 'session' object.



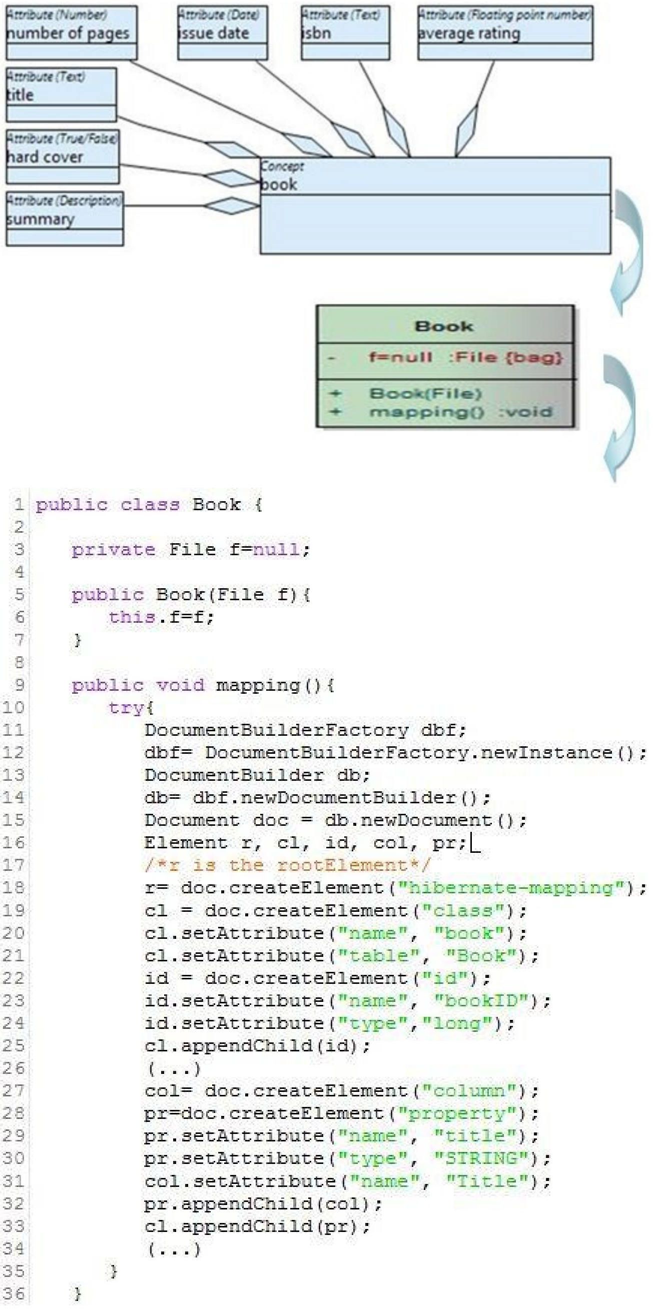
```
public class BookDAO {

   public BookDAO(){

   }
   private static Session session;
   private static Transaction tx;

   Book book=new Book();

   public void common(){
      Configuration cfg=new Configuration();
      cfg.configure();
      SessionFactory sf=cfg.buildSessionFactory();
      session=sf.openSession();
      tx=session.beginTransaction();
   }

   public void readBook(Class clazz, long id){
      Object obj;
      try{
         common();
         obj= session.load(clazz, id);
         (...)
      }
   }
   (...)   /*The other CRUD operations*/
}
```

Fig. 3. Generation of the “BookDAO” class with CRUD operations

- **Rule R2:** Every “Concept” with its attributes in the RSL domain model is translated into a mapping class. The class’ name is derived from the concept’s name. Each mapping class contains a constructor with a parameter, and the “mapping” operation. Rule R2 is illustrated in Figure 4. The example shows a fragment of the code for the “mapping” operation. This operation is responsible for mapping of the Book class into the book table which exists already in a database, this table of course has a unique identifier (ID). Each attribute is mapped into a column in a database table (see lines 27-31).

### B. Transformation Algorithm

The presented rules define the expected outcome of a transformation from RSL’s domain models to database access code. To implement these rules, we have developed appropriate transformation algorithms. Here we introduce them by using MOLA, a language dedicated to model transformations, developed at the University of Latvia [17], [18]. The MOLA notation is graphical, as illustrated in Figures 5-7 and is based on graph-grammar rules that are defined in the context of UML activity diagrams.



```
 1 public class Book {
 2
 3    private File f=null;
 4
 5    public Book(File f){
 6       this.f=f;
 7    }
 8
 9    public void mapping(){
10       try{
11          DocumentBuilderFactory dbf;
12          dbf= DocumentBuilderFactory.newInstance();
13          DocumentBuilder db;
14          db= dbf.newDocumentBuilder();
15          Document doc = db.newDocument();
16          Element r, cl, id, col, pr;
17          /*r is the rootElement*/
18          r= doc.createElement("hibernate-mapping");
19          cl = doc.createElement("class");
20          cl.setAttribute("name", "book");
21          cl.setAttribute("table", "Book");
22          id = doc.createElement("id");
23          id.setAttribute("name", "bookID");
24          id.setAttribute("type","long");
25          cl.appendChild(id);
26          (...)
27          col= doc.createElement("column");
28          pr=doc.createElement("property");
29          pr.setAttribute("name", "title");
30          pr.setAttribute("type", "STRING");
31          col.setAttribute("name", "Title");
32          pr.appendChild(col);
33          cl.appendChild(pr);
34          (...)
35       }
36    }
```

Fig. 4. Generation of Hibernate mapping code for the ”Book” concept

MOLA is a procedural language, and Figure 5 shows a sequence of 6 procedure calls. This forms the main procedure of our algorithm. After cleaning-up the target model, the procedure creates a general package structure. Then, it creates appropriate Data Transfer Objects and Data Access Objects. Finally, it generates Hibernate mappings and configuration files.

In this current short introduction to the transformation algorithm we will concentrate on creating the DAOs (cf. Rule 1 in the previous section). The appropriate procedure for

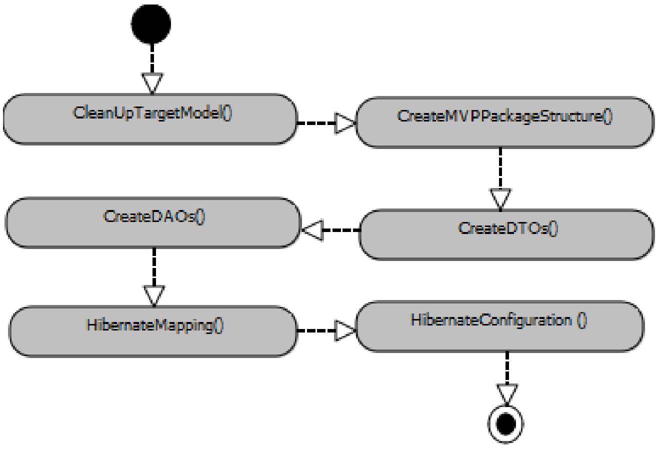


Fig. 5. Main steps of the algorithm expressed in MOLA

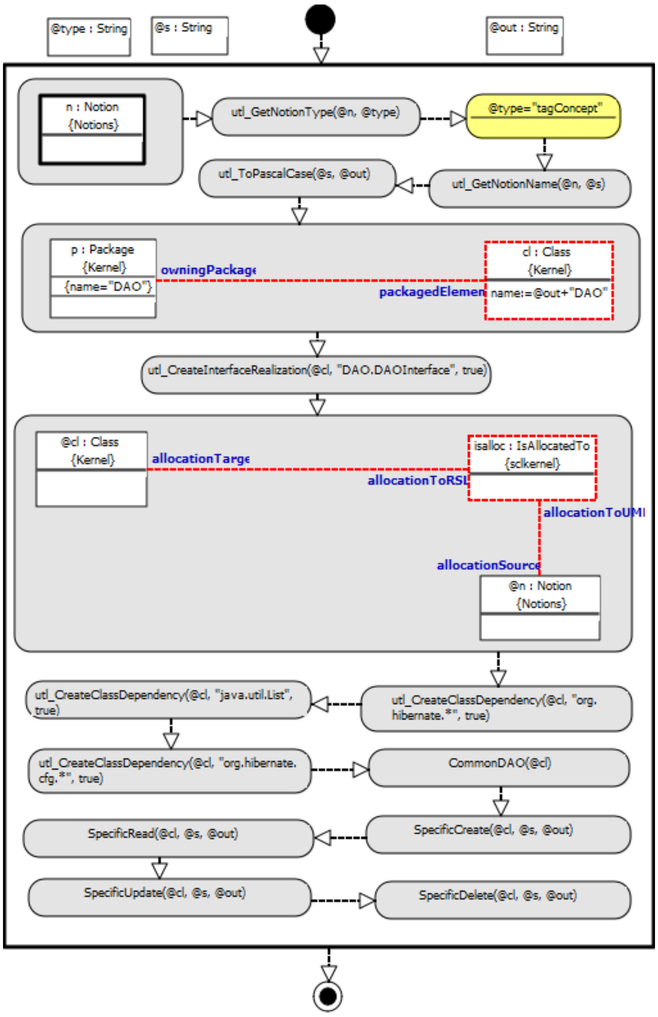

@type : String
@s : String
@out : String
n : Notion
{Notions}
utl_GetNotionType(@n, @type)
@type="tagConcept"
utl_ToPascalCase(@s, @out)
utl_GetNotionName(@n, @s)
p : Package
{Kernel}
{name="DAO"}
owningPackage
packagedElemen
cl : Class
{Kernel}
name:=@out+"DAO"
utl_CreateInterfaceRealization(@cl, "DAO.DAOInterface", true)
@cl : Class
{Kernel}
allocationTarge
allocationToRSL
isalloc : IsAllocatedTo
{sclkernel}
allocationToUMI
allocationSource
@n : Notion
{Notions}
utl_CreateClassDependency(@cl, "java.util.List", true)
utl_CreateClassDependency(@cl, "org.hibernate.*", true)
utl_CreateClassDependency(@cl, "org.hibernate.cfg.*", true)
CommonDAO(@cl)
SpecificRead(@cl, @s, @out)
SpecificCreate(@cl, @s, @out)
SpecificUpdate(@cl, @s, @out)
SpecificDelete(@cl, @s, @out)

Fig. 6. Procedure for creating DAOs (“CreateDAO”) expressed in MOLA

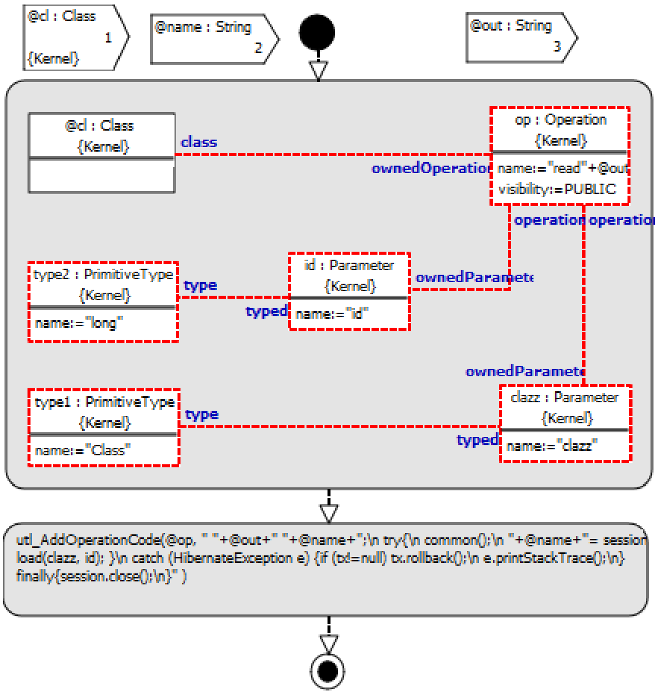
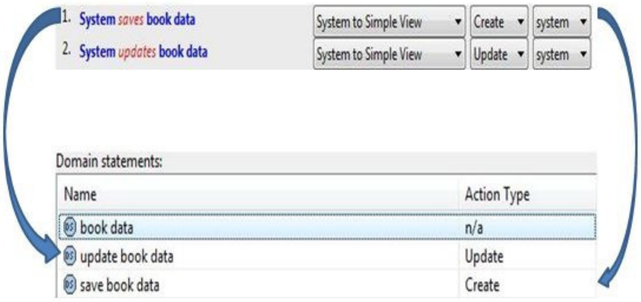


Fig. 7. MOLA rule for “SpecificRead” operation

implementing this is presented in Figure 6. The procedure iterates over all the **Notion** objects (refer to Figure 1) found in the source model (see the object ‘n:Notion’ in the top-left of the figure). Note that in MOLA, loops are represented by rectangles with thick borders which encompass all the actions to be performed within them. For each notion, the loop determines the notion’s type, and if it is of type ‘concept’ it creates a new **Class** type object (see ‘cl: Class’). This object is placed inside an existing **Package** (see ‘p:Package’) named ‘DAO’. Additionally, the newly created class is related through a realisation relationship to a DAO interface object and through a mapping relationship (see ‘isalloc:IsAllocatedTo’) – to the original notion object. It is also completed by generating some relations to other elements that allow for importing certain elements specific to the Hibernate framework.

The main loop concludes by generating four CRUD operations within the new class. One of these procedures is illustrated in Figure 7. This procedure takes two parameters as input – the class and its name. In the class it creates an operation (see ‘op:Operation’ with two parameters (see ‘id:Parameter’ and ‘clazz:Parameter’). As we can see, the main MOLA rule presented in Figure 7 shows a configuration of 5 objects to be created (one operation, two parameters and two primitive types). This configuration is consistent with the UML’s metamodel which can be found in its official specification [21].

Note that the generated operation matches code illustrated in line 19 in Figure 3. In addition, by using a simple text processing statement (see ‘utl_AddOperationCode’), the procedure appends the method of the operation with the code like in lines 20-24 in the same figure.



1. System saves book data | System to Simple View ▾ | Create ▾ | system ▾
2. System updates book data | System to Simple View ▾ | Update ▾ | system ▾

Domain statements:

| Name | Action Type |
|---|---|
| book data | n/a |
| update book data | Update |
| save book data | Create |

Fig. 8. Linking domain elements from scenarios

## IV. Conclusion and Future Work

To validate the presented approach we have implemented the above introduced algorithm within the framework of the the ReDSeeDS (Requirements-Driven Software Development System) tool suite [19] (see: www.redseeds.eu). The tool offers a full Model-Driven Software Development (MDSD) life cycle: i.e. from requirements to UML models down to Java code. In our application, the source RSL domain model with its notions were specified using the ReDseeDS editor (see Fig. 2). After writing and testing the transformation rules using the MOLA environment (see: mola.mii.lu.lv) we have integrated them into the ReDSeeDS tool. The MOLA transformation has been compiled and made available within the ReDSeeDS transformation menu. More details about this integration process can be found in the book by Śmiałek and Nowakowski [11]. The transformations generate UML classes with embedded method code. These UML classes are then handled by the UML tool (Enterprise Architect) and its standard code generator, to produce Java code.

The resulting transformation from RSL to Hibernate ORM produced good quality, consistent code that could be used directly to implement the data access layer. In current work we did not approach at generating the database tables, but it can be noted that our solution shows that a complete persistence layer could be generated automatically from domain models in RSL. This is an interesting research direction and we treat this as future work.

Moreover, our future work will also include integration of the persistence layer within the Model-View-Presenter (MVP) [12] architectural pattern. The upper layers (View and Presenter) of a complete software system can be fully generated from RSL models as shown by Śmiałek et al. [10], [11], [26]. This approach does not generate any contents of the Model layer. However, we can approach at generating meaningful code for the CRUD operations. Such operations are used frequently in RSL scenarios, as illustrated in Figure 8. Links between scenario sentences (e.g. 'System saves book data') and domain statements (e.g. 'save book data') can be transformed into calls from the Presenter layer to the Model layer. The RSL environment allows for determining the type of operation (e.g. Create or Update) and thus appropriate database access code can be provided in proper places.

## References

[1] B. Berenbach, "A 25 year retrospective on model-driven requirements engineering," in *IEEE Model-Driven Requirements Engineering Workshop (MoDRE'12)*, 2012, pp. 87–91, DOI: 10.1109/MoDRE.2012.6360078.

[2] B. A. Berenbach, "Comparison of UML and text based requirements engineering," in *Companion 19th OOPSLA Conference*, 2004, pp. 247–252, DOI: 10.1145/1028664.1028766.

[3] D. Bjôrner, "Rôle of domain engineering in software development. why current requirements engineering is flawed!" *Lecture Notes in Computer Science*, vol. 5947, pp. 2–34, 2010, DOI: 10.1007/978-3-642-11486-1_2.

[4] H. Kaindl, M. Smialek, P. Wagner *et al.*, "Requirements specification language definition," ReDSeeDS Project, Project Deliverable D2.4.2, 2009, www.redseeds.eu.

[5] M. Śmiałek, A. Ambroziewicz, J. Bojarski, W. Nowakowski, and T. Straszak, "Introducing a unified requirements specification language," in *Proc. CEE-SET'2007, Software Engineering in Progress*. Nakom, 2007, pp. 172–183.

[6] *OMG Meta Object Facility (MOF) Core Specification, version 2.4.1, formal/2013-06-01*, Object Management Group, 2013.

[7] M. Śmiałek, J. Bojarski, W. Nowakowski, A. Ambroziewicz, and T. Straszak, "Complementary use case scenario representations based on domain vocabularies," *Lecture Notes in Computer Science*, vol. 4735, pp. 544–558, 2007, MODELS'07, DOI: 10.1007/978-3-540-75209-7.

[8] I. Jacobson, M. Christerson, P. Jonsson, and G. Övergaard, *Object-Oriented Software Engineering: A Use Case Driven Approach*. Addison-Wesley, 1992.

[9] M. Smialek, W. Nowakowski, N. Jarzebowski, and A. Ambroziewicz, "From use cases and their relationships to code," in *Second IEEE International Workshop on Model-Driven Requirements Engineering, MoDRE 2012*, 2012, pp. 9–18, DOI: 10.1109/MoDRE.2012.6360084.

[10] M. Smialek, N. Jarzebowski, and W. Nowakowski, "Translation of use case scenarios to Java code," *Computer Science*, vol. 13, no. 4, pp. 35–52, 2012, DOI: 10.7494/csci.2012.13.4.35.

[11] M. Śmiałek and W. Nowakowski, *From Requirements to Java in a Snap: Model-Driven Requirements Engineering in Practice*. Springer, 2015.

[12] M. Potel, "MVP: Model-View-Presenter the Taligent programming model for C++ and Java," Taligent Inc., Tech. Rep., 1996.

[13] C. Bauer and G. King, *Hibernate in Action (In Action Series)*. Greenwich, CT, USA: Manning Publications Co., 2004.

[14] H. Feddema, *DAO Object Model: The Definitive Reference*. O'Reilly Media, 2000.

[15] T. Stahl, M. Voelter, and K. Czarnecki, *Model-Driven Software Development: Technology, Engineering, Management*. Wiley, 2006.

[16] A. G. Kleppe, J. B. Warmer, and B. Wim, *MDA Explained, The Model Driven Architecture: Practice and Promise*. Addison-Wesley, 2003.

[17] A. Kalnins, J. Barzdins, and E. Celms, "Model transformation language MOLA," *Lecture Notes in Computer Science*, vol. 3599, pp. 62–76, 2005, MDAFA'04, DOI: 10.1007/11538097_5.

[18] *The MOLA Language, Reference Manual, Version 2.0 final*, University of Latvia, 2007, http://mola.mii.lu.lv/.

[19] M. Smialek and T. Straszak, "Facilitating transition from requirements to code with the ReDSeeDS tool," in *20th IEEE Requirements Engineering Conference (RE'12)*, 2012, pp. 321–322, DOI: 10.1109/RE.2012.6345825.

[20] W. Nowakowski, M. Śmiałek, A. Ambroziewicz, and T. Straszak, "Requirements-level language and tools for capturing software system essence," *Computer Science and Information Systems*, vol. 10, no. 4, pp. 1499–1524, 2013, DOI: 10.2298/CSIS121210062N.

[21] *OMG Unified Modeling Language, version 2.5, ptc/2013-09-05*, Object Management Group, 2013.

[22] J. van Wijngaarden and E. Visser, "Program transformation mechanics: A classification of mechanisms for program transformation with a survey of existing transformation systems," Utrecht University, Tech. Rep. UU-CS-2003-048, 2003.

[23] A. Kleppe, *Software Language Engineering: Creating Domain-Specific Languages Using Metamodels*. Addison-Wesley, 2008.

[24] J. Evermann and Y. Wand, "Toward formalizing domain modeling semantics in language syntax," *IEEE Transactions on Software Engineering*, vol. 31, no. 1, pp. 21–37, 2005, DOI: 10.1109/TSE.2005.15.

[25] M. Fowler, *Patterns of Enterprise Application Architecture*. Addison-Wesley, 2002, p. 401.

[26] M. Śmiałek, N. Jarzebowski, and W. Nowakowski, "Runtime semantics of use case stories," in *IEEE Symp. Visual Languages and Human-Centric Computing (VL/HCC'12)*. IEEE, 2012, pp. 159–162, DOI: 10.1109/VLHCC.2012.6344506.

DOI: 10.15439/2015F301
ACSIS, Vol. 5

# Sharing Developers' Mental Models through Source Code Annotations

Matúš Sulír, Milan Nosáľ
Department of Computers and Informatics
Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9, 042 00 Košice, Slovakia
Email: matus.sulir@tuke.sk, milan.nosal@gmail.com

***Abstract*—*Context:* Developers possess mental models containing information far beyond what is explicitly captured in the source code. *Objectives:* We investigate the possibility to use source code annotations to capture parts of the developers' mental models and later reuse them by other programmers during program comprehension and maintenance. *Method:* We performed two studies and a controlled experiment. *Results:* Developers' mental models overlap and thus can be shared. Possible use cases of shared annotations are hypotheses confirmation, feature location, obtaining new knowledge, finding relationships and maintenance notes. In the experiment, the presence of annotations reduced program comprehension and maintenance time by 34%. *Conclusion:* Annotations are a viable way to share programmers' thoughts.**

## I. Introduction

When programmers develop a new program, they continuously create a mental model, which is a representation of the program in their mind. They try to express the mental model in the programming language constructions. However, not all parts of the mental model are transferred into the source code and some details are lost.

### A. Motivation

For example, in small parts of a large information system, we can be forced to deal with a character encoding different than in the rest of the system because of the legacy library limitations.

Later, during the program maintenance phase, developers make a tremendous effort to recover such types of information from the source code [2]. However, they rarely persist their findings [3] and the situation occurs again and again.

We could record the information about the encoding from the aforementioned example in a form of a Java annotation `@Encoding("win-1250", reason="myOldLib 2.4")` above all affected classes or methods. It would be later possible to use the IDE (integrated development environment) searching capabilities on the given annotation to find all methods where this encoding is used if an encoding-related bug is reported or refactoring is planned to replace the legacy library.

### B. Aim

The purpose of this paper is to investigate the viability of annotations as a medium to share parts of developers' mental models.

**Hypothesis**: Annotations created by one group of developers are useful for comprehension and maintenance tasks performed by other developers.

We formulate the research questions as follows:

- **RQ1**: Do programmers' mental models overlap?
- **RQ2**: How do developers use shared annotations when they are available?
- **RQ3**: Does using annotations created by others improve program comprehension and maintenance correctness, time and confidence?

To answer each of the questions, we used an appropriate empirical research methodology [4].

## II. Concern Annotations

Before describing the studies, we briefly introduce the notion of concerns and their kinds.

### A. Basic Concepts

A *concern* can be characterized as a developer's intent of a particular piece of code: What should this code accomplish? How would I tersely characterize it? Is there something special about it? Some concerns can be obvious by looking at the code itself (chiefly from names of the identifiers), but many concerns are hidden.

A concern is similar to an aspect in aspect-oriented programming. In contrast to aspects, concerns can overlap – one piece of code can belong to multiple concerns [5]. This complicates the situation: We cannot simply name all classes and methods according to their concerns since one identifier can have only one name. Thus, we can also look at concerns as alternative names for identifiers.

It is possible for a class or method in Java to have more than one annotation. Therefore, source code annotations (attributes in C# terminology) are an ideal candidate to implement concerns.

This work was supported by VEGA Grant No. 1/0341/13 Principles and methods of automated abstraction of computer languages and software development based on the semantic enrichment caused by communication. The first two research questions are also elaborated in the second author's PhD thesis [1].

978-83-60810-66-8/$25.00©2015, IEEE

For each distinct concern, we recommend to create one Java *annotation type*. For example, we can create an annotation type `@Persistence` which tells us the code marked with it fulfills the task of persistent storage of objects.

Subsequently, in our imaginary program, we could mark the methods like `FileDialog.open()`, `Note.load()`, `Note.save()` and a class `FileFormat` with it. We will call them *annotation occurrences*.

At the same time, the first of the mentioned methods could be also annotated with the `@GUI` concern, as it presents a GUI (graphical user interface) dialog to a user.

Compared to traditional source code comments, concern annotations are more formal. We can use standard IDE features like navigating to the declaration, usages searching, refactoring [6] and other on them.

Concern annotations can have parameters to further specify their properties. They may be also commented by natural language comments if needed.

### B. Kinds of Concern Annotations

*Domain annotations* document concepts and features of the application (problem) domain. For example, all source code elements representing the feature of filtering a list of items can be marked with an annotation `@Filtering`. Similarly, all code related to bibliographic citations could be annotated by `@Citing`.

*Design annotations* document design and implementation decisions like design patterns, e.g., `@Observer`.

*Maintenance annotations* are intended to replace the traditional TODO and related comments. An example is the `@Unused` annotation for parts of code not used in a project at all.

## III. Mental Model Overlapping

Our first goal is to find out whether at least some of the concerns recognized by one programmer can be recognized by other developers. If so, then the mental model of one developer at least partially overlaps with the other persons' mental model. This is a necessary condition for concern annotation sharing to be useful.

### A. Method

We asked multiple programmers to independently annotate the source code of an existing program. Then we measured to what extent their annotations overlap.

*1) Materials:* For this and subsequent studies, we used EasyNotes[1] – a desktop application for bibliographic note-taking. It is a small-scale Java project consisting of around 2500 lines of code located in 33 classes. Except for scarce source code comments, it has no documentation available.

TABLE I
The number of recognized concerns and annotation occurrences per individual subjects

| Subject | Concerns | Occurrences |
|---|---|---|
| A | 11 | 70 |
| B | 12 | 56 |
| C | 24 | 108 |
| D | 20 | 79 |
| E | 12 | 56 |
| F | 14 | 89 |
| G | 17 | 140 |
| **Total (distinct)** | **46** | **464** |

*2) Participants:* This study had 7 participants:

| | |
|---|---|
| A | a researcher and lecturer with PhD in Computer Science, |
| B | an industrial Java programmer, |
| C | a postdoc and Java programmer, |
| D | an associate professor with extensive Java experience, |
| E, F | first-year PhD students, |
| G | the author of EasyNotes. |

None of the subjects, except for the author, had a previous experience with EasyNotes. The activity was individual and the participants were not allowed to interact during the experiment.

*3) Procedure:* First, the participants were given the original source code of EasyNotes without any annotations (commit `a299e64`). They had an arbitrary amount of time available to become familiar with the application both from an end-user and programmer perspective.

Next, they were asked to create a custom annotation type for each concern they recognized in the application and to mark classes, member variables and methods with the annotations they just created whenever they thought it was appropriate.

*4) Analysis:* Finally, we collected the modified projects and analyzed them semi-automatically for an overlap in annotation types and the use of the annotations on specific elements. A manual intervention was necessary because some participants used a slightly different name for the same concern – e.g., `@Tags` and `@Tagging` both represent the "tagging" concern.

### B. Results

For each participant, a set of created concerns (annotation types) was constructed. The number of concerns created by individual participants ranged from 11 to 24 and the number of annotation occurrences from 56 to 140 (see Table I).

*1) Concern Sharing:* We constructed a set of distinct concerns, i.e., an union of all sets of the concerns recognized by the participants. The size of this set, i.e., a number of distinct concerns, is 46 (the Total row in Table I). More than a half of them (26) was shared by at least two participants. We will call them *shared concerns*. A list of all shared concerns is in Table II.

[1] http://github.com/MilanNosal/easy-notes

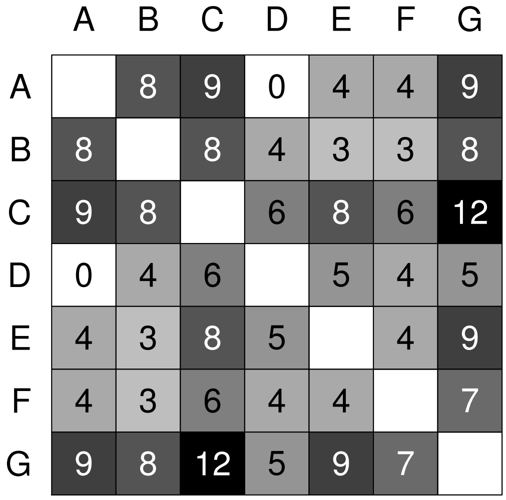
TABLE II
A LIST OF ALL SHARED CONCERNS IN EASYNOTES

| | Concern | Shared by *n* participants | Effective agreement |
|---|---|---|---|
| 1 | Searching | 6 | 61% |
| 2 | Note editing | 5 | 60% |
| 3 | Note change observing | 5 | 50% |
| 4 | Note presenting | 5 | 21% |
| 5 | Unused code | 4 | 67% |
| 6 | Tagging | 4 | 64% |
| 7 | Persistence | 4 | 41% |
| 8 | Links | 4 | 39% |
| 9 | Note adding | 4 | 38% |
| 10 | Data model | 4 | 36% |
| 11 | Loading notes | 4 | 35% |
| 12 | Saving notes | 4 | 31% |
| 13 | GUI | 4 | 26% |
| 14 | Note deleting | 4 | 18% |
| 15 | UI-model mapping | 4 | 4% |
| 16 | Filter implementation | 3 | 18% |
| 17 | TODO | 3 | 8% |
| 18 | Exceptions | 2 | 100% |
| 19 | Utilities | 2 | 50% |
| 20 | Model change watching | 2 | 17% |
| 21 | Filters management | 2 | 12% |
| 22 | Notes manipulation | 2 | 8% |
| 23 | Questions about code | 2 | 0% |
| 24 | Coding by convention | 2 | 0% |
| 25 | BibTeX | 2 | 0% |
| 26 | Domain entity | 2 | 0% |

*2) Occurrence Sharing:* Similarly, we constructed a set of all distinct annotation occurrences with a total size of 464 (as noted in Table I). 128 of them were so-called *shared annotation occurrences* which means they were shared by at least two participants.

We define an *effective agreement* as the number of shared annotation occurrences divided by the total number of annotation occurrences. For our EasyNotes study, the overall effective agreement was 27.59%. It is possible to see the values of effective agreement for individual concerns in Table II. We can consider the effective agreement of a specific concern its quality indicator – to what extent multiple developers agree about the mapping of this concern to source code elements.

*3) Agreement between Participants:* Fig. 1 shows the number of shared concerns between each pair of participants. We can obtain interesting insights from this matrix. The subjects A and D did not create any common annotation type in our study – this could be an indication of a huge difference in the mental models of these two people. In fact, D is a former or current supervisor of all other participants except A. On the other hand, G (the author of EasyNotes) shares the most concerns with all other people. This could mean that the source code author is the best annotator.



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | | 8 | 9 | 0 | 4 | 4 | 9 |
| B | 8 | | 8 | 4 | 3 | 3 | 8 |
| C | 9 | 8 | | 6 | 8 | 6 | 12 |
| D | 0 | 4 | 6 | | 5 | 4 | 5 |
| E | 4 | 3 | 8 | 5 | | 4 | 9 |
| F | 4 | 3 | 6 | 4 | 4 | | 7 |
| G | 9 | 8 | 12 | 5 | 9 | 7 | |

Fig. 1. The number of shared concerns between individual subjects

A similar matrix was constructed for concern occurrences. Qualitatively, it resembled the annotation type matrix.

### C. Threats to Validity

*1) Internal Validity:* While some hypotheses were outlined in this section, being an exploratory study [7], there is a need to properly quantify and statistically confirm or reject them. This suggest interesting future research directions.

*2) External Validity:* We performed the study only on a small-scale Java project. The participants were all from the same department which could affect their mental models. A more extensive study should be conducted in the future.

### D. Conclusion

We studied mental model overlapping, where parts of the mental model were represented by source code annotations. In our study, about

- 57% of all concerns
- and 28% of concern occurrences

were shared by at least two participants. This means there is a potential in recording and subsequent reuse of these data.

## IV. CONCERN ANNOTATIONS USE CASES

The goal of the second study is find out how third-party developers (i.e., not the annotation authors) use the annotations in the source code if they are available.

### A. Method

We conducted an observational study.

*1) Materials:* We copied all annotation types shared by at least two subjects (from the first study, see Table II for a list) into the EasyNotes project. For each of these annotation types, we merged all its occurrences (recognized by at least one developer) into the project. The resulting source code was manually edited for consistency by the EasyNotes author. It is published as the commit `f52872b`.

*2) Participants:* There were three participants:

K a first-year Computer Science PhD student,

L a masters degree student with a minor industrial experience,

M a professional developer with 2 years of industrial experience.

None of them had a previous knowledge of EasyNotes.

*3) Procedure:* First, all annotations were briefly introduced by the EasyNotes author. The subjects were reminded about common feature location possibilities of the NetBeans IDE.

Each participants was given the same task: To add a "note rating" (one to five stars) feature to EasyNotes. The fulfillment of this task required a modification of multiple application layers – from the model to the user interface.

We used a think-aloud method [7], i.e., the participants were kindly requested to comment their thoughts when comprehending the code.

### B. Results

We will now look at typical use cases of concern annotations during our study.

*1) Confirming Hypotheses:* The most common use of concern annotations was to confirm hypotheses about the code. For example, the participant K used the `@NotesSaving` annotation to confirm that a particular piece of stream-writing code actually saves notes.

*2) Feature Location:* In contrast to traditional comments, it is possible to use the Find Usages feature on an annotation type to find all concern occurrences. Our participants were finding the occurrences of the "filtering", "note adding", "note saving" concerns and others. This was considered helpful especially to find if they did not forget to implement the necessary methods for a particular aspect of the note rating feature.

*3) Non-Obvious Concerns:* The developers also used annotations to obtain new knowledge about the source code. For instance, a UI (user interface) code contained a method used both when adding a new note and when editing an existing one. However, just a note editing concern was obvious from a brief source code inspection. Only thanks to the concern annotation `@NoteAdding`, the participant M noticed the code is used for note adding, too.

*4) Elements Relationship:* The subjects noticed that if two or more elements are marked with the same annotation type, there is an implicit relationship between them. For instance, when using the MVC (Model-View-Controller) design pattern, the code in the model marked with a specific annotation is linked with the UI code with the same annotation.

*5) Maintenance Notes:* The annotation `@Unused` marks methods not used in the rest of the code. This helped the participants to skip them when scanning the code and thus save time.

### C. Conclusion

*1) Advantages:* The participants stated that compared to traditional natural language comments, annotations are much shorter and thus easier to spot. They are also better structured and usually less ambiguous. The ability to find all usages of a particular concern through standard features present in contemporary IDEs was also appreciated.

*2) Disadvantages:* The participant with an industrial experience (M) remarked there is a possible scaling problem. Even in a small project like EasyNotes, 26 shared concerns were identified. In large projects, where this number is expected to grow, some sort of concern categorization would definitely be needed. As Java annotations do not support inheritance, marking them with meta-annotations or sorting them to packages are possible solutions.

## V. The Effect of Annotations on Program Maintenance

We performed a controlled experiment to study the effect of the annotated source code on program comprehension and maintenance.

The guidelines to perform software engineering experiments on human subjects [8] were used. To present our findings, the experiment reporting guidelines [9] were followed. We customized them to the specific needs of this experiment.

### A. Hypothesis

Similar to Barišić et al. [10], we were interested in correctness, time and confidence.

We hypothesize that the presence of concern annotations in the source code improves program comprehension and maintenance correctness, time and confidence. Thus we formulate the null and alternative hypotheses:

**$\mathbf{H1_{null}}$**: The correctness of the results of program comprehension and maintenance tasks on an annotated project = the correctness on the same project without concern annotations.

**$\mathbf{H1_{alt}}$**: The correctness of the results of program comprehension and maintenance tasks on an annotated project > the correctness on the same project without concern annotations.

**$\mathbf{H2_{null}}$**: Time to complete program comprehension and maintenance tasks on an annotated project = time to complete them on the same project without concern annotations.

**$\mathbf{H2_{alt}}$**: Time to complete program comprehension and maintenance tasks on an annotated project < time to complete them the same project without concern annotations.

**$\mathbf{H3_{null}}$**: Participants' confidence of their answers to program comprehension questions on an annotated project = their confidence on the same project without concern annotations.

**$\mathbf{H3_{alt}}$**: Participants' confidence of their answers to program comprehension questions on an annotated project > their confidence on the same project without concern annotations.

We will statistically test the hypotheses with a confidence interval of 95% ($\alpha = 5\%$).

### B. Variables

Now we will define independent variables, i.e., the factors we control, and dependent variables – the outcomes we measure.

*1) Independent Variables:* There is only one independent variable – the presence of concern annotations in the project. It has a nominal scale which means there is a finite number of possible values without any meaningful ordering [4]. The levels (possible values) of this variable are: yes ("annotated") and no ("unannotated").

*2) Dependent Variables:* The correctness was measured as a number of correct answers (or correctly performed tasks) divided by the total number of tasks (5). The tasks are not weighted, each of them is worth one point. The assessment is subjective – by a researcher.

The second dependent variable is the time to finish the tasks. Its scale is of a ratio type since the ratio between two values is meaningful [4]. Although it was technically measured with millisecond precision, we will use the unit "minutes" rounded to two decimal places in subsequent analysis. We are interested mainly in the total time, i.e., a sum of times for all tasks.

Instead of measuring just time alone, it is possible to define efficiency as a number of correct tasks and questions divided by time. On one hand, efficiency depends on correctness, which already is a dependent variable. On the other hand, efficiency can deal with participants who fill the answers randomly to finish quickly [11]. We decided to use efficiency only as an auxiliary metric to make sure that time differences are still significant even if the correctness is considered.

For each comprehension question, we also asked a subject how confident (s)he was on a 3-point Likert scale: from Not at all (1) to Absolutely (3). Since we asked a subject about the confidence equally for each task, we consider it meaningful to calculate the mean confidence, which is the third dependent variable.

### C. Experiment Design

*1) Materials:* Again, the EasyNotes project was used. This time, we prepared two different versions:

- with shared concern annotations (as in the second study)
- and without annotations.

As the project was only scarcely commented, we deleted all traditional source code comments from both versions to remove a potential confounding factor. Only comments for the annotation types themselves were left intact, as we regard them as their integral part.

During this experiment, we used the NetBeans IDE.

*2) Participants:* We used 18 first-year master's degree Computer Science students as participants. Carver et al. [12] recommend to integrate software engineering experiments performed on students with teaching goals. We decided to execute the experiment as a part of the Modeling and Generation of Software Architectures course, which contained Java annotations in its curricula.

The course was attended by students focused not only on software engineering, but also on other computer science subfields. Inclusion criteria were set to select mainly students with a prospective future career as professional programmers. Additionally, as EasyNotes is a Java project, a sufficient Java language knowledge was required.

The experiment was performed during three lessons in the same week – the first time with four students, then 9 and finally with the remaining 5 students. Each session lasted approximately 1.5 hours. The study was executed in a separate room, so the participants were not disturbed.

*3) Design:* When assigning the subjects to groups, we applied a completely randomized design [4]. This means that each group received only one treatment – either an annotated or an unannotated program – and the assignment was random. Each participant drew a piece of paper with a number on it. Subjects with an odd number were assigned to the "annotated" group, participants with an even number to the "unannotated" one. Our design was thus balanced, with n=9 per group.

*4) Instruments:* To both guide the subjects and collect the data, we designed an interactive web form[2]. All fields in the form were mandatory, so the participants could not skip any task.

We asked the subjects to install a NetBeans plugin, SSCE[3]. Although it is not its primary feature, it provides an option to record programming sessions – time elapsed, a list of open files and NetBeans windows gaining the focus. Just before the start of each task, a subject clicked the button to start a new session, named after the task (e.g., "Filter"). Immediately after the task was finished, (s)he clicked the button again to end the session. The collected data were written to an XML file which the participants uploaded to the web form at the end of the experiment.

### D. Procedure

*1) Training:* At the beginning of the experiment, the users were given 5 minutes to familiarize themselves with EasyNotes from an end-user perspective, without looking at the source code. This provided them an overview of the application domain and helped them to better understand their subsequent tasks. Then, the session monitoring plugin was briefly introduced.

A short presentation about concern annotations usage in the NetBeans IDE followed. A researcher presented how to show a list of all available concerns, how to use the Find Usages feature on an annotation type and how to navigate from the annotation occurrence to an annotation type.

Just before each of the two maintenance tasks, the researcher presented the participants a running application with the required feature already implemented. This had two positive effects:

- It significantly lowered the task ambiguity. While a natural-language description was available in the web form during the tasks, seeing the finished application gave the participants greater confidence, so almost nobody asked unnecessary questions.
- We consider this a replacement for unit tests. As GUI code is notoriously difficult to test [13], we decided not implement the test code. At the same time, the

[2] http://www.jotformeu.com/sulir/sharing-annotations

[3] http://github.com/MilanNosal/sieve-source-code-editor

researcher's discretion about the task correctness was not indispensable during the experiment. Students just knew they finished the task when their application did the same thing as we had showed them.

In addition, the participants later uploaded their modified version of the source code to the web form. This way, potential disputes could be resolved without time stress.

*2) Tasks:* The experiment comprised of:

- one additive maintenance task (we will name it *Filter*),
- three program comprehension questions (*Test*),
- one corrective maintenance task (*Cite*),

in that order.

The tasks were formulated as follows:

Filter In a running EasyNotes application, load the sample notes and look at the searching feature (the down-right part of the window). Try how searching in the note text works (the option "Search in": "text"). Your task will be to add a filter to the EasyNotes application, which will search in the notes title (the option "title").

Cite In the EasyNotes application, there is a field "Cite as:" (the down-right window part). Currently, it displays information in the form: *somePubID* where somePubID is the value of the "PubID:" field. Your task is to modify the program so the "Cite as:" field will display information in the form: *\cite{somePubID}*.

Both tasks were simple, although the *Filter* task was slightly more complex than the latter. It required the creation of a new class with approximately 15 lines of code, whereas the *Cite* task could be accomplished by modifying just one source code line.

The questions asked in the *Test* about program comprehension were:

Q1 What does the `runDirMenuItemActionPerformed` in the class `easynotes.swingui.EasyNotesFrame` do?

Q2 How is the class `easynotes.model.abstractModel.UTFStringComparator` used in the EasyNotes project?

Q3 What method/s (and in which class) do perform note deleting?

*3) Demographic Data Collection:* At the end of the experiment, the form contained three demographic questions about the subjects' abilities:

- a general programming experience,
- their experience with Java, annotations and NetBeans
- and their English level.

Each question had possible answers on a 5-point Likert scale: from Beginner to Expert. We did not perform any specialized tests to measure programming experience, as a subjective opinion is good enough [14].

*4) Debriefing:* We also included a question asking to what extent did the subjects use annotations when comprehending the code. Possible answers ranged from Never to Always on a 5-point Likert scale. Finally, the form also contained a free-form question where the participants could describe how the annotations helped them in their own words.

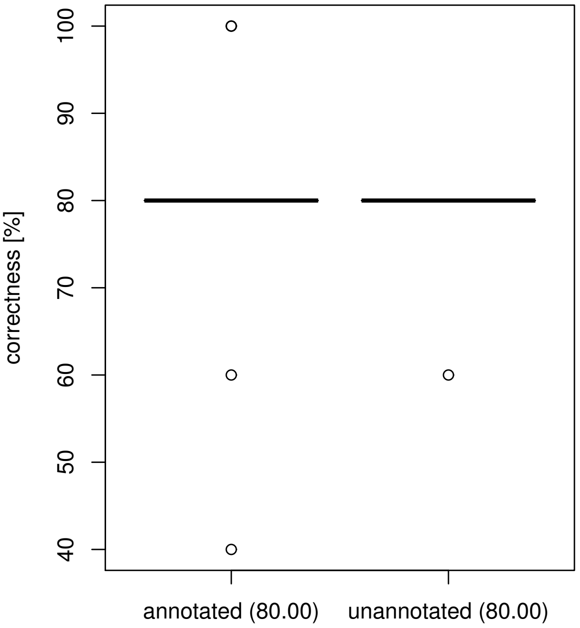


Fig. 2. The ratio of correct answers for each group.

### *E. Results*

The measured values and their summary is presented in Table III. To analyze the results, we used the R scripting language, auxiliary Ruby scripts and spreadsheets.

Each specific hypothesis considers one independent variable on a nominal scale with two levels (annotated, unannotated) and one dependent variable (either correctness, time or confidence). For each dependent variable, we displayed the values on a histogram and a normal Q-Q plot. None of the variables looked normally distributed, so we used the Mann-Whitney U test as a statistical test for our hypotheses.

*1) Correctness:* The median of correctness for both the "annotated" and "unannotated" group was 80%. Except for a few outliers, all subjects answered exactly 4 out of 5 questions correctly. See Fig. 2 for a plot.

The computed p-value (roughly speaking, the probability that we obtained the data by chance) is 0.3898, which is more than 0.05 (our significance level). This means we accept $\mathbf{H1_{null}}$. We did not prove that the presence of annotations has a positive effect on program comprehension and maintenance correctness.

As we can see in Table III (column Correctness), the most difficult question was Q2. Only two participants answered it correctly – both from the "annotated" group. The class of interest was not used in EasyNotes at all. This fact was noticeable by looking at the `@Unused` annotation.

TABLE III
THE EXPERIMENT RESULTS FOR INDIVIDUAL SUBJECTS

| The "annotated" group | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Correctness [true/false] | | | | | | Time [min] | | | | Efficiency [tasks/min] | Confidence [1-3] | | | | Files | Annotations useful? [1-5] |
| | Filter | Cite | Q1 | Q2 | Q3 | Total | Filter | Cite | Test | Total | | Q1 | Q2 | Q3 | Mean | | |
| 1 | 1 | 1 | 1 | 0 | 1 | 80% | 12.03 | 3.00 | 13.96 | 28.99 | 0.14 | 1 | 2 | 3 | 2.00 | 19 | 1 |
| 3 | 1 | 1 | 0 | 0 | 0 | 40% | 5.23 | 2.81 | 11.13 | 19.17 | 0.10 | 3 | 2 | 3 | 2.67 | 10 | 3 |
| 5 | 1 | 1 | 1 | 1 | 1 | 100% | 17.43 | 3.93 | 6.71 | 28.07 | 0.18 | 3 | 3 | 3 | 3.00 | 18 | 4 |
| 7 | 1 | 1 | 1 | 1 | 1 | 100% | 7.79 | 1.43 | 11.85 | 21.07 | 0.24 | 3 | 3 | 3 | 3.00 | 10 | 4 |
| 9 | 1 | 1 | 1 | 0 | 1 | 80% | 6.72 | 5.86 | 3.87 | 16.45 | 0.24 | 3 | 3 | 2 | 2.67 | 20 | 2 |
| 11 | 1 | 1 | 1 | 0 | 1 | 80% | 8.41 | 4.58 | 4.32 | 17.31 | 0.23 | 3 | 3 | 3 | 3.00 | 13 | 4 |
| 13 | 1 | 1 | 0 | 0 | 1 | 60% | 20.97 | 3.48 | 8.80 | 33.25 | 0.09 | 3 | 2 | 3 | 2.67 | 11 | 3 |
| 15 | 1 | 1 | 1 | 0 | 1 | 80% | 4.64 | 1.91 | 5.22 | 11.77 | 0.34 | 2 | 2 | 3 | 2.33 | 11 | 2 |
| 17 | 1 | 1 | 1 | 0 | 1 | 80% | 25.08 | 6.38 | 6.99 | 38.45 | 0.10 | 2 | 2 | 3 | 2.33 | 16 | 4 |
| **Median** | **1** | **1** | **1** | **0** | **1** | **80%** | **8.41** | **3.48** | **6.99** | **21.07** | **0.18** | **3** | **2** | **3** | **2.67** | **13** | **3** |
| **Std.dev.** | **-** | **-** | **-** | **-** | **-** | **18.56%** | **7.41** | **1.67** | **3.57** | **8.80** | **0.08** | **-** | **-** | **-** | **0.35** | **4.06** | **-** |

| The "unannotated" group | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Correctness [true/false] | | | | | | Time [min] | | | | Efficiency [tasks/min] | Confidence [1-3] | | | | Files | Annotations useful? [1-5] |
| | Filter | Cite | Q1 | Q2 | Q3 | Total | Filter | Cite | Test | Total | | Q1 | Q2 | Q3 | Mean | | |
| 2 | 1 | 1 | 1 | 0 | 1 | 80% | 2.84 | 4.72 | 10.66 | 18.22 | 0.22 | 3 | 2 | 3 | 2.67 | 9 | NA |
| 4 | 1 | 1 | 1 | 0 | 1 | 80% | 8.76 | 23.45 | 8.10 | 40.31 | 0.10 | 3 | 2 | 3 | 2.67 | 19 | NA |
| 6 | 1 | 1 | 1 | 0 | 1 | 80% | 18.24 | 5.23 | 5.62 | 29.09 | 0.14 | 3 | 2 | 1 | 2.00 | 17 | NA |
| 8 | 1 | 1 | 1 | 0 | 1 | 80% | 6.47 | 5.59 | 11.23 | 23.29 | 0.17 | 3 | 2 | 3 | 2.67 | 8 | NA |
| 10 | 1 | 1 | 1 | 0 | 1 | 80% | 4.82 | 9.64 | 17.50 | 31.96 | 0.13 | 3 | 2 | 3 | 2.67 | 8 | NA |
| 12 | 1 | 1 | 1 | 0 | 1 | 80% | 11.11 | 2.09 | 11.30 | 24.50 | 0.16 | 2 | 2 | 2 | 2.00 | 13 | NA |
| 14 | 1 | 1 | 0 | 0 | 1 | 60% | 30.73 | 7.19 | 5.50 | 43.42 | 0.07 | 2 | 2 | 3 | 2.33 | 17 | NA |
| 16 | 1 | 1 | 1 | 0 | 1 | 80% | 12.56 | 18.39 | 16.07 | 47.02 | 0.09 | 3 | 2 | 3 | 2.67 | 14 | NA |
| 18 | 1 | 1 | 1 | 0 | 1 | 80% | 25.54 | 9.59 | 12.94 | 48.07 | 0.08 | 3 | 2 | 3 | 2.67 | 16 | NA |
| **Median** | **1** | **1** | **1** | **0** | **1** | **80%** | **11.11** | **7.19** | **11.23** | **31.96** | **0.13** | **3** | **2** | **3** | **2.67** | **14** | **NA** |
| **Std.dev.** | **-** | **-** | **-** | **-** | **-** | **6.67%** | **9.56** | **6.98** | **4.18** | **11.06** | **0.05** | **-** | **-** | **-** | **0.30** | **4.22** | **-** |

*2) Time:* The differences in the total time for all tasks between two groups are graphically depicted in the box plot in Fig. 3. The median time changed from 31.96 minutes for the "unannotated" group to 21.07 minutes for the "annotated" one, which is a decrease by 34.07%.

The p-value of time is 0.0252, that is less than 0.05. The difference is statistically significant, therefore we reject $\mathbf{H2_{null}}$ and accept $\mathbf{H2_{alt}}$. The presence of concern annotations improves the program comprehension and maintenance time.

It is possible to see from Table III (column Time) that the median time for each individual task was better for the "annotated" group. The most prominent difference was for the task *Cite*. This can be due to the fact that the project contained the concern annotation `@Citing` which helped the participants find the relevant code quickly.

The median of efficiency, which we defined as the number of correctly performed tasks (and answers) divided by total time, raised by 42.32% (p=0.0385). This means the time improvement is significant even if we take correctness into account.

*3) Confidence:* The median of mean confidence is the same for both groups (2.67), as obvious from Table III, column Confidence / Mean and Fig.4. The p-value is 0.1710 (>

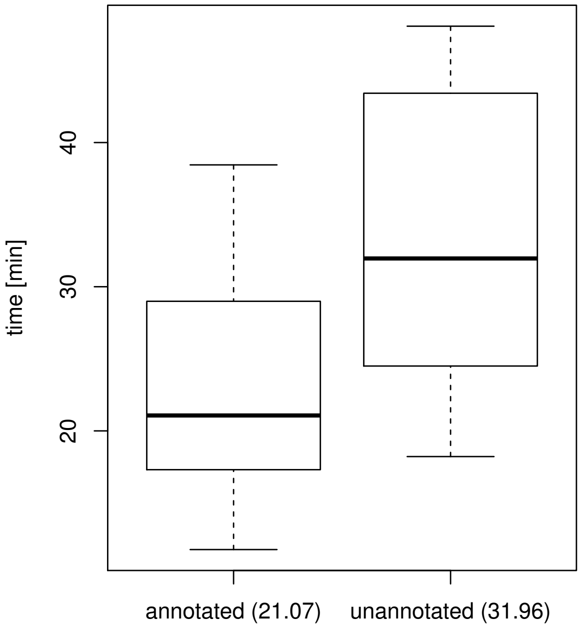


Fig. 3. Time to complete comprehension and maintenance tasks on an annotated vs. unannotated project

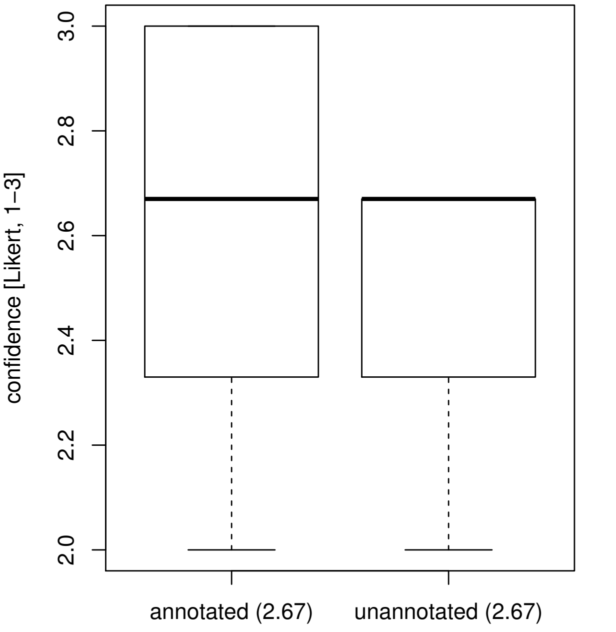


Fig. 4. The mean confidence for the "annotated" and "unannotated" group

0.05) and therefore we accept $\mathbf{H1_{null}}$. The effect of concern annotations on confidence was not demonstrated.

Looking at individual questions, Q2 was clearly perceived the most difficult. This corresponds with our previous finding about the correctness. An interesting fact is that no participant in the "unannotated" group was confident enough to select the level 3 (Absolutely), while in the "annotated" group, there were 4 such subjects and two of them really answered correctly.

*4) Other Findings:* Although not included in our hypotheses, we also measured how many unique Java source files did the subjects open in the IDE during the whole experiment. We excluded the annotation type files in these numbers. The results are in Table III, column Files. There was only a marginal and statistically insignificant improvement in medians (from 14 to 13).

As seen from Table III, column "Annotations useful?", concern annotations were perceived relatively helpful by the participants (median 3 from 5-point Likert scale).

Answers to a free-form question asking how specifically were the annotations useful included:

- faster orientation in a new project,
- faster searching (mainly through Find Usages on annotations),
- less scrolling,
- "they helped me to understand some methods",
- "annotations could be perfect, but I would have to get used to them".

### *F. Threats to Validity*

To analyze threats to validity of this experiment, we used [15] as guidelines.

*1) Construct Validity:* Similar to Kosar et al. [11], we compensated the students with points for the participation in the experiment, which increased their enthusiasm. Unlike them, we did not reward the participants with bonus points for good results because our experiment spanned several days with the same tasks and this would motivate the students to discuss the task details with classmates, which could negatively affect the construct validity (interaction between subjects). Furthermore, the participants were explicitly told not to share experiment details with anyone. Therefore, we do not consider executing the experiment in three separate sessions an important validity threat.

To measure confidence, we used only a 3-point Likert scale. This decision was not optimal. Because subjects rarely select the value of 1 (which can be interpreted as guessing an answer), there were only two values left. This could be one of the reasons we did not found a statistically significant difference in confidence.

*2) Internal Validity:* There could be a selection bias in the experiment because we selected the participants using subjective criteria. As we already mentioned, we concentrated on subjects with a potential future career as developers.

We divided the subjects into groups randomly. Another possibility was a quasi-experiment (nonrandom assignment), i.e., to divide the subjects evenly according to the most important co-factor affecting the results, like their programming experience. However, random assignments tend to have larger effect sizes than quasi-experiments [16].

We did not perform a full pilot testing with third-party participants, only tested the comprehension questions on one of the researchers. We rejected 3 out of the 6 prepared questions because we considered them too difficult and ambiguous. Despite this, all tasks were either completed by almost all participants (Filter, Q1, Q3, Cite) or by almost none (Q2) during the experiment. This negatively affected the results for the "correctness" variable.

During the actual experiment, we used a native language (Slovak) version of the web form to eliminate the effect of English knowledge on the results. While the source code was in English, this did not present a validity threat since all participants had at least a medium level (3 on a 5-point Likert scale) of English knowledge.

*3) External Validity:* We invited only students to our experiment, no professional developers. We can take this fact positively – concern annotation consumers are expected to be mainly junior developers, whereas potential annotation creators are mostly senior developers. Furthermore, some students may already work in companies during their study.

EasyNotes is a small-scale Java project – around 3 KLOC (thousands of lines of code), including annotations. The effect of concern annotations on larger programs should be investigated.

*4) Reliability:* The concern annotations training (tutorial) was presented manually by a researcher. However, there were two independent researchers which took turns.

The experiment is replicable, as we published the data collection form (`http://www.jotformeu.com/sulir/sharing-annotations`) which contains both the guidelines and links to the materials (two versions of the EasyNotes project).

*5) Conclusion Validity:* A small number of subjects (n=9 per group) is the most important conclusion validity threat. If we used a paired design (to assign both treatments to every subject), we could easily reach n=18. However, the participants would quickly become familiar with EasyNotes and the second set of tasks would be affected by their knowledge.

### G. Conclusion

We successfully confirmed the hypothesis that concern annotations have a positive effect on program comprehension and maintenance time. The group which had concern annotations available in their project reached a time more than 34% shorter than the group without them ($p < 0.05$).

On the other hand, we did not discover a significant change in correctness and confidence. The difference was neither negative (which could mean the annotations are confusing because of their "visual noise") nor positive and it was probably a result of the discussed validity threats.

## VI. Related Work

### A. Concerns

Reinikainen et al. [17] present concern-base queries to reason about the program. However, their approach is based on UML (Unified Modeling Language) models, while we use the source code of a system.

Niu et al. [18] propose the application of HFC (hierarchical faceted categories) on source code. Their approach requires specialized tools whereas we use source code annotations which have a standard IDE support.

### B. Source Code Projections

In [5], we used concern annotations as one of the ways to perform source code projections. Projections allow to look at one source code from multiple different perspectives. For example, an IDE plugin can filter all code marked with a specific annotation type and display it in an editable window – even if it is spread across multiple files.

In this work, we take a more pragmatic approach. We investigate the possibility to use concern annotations in contemporary IDEs, using the features already available, like Find Usages. Above all, we perform three empirical studies to assess the viability and find possible use cases of concern annotations in everyday developer's life.

### C. Annotations

Today, one of the most common applications of annotations is configuration. A programmer marks selected source code elements with appropriate annotations. Later, they are processed either by an annotation processor (during compilation) or by reflection (at runtime). For example, annotations can be used to define concrete syntax in parser generators [19] and to declare references between elements in a domain-specific language [20]. This way, annotations can indirectly modify the annotated program semantics. In contrast, our approach utilizes annotations just as clues for a programmer which are processed by an IDE when needed.

Sabo and Porubän [21] use source code annotations to preserve design patterns. Therefore, their approach does not include recording and sharing domain and maintenance annotations.

In Java, annotations can be only applied to packages, classes, member variables, methods and parameters. @Java [22], an extension to the standard Java language, brings annotations below the method level. It allows to mark individual source code statements like assignments, method calls, conditions and loops with annotations. This could be useful to capture e.g., algorithmic design decisions with concern annotations.

### D. Mental Model Overlapping

Revelle et al. [23] also studied concern overlapping. However, their study included only two concern sets (compared to our 7). Their results are positive, too. This further confirms our hypothesis that it is possible to share mental models.

### E. Code Bookmarks

Source code bookmarks are one of the standard features present in today IDEs. They allow a developer to mark specific source code lines with notes and later list the bookmarks and jump to each of them.

Collective code bookmarks [24] provide a way to share parts of developers' mental models. The plugin allows to share code bookmarks between developers in a team. However, this approach is IDE-dependent and the bookmarks are not saved to the original source code files which complicates standard procedures like versioning and merging.

### F. Maintenance Notes

Developers often write "TODO comments" like `// TODO: fix this` to mark parts of source code which need their attention [25]. IDEs can then try to parse and display these notes in a task list window.

Our approach is more formal, as annotations are a part of the standard Java grammar and can be parsed unambiguously. Furthermore, it is possible to distinguish between multiple maintenance note types through individual annotation types.

## VII. Conclusion and Future Work

In this paper, we presented an idea to use Java source code annotations to capture parts of developers' mental model, namely their concerns (intents), thus the name "concern annotations". Each concern (e.g., "searching", "editing", "GUI code", "unused code") is implemented as an annotation type. Subsequently, all classes and methods relevant to that concern are marked with the given annotation. Two studies and one experiment were conducted to assess the practical implications of this approach.

It is possible to share these annotations because the developers' mental models overlap: More than a half of the concerns created by one of 7 developers in our study was recognized by at least two of them. More than 1/4 of concern occurrences (locations in source code where a particular annotation is used) were shared by at least two participants.

An interesting future research question is to what extent the source code itself, when the same program is created separately by multiple people, overlaps.

In the second study, we discovered that concern annotations are particularly useful to confirm hypotheses about the code, locate the features, find out non-obvious concerns which a method fulfills, discover hidden relationships between elements. Concern annotations can be also used as a replacement of traditional TODO comments.

In the controlled experiment, we showed there is a statistically significant improvement of development time when performing program comprehension and maintenance tasks on a small-scale Java project. The group which had an annotated version of the same program available, consumed 1/3 less time than the group which did not have concern annotations present in the source code.

In our studies, the source code was commented scarcely or not at all. An interesting future comparison would consider an annotated program vs. a program without annotations, but with high-quality traditional source code comments instead.

Currently, the source code is annotated manually by a programmer, which is a time-consuming task. An interesting future work would be a method of (semi-)automatic annotation.

## References

[1] M. Nosáľ, "Leveraging program comprehension with concern-oriented projections," PhD thesis, Technical University of Košice, Apr. 2015.

[2] T. D. LaToza, G. Venolia, and R. DeLine, "Maintaining mental models: A study of developer work habits," in *Proceedings of the 28th International Conference on Software Engineering*, ser. ICSE '06. New York, NY, USA: ACM, 2006, pp. 492–501. http://dx.doi.org/10.1145/1134285.1134355

[3] W. Maalej, R. Tiarks, T. Roehm, and R. Koschke, "On the comprehension of program comprehension," *ACM Trans. Softw. Eng. Methodol.*, vol. 23, no. 4, pp. 31:1–31:37, Sep. 2014. http://dx.doi.org/10.1145/2622669

[4] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Springer Publishing Company, Incorporated, 2012.

[5] J. Porubän and M. Nosáľ, "Leveraging program comprehension with concern-oriented source code projections," in *3rd Symposium on Languages, Applications and Technologies*, ser. OpenAccess Series in Informatics (OASIcs), M. J. V. Pereira, J. P. Leal, and A. Simões, Eds., vol. 38. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2014, pp. 35–50. http://dx.doi.org/10.4230/OASIcs.SLATE.2014.35

[6] J. Kollár, I. Halupka, S. Chodarev, and E. Pietriková, "pLERO: Language for grammar refactoring patterns," in *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, Sept 2013, pp. 1503–1510.

[7] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical Software Engineering*, vol. 14, no. 2, pp. 131–164, 2009. http://dx.doi.org/10.1007/s10664-008-9102-8

[8] A. Ko, T. LaToza, and M. Burnett, "A practical guide to controlled experiments of software engineering tools with human participants," *Empirical Software Engineering*, vol. 20, no. 1, pp. 110–141, 2015. http://dx.doi.org/10.1007/s10664-013-9279-3

[9] A. Jedlitschka and D. Pfahl, "Reporting guidelines for controlled experiments in software engineering," in *Empirical Software Engineering, 2005. 2005 International Symposium on*, Nov 2005, pp. 95–104. http://dx.doi.org/10.1109/ISESE.2005.1541818

[10] A. Barišić, V. Amaral, M. Goulão, and B. Barroca, "Quality in use of domain-specific languages: A case study," in *Proceedings of the 3rd ACM SIGPLAN Workshop on Evaluation and Usability of Programming Languages and Tools*, ser. PLATEAU '11. New York, NY, USA: ACM, 2011, pp. 65–72. http://dx.doi.org/10.1145/2089155.2089170

[11] T. Kosar, M. Mernik, and J. C. Carver, "Program comprehension of domain-specific and general-purpose languages: comparison using a family of experiments," *Empirical Software Engineering*, vol. 17, no. 3, pp. 276–304, 2012. http://dx.doi.org/10.1007/s10664-011-9172-x

[12] J. Carver, L. Jaccheri, S. Morasca, and F. Shull, "A checklist for integrating student empirical studies with research and teaching goals," *Empirical Software Engineering*, vol. 15, no. 1, pp. 35–59, 2010. http://dx.doi.org/10.1007/s10664-009-9109-9

[13] A. Memon, I. Banerjee, and A. Nagarajan, "GUI ripping: reverse engineering of graphical user interfaces for testing," in *Reverse Engineering, 2003. WCRE 2003. Proceedings. 10th Working Conference on*, Nov 2003, pp. 260–269. http://dx.doi.org/10.1109/WCRE.2003.1287256

[14] J. Feigenspan, C. Kastner, J. Liebig, S. Apel, and S. Hanenberg, "Measuring programming experience," in *Program Comprehension (ICPC), 2012 IEEE 20th International Conference on*, June 2012, pp. 73–82. http://dx.doi.org/10.1109/ICPC.2012.6240511

[15] A. A. Neto and T. Conte, "Threats to validity and their control actions – results of a systematic literature review," Universidade Federal do Amazonas, Technical Report TR-USES-2014-0002, Mar. 2014.

[16] V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. K. Sjøberg, "A systematic review of quasi-experiments in software engineering," *Information and Software Technology*, vol. 51, no. 1, pp. 71 – 82, 2009, special Section - Most Cited Articles in 2002 and Regular Research Papers. http://dx.doi.org/10.1016/j.infsof.2008.04.006

[17] T. Reinikainen, I. Hammouda, J. Laiho, K. Koskimies, and T. Systa, "Software comprehension through concern-based queries," in *Program Comprehension, 2007. ICPC '07. 15th IEEE International Conference on*, June 2007, pp. 265–270. http://dx.doi.org/10.1109/ICPC.2007.36

[18] N. Niu, A. Mahmoud, and X. Yang, "Faceted navigation for software exploration," in *Program Comprehension (ICPC), 2011 IEEE 19th International Conference on*, June 2011, pp. 193–196. http://dx.doi.org/10.1109/ICPC.2011.18

[19] J. Porubän, M. Forgáč, M. Sabo, and M. Běhálek, "Annotation based parser generator," *Computer Science and Information Systems*, vol. 7, no. 2, pp. 291–307, Apr. 2010. http://dx.doi.org/10.2298/CSIS1002291P

[20] D. Lakatoš, J. Porubän, and M. Bačíková, "Declarative specification of references in DSLs," in *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, Sept 2013, pp. 1527–1534.

[21] M. Sabo and J. Porubän, "Preserving design patterns using source code annotations," *Journal of Computer Science and Control Systems*, vol. 2, no. 1, pp. 53–56, 2009.

[22] W. Cazzola and E. Vacchi, "@Java: Bringing a richer annotation model to Java," *Computer Languages, Systems & Structures*, vol. 40, no. 1, pp. 2–18, 2014, special issue on the Programming Languages track at the 28th ACM Symposium on Applied Computing. http://dx.doi.org/10.1016/j.cl.2014.02.002

[23] M. Revelle, T. Broadbent, and D. Coppit, "Understanding concerns in software: insights gained from two case studies," in *Program Comprehension, 2005. IWPC 2005. Proceedings. 13th International Workshop on*, May 2005, pp. 23–32. http://dx.doi.org/10.1109/WPC.2005.43

[24] A. Guzzi, L. Hattori, M. Lanza, M. Pinzger, and A. van Deursen, "Collective code bookmarks for program comprehension," in *Program Comprehension (ICPC), 2011 IEEE 19th International Conference on*, June 2011, pp. 101–110. http://dx.doi.org/10.1109/ICPC.2011.19

[25] M. Storey, J. Ryall, R. Bull, D. Myers, and J. Singer, "TODO or to bug," in *Software Engineering, 2008. ICSE '08. ACM/IEEE 30th International Conference on*, May 2008, pp. 251–260. http://dx.doi.org/10.1145/1368088.1368123

DOI: 10.15439/2015F364
ACSIS, Vol. 5

# Feature Model Driven Generation of Software Artifacts

Roman Táborský and Valentino Vranić
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava, Bratislava, Slovakia
E-mail: crudecrude@gmail.com, vranic@stuba.sk

***Abstract*—The objective of feature modeling is to foster software reuse by enabling to explicitly and abstractly express commonality and variability in the domain. Feature modeling alone is not sufficient to express all the aspects of the software being developed. Other models and, eventually, code is necessary. These software assets are being configured by the feature model based on the selection of variable features. However, selecting a feature is far from a naive component based approach where feature inclusion would simply mean including the corresponding component. More often than not, feature inclusion affects several places in models or code to be configured requiring their nontrivial adaptation. Feature inclusion recalls *transformation* and this is at heart of the approach to feature model driven generation of software artifacts proposed in this paper. Features are viewed as solution space transformations that may be executed during the generative process conducted by the feature model configuration.**

***Index Terms*—feature modeling; transformation; metatransformation; generative process; reuse**

## I. Introduction

FEATURE modeling is an approach used in software development proposed in 1990's [1] with the growing popularity of software product lines. The objective of this approach to modeling is to foster software reuse by enabling to explicitly and abstractly express commonality and variability in the domain. Based on commonality and variability, appropriate implementation mechanisms can be selected [2].

Feature modeling alone is not sufficient to express all the aspects of the software being developed. Other models and, eventually, code is necessary. These software assets are being configured by the feature model based on the selection of variable features. However, selecting a feature is far from a naive component based approach where feature inclusion would simply mean including the corresponding component. More often than not, feature inclusion affects several places in models or code to be configured requiring their nontrivial adaptation. Feature inclusion recalls *transformation* and this is at heart of the approach to feature model driven generation of software artifacts proposed in this paper.

The rest of the paper is organized as follows. Section II discusses the possibilities of representing software artifacts in feature modeling. Section III identifies specifics of feature modeling necessary for the employment of this technique in the generative process. Section IV explains how features can be perceived as transformations, which is the essence of the approach proposed in this paper. Section V presents the overall process of feature model driven generation of software artifacts. Section VI reports the implementation of the approach. Section VII presents the evaluation. Section VIII discusses related work. Section IX concludes the paper.

## II. Representing Software Artifacts in Feature Modeling

Feature modeling can be used to configure software assets—models and code—in order to create software instances that exhibit desired features. One way to achieve this is by employing so-called superimposed variants [3] where the software models or code contain all the variants that are being reduced based on the features selected—or not selected—in the corresponding feature model [4], [5]. The FeatureHouse project [6] implements an approach that uses these models and allows language independent source code generation.

The pure::variants software tool [7] uses a specialized family model to represent a feature to architecture mapping. In this model, it is necessary to specify the type of the impact on the software instance. There are several possible impact specifications that allow for a wide scope of software artifact to be created, such as files, file fragments, XSLT transformations, conditional XML or text, C/C++ flag files, makefiles, class alias files, or symbolic links to folders or files.

These types of interaction can be effectively used to describe various architectural parts of elements, but provide no way of direct implementation of quality attributes [8]. Extra-functional[1] features can be mapped to specific modules/components or software artifacts in general by implying rules the software artifacts have to comply with, such as specific testing procedures or documentation requirements, or by specifying a human interaction task to be performed, such as evaluating a managerial decision or performing human assessment of a feature instance in a finalized product (e.g., evaluating user experience or GUI usability) [10].

In automatic software composition, software artifacts represent the reusable and generated parts of software. By putting together these parts and rules expressed by extra-functional

[1] We use term *extra-functional*—as proposed by Mary Shaw [9]—to refer to requirements and features that go beyond software system functionality instead of more widely used, but potentially confusing term non-functional.

978-83-60810-66-8/$25.00©2015, IEEE

features, it is possible to automatically compose large parts of software.

## III. Feature Modeling for the Generative Process

To drive the generative programming by a feature model, it is necessary to represent the software assets in a configurable manner. Also, the feature model must be capable of expressing different transformations of the software assets to produce the software artifacts that correspond to the features that have been selected.

### A. Feature Types

As has been explained in Section II, features can be functional or extra-functional. Some functional features exhibit a crosscutting nature, i.e., they affect several distinct and often unrelated parts of the system. Extra-functional features commonly correspond to the quality requirements (e.g., security or performance) or to the impact of the software being created to its environment.

Furthermore, internal and external features can be distinguished. Internal features are contained within the software being created and they can be changed during the development process. External features represent environment functionality and quality implications and these are part of the deployment environment and thus are not affected by the development process. An example of an internal feature is a configuration file for a web server to which the product is going to be deployed. The web server and its running environment is an example of an external feature.

### B. Feature Implications

The inclusion or exclusion of specific features in a particular feature model configuration has an impact on the final software product being created. Apart from the features that serve purely organizational purpose, which is mostly to group other features, the features in a feature model can be viewed as abstract elements having some impact on the final software product. This idea is crucial to automated software generation and configuration because a feature can be viewed as prescribing a change to the configuration or generative process. This means that the whole generative process consists of a set of events that are implied by the inclusion or exclusion of the features in a specific configuration.

The set of features that are included in the generative process is a result of a feature model configuration process. This process resolves variability in the feature model and a fully-specified feature model configuration can be used as an input to the generative process.

A feature inclusion can have various kinds of impact on a final software product: file or folder operations, deployment rules to provide documentation, realizing the testing requirements, etc. It is necessary to distinguish between a fully autonomous feature impact that requires no human interaction to deliver a final software asset (e.g., source code generation) and an impact that requires human interaction (e.g., designing a splash screen). With respect to the fully automated generative programming approach, the latter is not applicable. One way of integrating human interaction into the generative process is to create task placeholders that notify developers their input is needed for the process to continue.

Autonomous actions can be described by a set of events that perform a particular computer operation. This leads to the concept of extending the feature model by including the information of these events that are implied by the specific features in the model. By this, we get the description of the generative process event chain that has to be executed when processing a particular configuration of a feature model.

### C. Generative Process

A generative process based on a feature model is driven by its configuration, i.e., by selecting the features. This process represents the transformation of the input models or code into resulting software artifacts according to the feature selection. The process can be restricted to only a single solution space transformation or it can represent a complex multi-tier set of intermediate actions where each of these can be perceived as a stand-alone process with its own inputs and outputs. In any case, the key issue is how to realize the impact of the feature inclusion on the underlying software assets. This is different for functional and extra-functional features.

*1) Functional Feature Inclusion:* Functional features can be directly mapped to software artifacts such as source code or resource files. The relationships between functional features can be quite complex and the problem of feature interaction can aries.

A simple example of a functional feature is the choice of the data provider for some other feature. The corresponding feature diagram is displayed in Figure 1. In this paper, a simple FODA-style feature diagram notation [1] is used. A feature diagram is a tree whose nodes represent features that can be selected or not for the resulting configuration. For a feature to be selected, its parent feature has to be selected. Empty circle ended edges connect parent features with their optional features. An arc over a group of edges means the corresponding features are mutually exclusive (alternative). Non-decorated edges connect mandatory features. We do not elaborate on textually expressed constraints and default dependency rules that are necessary to overcome the limitations of the tree structure of feature diagrams [11], [12], as for the purposes of the approach proposed here, these can be considered as any other feature relationships to be applied in feature model configuration.

A configurable piece of code corresponding to this model could in C# look like this:[2]

```
DataProvider provider = new DataProvider();
provider.DataSource = new <Template Field>();
Page.GridView.DataSource = provider;
Page.Databind();
```

Choosing the *XML Data source* changes the second line of the code sample to the following one:

[2] The further examples in this paper are in C#, too, if not stated otherwise.

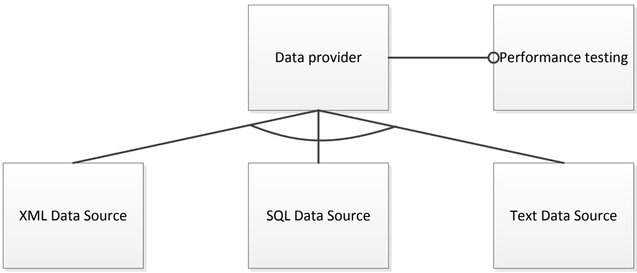


Fig. 1. The choice of the data provider as a functional feature and performance testing as an extra-functional feature.

```
provider.DataSource = new XMLDataSource(params []);
```

Another example would be an inclusion of the logging feature. This example is a little bit more complex as the points in source code to which the feature has to be bound are changing with as the code base grows. Aspect-oriented programming can be applied here to express so-called join points declaratively and address them without having to actually modify their source code representation.

*2) Extra-Functional Feature Inclusion:* Extra-functional features are a more difficult problem than functional features as it is rarely possible to map them to functional software artifacts. However, it is possible to implement the tests of the corresponding software artifacts against the conditions stated by extra-functional features. Consider the performance testing requirement as an example. The corresponding optional performance testing feature (see Figure 1) represents the fact that the software product has to conform to performance criterion and therefore needs to be tested against this criterion. This type of a feature can be directly represented in source code:

```
class XMLDataSource {
}
```

Having the *XMLDataSource* class to represent a functional feature when the extra-functional feature *Performance testing* is included, the source code can be transformed as follows:

```
class XMLDataSource {
    private TestContext testContextInstance;
    ///<summary>
    ///Gets or sets the test context which provides
    ///information about and functionality for the current test run.
    ///</summary>
    public TestContext TestContext {
        get {
            return testContextInstance;
        }
        set {
            testContextInstance = value;
        }
    }
    [TestMethod]
    public void PerformanceTesting() {
        throw new NotImplementedException();
    }
}
```

The *XMLDataSource* class has been enriched by methods and properties that support the performance testing. Among these is the *PerformanceTesting()* method, which has a placeholder that states it must be implemented manually.

## IV. Features as Transformations

Fully configured feature model provides a list of all features that are affecting the generative process. Each of these features that are included provides a partial information on what actions shall be taken during the automated software creation. A transformation is an entity that represents these actions to be taken for a particular feature. Consequently, the generative process becomes a composition of all transformations included in the particular configuration that was the input for generative process. A problem with this approach is that it is necessary to define the order in which the transformations will be executed. There are several possible solutions:

- Explicitly adding the ordering information in the input feature model configuration
- Traversing the feature model configuration structure in a predefined way
- Providing a priority property to transformations

An explicitly stated order in the feature model depends on the knowledge of priority in which the transformations have to be executed. This means that if we add a new feature into the model, the whole model needs to be examined to accommodate the changes in priority. Traversing the feature model in a predefined way is based on an idea that a feature model is a tree, so it can be scanned breadth-first or depth-first. In both cases, a priority parameter has to be introduced into features.

Many actions are common to different features. Consider creating a file, creating a folder, or making a text input text into a file. These transformations are actually generic, but rely on specific parameters in achieving their result. Therefore, a transformation is in fact a transformation template that uses the information provided to create a specific transformation instance. Consequently, a transformation in our approach is an entity consisting of the following elements:

- List of events, i.e., actions to be performed
- List of requirements that are imposed on the feature model
- List of metatransformations that are included in this transformation

Transformations are bound one-to-one to features in the feature model. This means that every feature that is included in the feature model and is relevant to the generative process of the final solution has its a transformation assigned to it. The feature node in the feature model also carries all the additional information that is bound to the transformation assigned to it.

### A. Transformation Reusability

The process of the transformation design requires an interaction on part of a domain engineer to provide the necessary domain information specific to the project and a software engineer to design the transformations in such a way that they implement the information provided by the domain engineer and the requirement analysis of the corresponding features. For an effective cooperation between the domain and software engineer, it is useful to distinguish different levels of transformations with respect to reuse:

- Specialized transformations that can be used only for the specific features in the specific configuration of the feature model
- Specialized transformations that can be used only for the specific features, but in any configuration of the feature model
- Domain dependent generic transformations, which can be used across multiple software product lines in the same domain
- Domain independent generic transformations, which are the most reusable transformations as they can be included in different software product lines across multiple domains

Distinguishing these transformation types helps designing transformations that are as generic as possible at their level.

### B. Transformation Hierarchy

An atomic transformation is a transformation impossible or unwanted to be decomposed into smaller transformations. Thus, even though such a transformation may consist of differentiable actions, the transformation is conceptually perceived as a one. Complex transformations are transformations that can be decomposed into a set of transformations where each of these lower level transformations represent an atomic transformation or a complex one (see Figure 2). Complex transformations themselves can have their own event chain besides the event chains of atomic transformations they embrace.

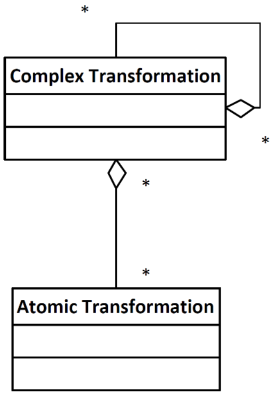


Fig. 2. Atomic and complex transformations (UML).

### C. Transformation Inheritance

The inheritance relationship known from object-oriented programming can be applied to transformations, too (see Figure 3). It is possible to use inheritance as a way of combining ancestor and descendant event chains. There are several possible approaches to inheriting event chains:

- The ancestor event chain remains the same
- Items are added or removed to the event chain
- The event chain is completely overridden in the descendant transformation

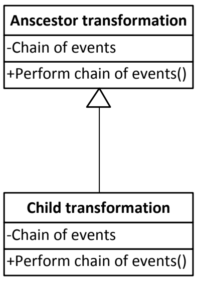


Fig. 3. Inheritance between transformations (UML).

The inheritance model can also be perceived as a way of organizing transformations into logical groups or packages. With respect to this, the inheritance is purely a tool of categorization and it is not necessary to maintain the typical parent–child class relationship, i.e., if transformations are represented as classes, it is not necessary to support inheritance mechanism at the level of methods and attributes.

### D. Metatransformations

There are some features that have a global effect that spans throughout the whole software product (or its significant part). Quality features, such as logging requirement or performance and security constraints, represent a typical example. Including such a feature with the corresponding transformation into the configuration leads to the necessity of modifying other transformations in one of the following ways:

- Modifying the event chain of a transformation
- Modifying the requirements of a transformation
- Providing information that is defined by the transformation requirements

The modification of the event chain means that some actions are added or removed to or from the transformation being modified, or some properties of these actions are changed. This leads to the connection with the other two types of modification that can be defined on their own or are a result of this first type of modification.

### E. Transformation Requirements

Seeing a transformation as a transformation template that instantiates a specific transformation based on external information, such as the file name in the create file transformation, constitutes the need for the external information required to perform this transformation. Therefore, it is necessary to include the information in the feature model or the associated feature model configuration to allow for this. However, having two features represented by the same type of transformation, the information provided can vary (see Figure 4).

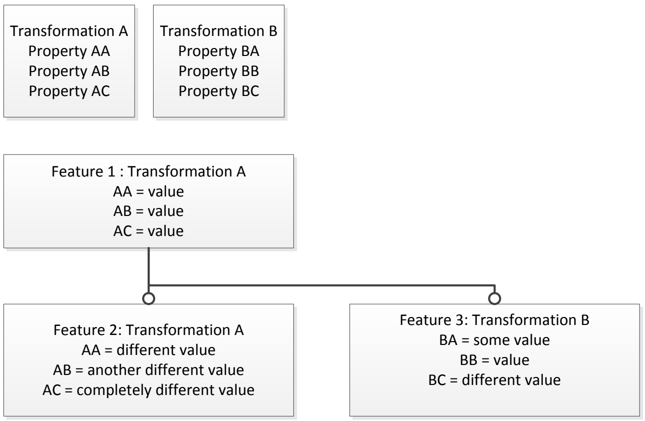


Fig. 4. Mapping transformations and properties to features.

Each transformation defines these requirements as a list of items that contain the particular facts about the requirements. This means that upon including a transformation in the feature model, it is necessary to specify its requirements.

## *F. Transformation Granularity*

There are two extreme approaches with respect to extending feature models with transformations. One approach is to target the separation of concerns with a large number of small transformations. This allows us to isolate the implications of transformations on the final software product into many small groups independent of each other (see Figure 5). The advantage of this approach is that the transformation model is easily changed without the need to analyze the impact on the whole transformation/feature model tree.

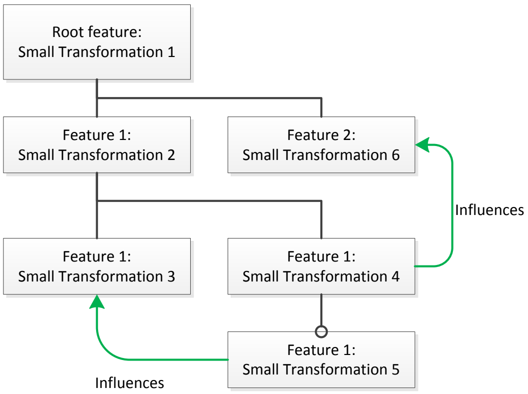


Fig. 5. Relationships between small independent transformations.

The other approach is to describe most actions in one global transformation that can be assigned to the root feature in the feature model. This global transformation is afterwards modified by the metatransformations that are assigned to the rest of the feature model nodes and their execution is based on the inclusion or exclusion of these features in a specific configuration (see Figure 6).

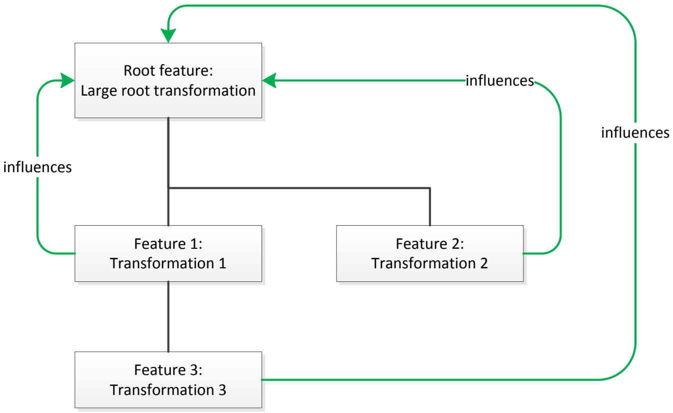


Fig. 6. The transformations influencing a large root transformation.

With respect to the transformation granularity, it is also important to consider the number of complex transformations in the model. Again, there are several approaches that can be used. One approach is to define only the simplest actions as atomic transformations and then to compose other transformations out of these atomic transformations. Another approach is to provide fairly large transformations that perform a substantial part of the tasks connected to a particular feature. It is also possible to employ an approach that lies somewhere inbetween of these extreme viewpoints. This is possible for the high or low granularity extremes and also for the complexity aspect of transformations. For example, metatransformations can be avoided by providing specific transformations for each feature.

## *G. Including Transformations in the Feature Model*

Since one transformation corresponds to one feature, their inclusion into the feature model means simply assigning the corresponding information to each feature (see Figure 7).

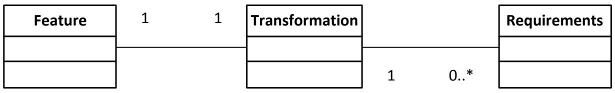


Fig. 7. Transformations are associated with features (UML).

One way of connecting the feature model with transformations is to include it directly in the feature model (an XML representation in used in this example):

```
<feature>
    <feature>
        <transformation>
            <!–– Transformation information including
            the event chain, metatransformation
            information, requirements, etc. ––>
        </transformation>
    </feature>
</feature>
```

There are two main problems connected with this approach: the degree of the transformation reuse between different models is reduced and it is necessary to parse the transformation information separately for every feature, even though the type of the transformation they use can be the same (e.g., create a file).

Another approach is to store the transformation definitions outside the feature model. A sample structure of this can be a feature model represented in XML and the transformations defined as C# classes that are used by the generator:

```
<feature transformationClassName=”CreateFile”
   fileName=”Samplefile.cs”>
</feature>

public class CreateFile : Transformation {
   public override void ExecuteTransformation() {
      ... // Create a file
   }
}
```

With this approach, it is necessary only to specify the transformation type and requirements of this transformation. The information about the metatransformations is not included in the feature model as it is internal to the transformation system and including this information in the feature model would be redundant.

## V. The Overall Process

The overall process of employing the feature model driven generation of software artifacts is as follows:

1) The input to the process is a feature model
2) Transformations are assigned to the feature model
3) Transformation requirements are provided where possible
4) A specific configuration of the feature model is created
5) Configuration specific requirements of the transformations are provided
6) The metatransformations are executed
7) The requirements and changes to the configuration model are incorporated or provided
8) The generator processes the fully specified configuration of the feature model and performs the event chain of the transformations included in it
9) The output of this process is a generated instance of the software system based on the input feature model, its configuration, and the transformations specified in the feature model

Transformation requirements are being fulfilled at three levels:

1) Feature model level
2) Feature model configuration level
3) Feature model configuration level after the execution of metatransformations

One may wonder whether it is not possible to merge the latter two levels into one. This is not possible because the metatransformations depend on the information provided at level 2 and therefore it is necessary to provide this information before executing level 3.

Often, the impact of a feature inclusion is not easy to analyze. Therefore, a stepwise approach to the transformation defining the impact of a feature design can help to analyze the impact of each atomic operation that a transformation consists of and also it can help to decide the order in which the transformations are applied. This is important because in many cases the final impact of a transformation depends on the order of the transformation in the sequence in which transformations are executed.

In the manual transformation application, it is not necessary to have this order predefined, but when a software asset generator employs an extended feature model as an input to perform the execution sequence, it must be deterministic and specified before the generative process starts as the generator is not aware of the final implications during the generative process: it merely executes the transformations that result from the input feature model configuration.

In summary, the main objectives of the stepwise transformation design are:

- Create complex transformation by a stepwise application of low-level transformations
- Assess the solution space after each low-level transformation in a stepwise manner
- Analyze the whole generative process by exploring it by this stepwise approach

When applying this approach, it is necessary to record the steps taken in the manual application of transformations. This recording can be further specified to create a complex transformation that can be used in the generative process. It also allows to analyze the impact of transformations that perform mass file renaming or other large-scale operations. The last important information contained in this recording is the transformation order, which allows to analyze and prioritize transformations in the generative process.

## VI. Implementing the Transformation Approach

The proposed approach of the feature model driven generation of software artifacts has been implemented in the .NET framework. A study of implementing the family of simple web sites has been performed.

### A. Applying the Transformation

It takes three steps to apply a transformation:

- Process all metatransformations in the model
- Check for requirements
- Execute the transformation

Consider the transformation called *CreateStaticHTMLPage* as an example. This transformation utilizes its parameters to fill a predefined HTML template string that is processed with the *String.Replace()* method. The template defines the HTML file and contains placeholders that are replaced with the values of the transformation parameters:

```
#region html text template
protected string htmlContent =
      ”<!DOCTYPE html PUBLIC \”−//W3C//DTD XHTML 1.0
            Strict//EN\”
      \”http://www.w3.org/TR/xhtml1/DTD/xhtml1−strict.dtd\”>” +
      ”<html xmlns=\”http://www.w3.org/1999/xhtml\”
```

```
        xml:lang=\”en\”> <head> ” +
    ”<meta http−equiv=\”content−language\”
        content=\”en\” />” +
    ”” +
    ”<title>Title placeholder</title>” +
    ”</head>” +
    ”” +
    ”<body>” +
    ”Content placeholder” +
    ”</body>” +
    ”</html>”;
#endregion
```

This transformation has four parameters:

- *PageName*: the file name that is used when creating the HTML file
- *htmlTemplate*: the template that is set up can be customized with this parameter
- *htmlTitle*: the text that replaces the *Title* placeholder
- *htmlContent*: the text that replaces the *Content* placeholder

If the transformation is included in the feature model, the configuration with these parameters:

```
<feature name=”StaticContent−History” ID=”2”
    Transformation=”org.crd.dp.CaseStudy.SimpleWebFinal,
        org.crd.dp.CaseStudy.
    SimpleWebFinal.Transformations.CreateStaticHTMLPage”
    htmlContent=”&lt;h1&gt;History&lt;/h1&gt;&lt;p&gt;
        Lorem Ipsum&lt;/p&gt;”
    htmlTitle=”StaticPage− History”
        PageName=”Site\\History.html” />
```

creates the following HTML file:

```
<!DOCTYPE html PUBLIC ”−//W3C//DTD XHTML 1.0
    Strict//EN”
    ”http://www.w3.org/TR/xhtml1/DTD/xhtml1−strict.dtd”>
<html xmlns=”http://www.w3.org/1999/xhtml” xml:lang=”en”>
    <head>
        <meta http−equiv=”Content−Type”
            content=”text/html; charset=utf−8” />
        <meta http−equiv=”content−language” content=”en” />
        <title>StaticPage− History
        </title>
    </head>
    <body><h1>History</h1>
        <p>Lorem Ipsum
        </p>
    </body>
</html>
```

### B. Feature Model Configuration

In this sample implementation, the feature model configuration is represented by an XML file. The structure of this file follows the tree structure of the feature diagram:

```
<?xml version=”1.0” encoding=”utf−8” ?>
<featuremodel>
<feature>
    <feature>
        <feature />
        <feature />
    </feature>
    <feature />
< /feature>
</featuremodel>
```

Each XML feature node has three compulsory attributes:

- name
- ID
- Transformation

Accordingly, the simplest feature node looks like this:

```
<feature name=”DynamicContentProvider” ID=”7”
    Transformation=”Transformations.Empty” />
```

The *name* and *ID* attributes have solely the purpose of identifying the node when the node is processed. The *Transformation* attribute specifies the transformation that will be used with this feature.

The transformation attribute consists of two parts delimited by a comma. The first part represents the dynamic link library that contains the transformation, and the second part the full class name of the transformation. The dynamic link library has to be a .NET managed library. Therefore, a filled transformation attribute looks like this:

```
Transformation=”org.crd.dp.CaseStudy.SimpleWebFinal, ...
    Transformations.Empty”
```

This model is afterwards transformed into an object model contained within the *TransformationStore* object.

### C. Transformations

The transformations in this implementation are based on a single class called *Transformation*. This class implements the *ITransformation* interface. This interface provides the basic methods needed by the generator to process the transformation:

```
public interface ITransformation {
    CheckPrerequisites();
    ExecuteTransformation();
    GetParameterNames();
    GetMetaTransformations();
    ...
}
```

The *CheckPrerequisites()* method does basic requirement checking before a transformation is processed. In this implementation, only a basic verification whether the transformation requirements are fulfilled is performed. However, it is possible to write a model aware method at the level of a metatransformation that can check the transformation dependencies. A sample of a model aware transformation is provided with the dynamic access log page.

The *ExecuteTransformation()* method represents the action which is contained within the transformation. This method is called in the final step of model processing.

The *GetParameterNames()* method is used in the XML file parsing, when it provides the parameter names to be retrieved from the feature node attributes.

The *GetMetaTransformations()* method provides a way how to retrieve the metatransformations that are connected with this transformation. This allows to connect a metatransformation list with a transformation and by this provide it with model awareness, which means that it can influence other transformations in the model within the possibilities provided by the connected metatransformations.

Two lists are initialized in the transformation class constructor:

```
protected Transformation() {
    metaTransformations = new List<MetaTransformation>();
    parameters = new Dictionary<string, object>();
}
```

These two lists contain the metatransformation list and key–value pairs of parameters. As the metatransformation list is of type *List <MetaTransformation>* and the interface requires *List <IMetaTransformation>*, it has to be casted:

```
public virtual List<IMetaTransformation>
    GetMetaTransformations() {
    return metaTransformations.ConvertAll(
        mt => (mt as IMetaTransformation));
}
```

The implementation of the transformation class provides two other important methods, *SetParameter()* and *GetParameter()*, which are the basis for parameter provisioning:

```
public object GetParameter(string name) {
    return parameters[name];
}
public void SetParameter(string name, object value) {
    if(parameters.ContainsKey(name))
        parameters[name] = value;
    else
        parameters.Add(name,value);
}
```

### D. Composite transformations

To cope with complex transformations, composite transformations, i.e., transformations that contain other transformations can be used. In the following implementation sample, the composition is based on overridden methods in a descendant class:

```
public class CompositeTransformation : Transformation {
    private TransformationStore transformations;
    protected CompositeTransformation(): base() {
        transformations = new TransformationStore();
    }
    public override string CheckPrerequisites() {
        foreach(ITransformation transformation
            in transformations.Store)
        { transformation.CheckPrerequisites();
        }
        ...
    }
    public override void ExecuteTransformation() {
        foreach(ITransformation transformation
            in transformations.Store)
        { transformation.ExecuteTransformation();
        }
    }
    ...
}
```

As it is observable form the sample code, the composite transformation contains *TransformationStore*, which is basically an ordered list of transformations. The *ExecuteTransformation()* and *CheckPrerequisites()* methods operate on this list. In the case this list contains another composite transformation, the transformations are processed in a depth-first recursive way.

The problem that arose with implementing this list was that at the point of creating an instance of a composite transformation, the child transformations could not access the parent transformation parameters as these were not yet extracted from the XML file and often these child transformations rely on the information provided in the model. Therefore, a new method called *InstantiateChain()* was introduced to the composite transformation. This method is called after the transformation requirements are extracted to the parent transformation. The name of this method suggests that this child transformation list represents a list of events as described in Section IV, which can be modified by metatransformations as suggested in Section IV-D.

### E. Metatransformations

The metatransformations are implemented by the *MetaTransformation* class. The difference between this class and the *Transformation* class is that the *MetaTransformation* class is model aware. This means that it can traverse the feature model and make changes to it. The specific changes are shown with transformation implementation samples. The difference is in the *ExecuteTransformation method()*:

```
ExecuteTransformation(object model, TransformationStore store);
```

In this method, the *TransformationStore* object is passed, representing the feature model. To allow for the modification of composite transformations, this method is implemented in a recursive way:

```
ExecuteTransformation(objectmodel, TransformationStorestore) {
    ...
    foreach(var trans in store.Store) {
        if(IsSubclassOfClass(typeof(
            CompositeTransformation),trans.GetType())) {
            ExecuteTransformation(model,
                ((CompositeTransformation)trans).Store);
        }
    }
}
```

### F. Generator

The generative process consists of these steps:

- Parse the feature model configuration from the XML file

- Parse and execute the metatransformations contained within the transformation from the XML file
- Check the requirements of the parsed transformations
- Execute the transformation

The step of checking the requirements before parsing the metatransformation is omitted as it is contained within the metatransformation parsing step. To support these steps, a parser object is introduced. This object is represented by the *IFeatureModelParser* interface:

```
public interface IFeatureModelParser {
    ParseFeatureModel(...);
    ExecuteTransformationChain(...);
    ParseMetaTransformations(...);
    CheckPrerequisites();
}
```

This object contains the methods that provide the functionality required to perform these steps. The *XMLModelParser* class is used to represent this object. This class provides the functionality required based on an XML feature model representation. The steps will be described in separate sections.

*1) Parse the Feature Model Configuration File:* The first step that is necessary is to translate the transformations from the XML feature model configuration into an object model. The transformations are parsed in a top-down order. It is possible to override this behavior using the *Priority* attribute at transformation nodes. First, the assembly and class name are parsed from the XML node and basic reflection is performed to create an instance of the transformation:

```
var assembly = Assembly.Load(assemblyName);
var ttype = assembly.GetType(typeName);
var transformationInstance = ttype.GetConstructor(
    Type.EmptyTypes).Invoke(null) as ITransformation;
transformationInstance.SetID(Convert.ToInt32((string)
    node.Attributes[”ID”].Value));
```

After creating an instance of the specified transformation, the *GetParameterNames()* method is used to obtain the list of parameters that this transformation uses. Afterwards, the XML node attributes that correspond to this list are copied into the dictionary containing the parameter key–value pairs:

```
foreach (var parameter in parameterNames) {
    if (node.Attributes[parameter] != null)
    transformationInstance.SetParameter(
        parameter, parameter != null ?
        node.Attributes[parameter].Value : null);
}
if (node.Attributes[”Priority”] != null)
    transformationInstance.SetPriority(
        node.Attributes[”Priority”].Value);
```

The last step to be done is to add the transformation to the *TransformationStore* object. This object serves as an advanced list for storing transformations. The enhancements against a standard list lie in the *AddTransformation()* methods that allow priority based insertion of transformations into the list. Another change is that simple and composite transformations are added in a different way as with composite transformations it is necessary to call the *InstantiateChain()* method to create instances of child transformations:

```
if (IsSubclassOfClass(typeof(CompositeTransformation),
    transformationInstance.GetType())) {
    if (node.Attributes[”Priority”] == null)
        store.AddTransformation(((CompositeTransformation)
            transformationInstance));
    else
    store.AddTransformation(((CompositeTransformation)
        transformationInstance),
        transformationInstance.GetPriority());
}
else {
    if (node.Attributes[”Priority”] == null)
        store.AddTransformation((Transformation)
            transformationInstance);
    else
        store.AddTransformation((Transformation)
            transformationInstance,
            transformationInstance.GetPriority());
}
```

The *IsSubclassOfClass()* method uses .NET reflection to recursively check for a match in all ancestor classes up to the *Object* class. Reaching the *Object* class signals that we are at the top of inheritance chain as in .NET the *Object* class is the topmost class from which all classes implicitly inherit. This step ends by adding all the transformation objects to the store, by which they become a part of the object model making the XML model unnecessary.

*2) Parse and Execute the Metatransformations:* After adding the transformations into *TransformationStore*, it is possible to perform metatransformations over this object model. The metatransformations are extracted from all transformations preserving their order as in the store:

```
foreach (Transformation trans in store.Store) {
    foreach (var a in trans.GetMetaTransformations()) {
        string[] parameterNames =
            a.GetParameterNames().ToArray();
        foreach (var parameter in parameterNames) {
            if (trans.GetParameter(parameter) != null)
            a.SetParameter(parameter, parameter != null ?
                trans.GetParameter(parameter) : null );
        }
        if (trans.GetPriority() != null)
            a.SetPriority(trans.GetPriority());
        ...
        // Add the metatransformation to the temporary store
        ...
}
```

The code for adding metatransformation to the temporary metatransformation store is similar to the code regarding common transformations. As a metatransformation can also be a composite transformation, it is again necessary to call the *InstantiateChain()* method. After obtaining a complete metatransformation store, it is possible to proceed with checking the prerequisites and perform the execution of metatransformations:

```
foreach (IMetaTransformation trans in metaStore.Store) {
    trans.CheckPrequisites();
}
foreach (IMetaTransformation trans in metaStore.Store) {
    trans.ExecuteTransformation(model,store);
}
```

After performing this last step, the changes to the transformations contained in the processed metatransformations have been applied to the transformation object model and therefore it is possible to perform the final prerequisite check over the model and proceed with executing the transformations.

*3) Check the Requirements of Parsed Transformations:* The requirement checking is simple. The only thing that is necessary is to call the *CheckPrerequisites()* method over all the transformations in the *TransformationStore* object. The current implementation uses a simple fault detection mechanism that is based on raising an exception when a problem occurs. One of the signals used is the *TransformationParameterNullException* exception. This signal means that a parameter expected at the XML model level was not provided:

```
public class TransformationParameterNullException : Exception {
    public TransformationParameterNullException( string transID,
        string parameter): base(”Transformation” + transID +
            ”:Parameter ” + parameter + ” was not defined.”){}
}
```

This exception can be raised afterwards in the *CheckPrerequisites()* transformation method:

```
public override string CheckPrerequisites() {
    if (GetParameter(”PageName”) == null)
        throw new TransformationParameterNullException(
            this.GetID().ToString(), ”PageName”);
}
```

With metatransformations it is possible also to check for transformation dependencies using the enhanced model aware method with the necessary parameters: *CheckPrerequisites(**object** model, TransformationStore store).*

*4) Execute the Transformations:* The precondition for this step is that the *TransformationStore* object contains a list of transformations that is prepared in a way that the metatransformations have been applied and the prerequisites checked. Afterwards, the *ExecuteTransformation()* method is called in a loop for each of the transformations contained in the list:

```
public void ExecuteTransformationChain(
    TransformationStore store) {
    foreach (var transformation in store.Store) {
        transformation.ExecuteTransformation();
    }
}
```

The order in which the transformations are executed is defined by their order in the *TransformationStore* object. After this step, the software artifacts specified in the transformations are created.

## VII. Evaluation

To evaluate the approach of feature model driven generation of software artifacts, we developed a study of the family of simple web sites comprising all the possibilities that may arise with features and transformations described in the previous section. Figure 8 shows the corresponding feature diagram.

The page features (ID 2, 3, and 4) embrace two ways of creating the text content: statically, by an HTML document (ID 2 and 3) or dynamically, by a script that generates the HTML document (ID 4).

A model aware transformation is introduced with the dynamic page feature (ID 4), with the utilization of a metatransformation providing the model traversal. Variability is introduced with the data provider (ID 5) providing an XML or Microsoft SQL database backend.

Optional features connected to the root feature (ID 10, 11, and 12) represent crosscutting features that require either metatransformations to change the actual features (ID 10 and 11) or they represent a parameter influencing the generative process (ID 12) to show an implementation of the development or generative environment property.

## VIII. Related Work

The pure::variants approach [8], mentioned in Section II, embraces a large set of predefined transformations that are assigned to particular features in the family model. The difference lies in the implementation where pure::variants is relying on XML transformation definitions and the solution proposed here uses C# classes, which is more flexible, allowing for custom programmed transformations.

Edicts [13] is another approach that aims at the mapping of features to source code parts. In addition, Edicts supports different binding times. The concept of binding time [14] should be taken into account when creating feature binding points in the suggested superimposed architectural framework.

XANA [15] strives for bringing closer the development process to end users using feature modeling. It decouples software product line design and implementation, which is to be performed by more technically knowledgeable users or professional developers, from application derivation, which is intended to be manageable by non-technical end users. Application derivation assumes not merely feature selection, but also providing parameters for parameterized features. A similar kind of decoupling can be applied to the approach proposed here. Generic transformations could be provided as a framework. Accompanied by an appropriate development environment extension, these generic transformations could be accessible to end users.

The superimposed variants approach [16] provides a way of mapping features to variabilities in external models, which can be used to activate or deactivate particular parts of the superimposed architectural framework. The transformation based approach proposed here is related to the idea of superimposed variants with respect to the external system of transformations used as the superimposed architecture or model. Differently than in our approach, the superimposed variants approach utilizes external models that are configured [16]. It is also possible to create such transformations that would prepare

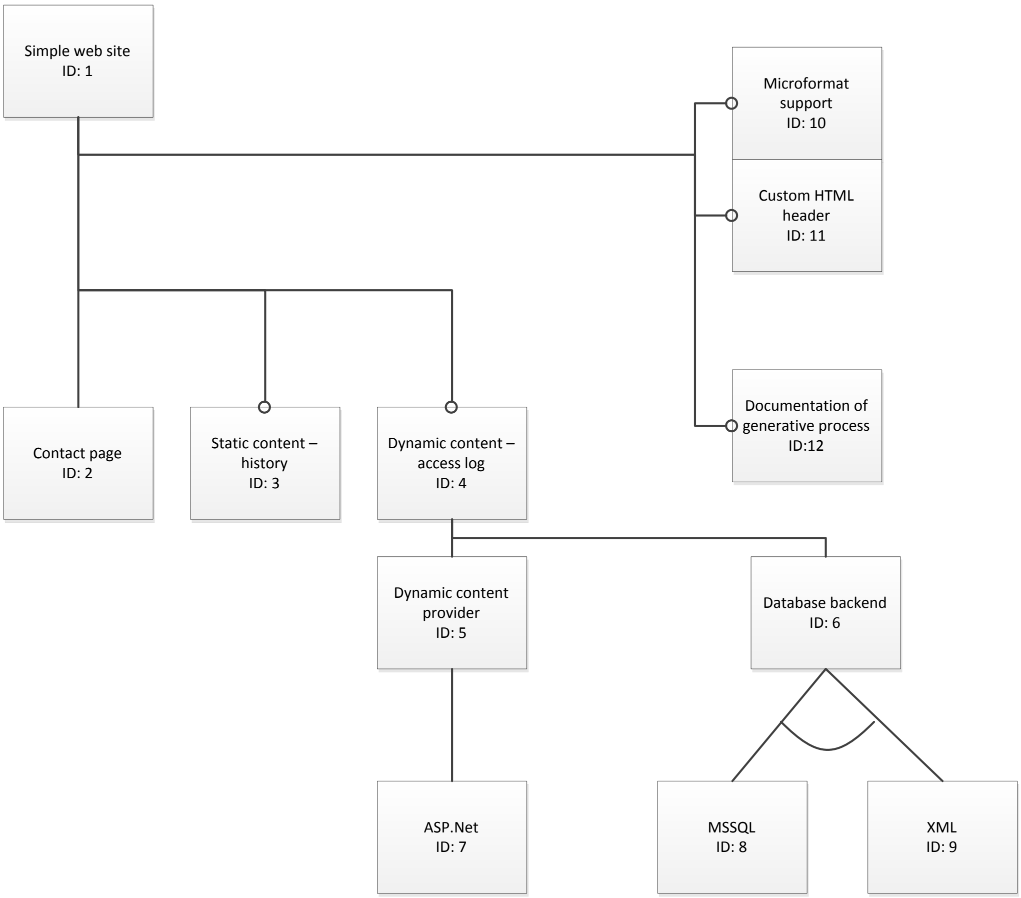


Fig. 8. The feature diagram of the family of simple web sites.

and configure additional models effectively emulating the superimposed variants approach.

One of the actions that is realized by metatransformations in the approach proposed here is the parameter replacement. This is similar to the template text replacement based on generic methods in generative programming for C and C++ [17]. However, the approach proposed here is different in the way that there are no restrictions on what is a transformation template parameter. Another difference is that in a template system, template fields are replaced and then the code is built, but in the approach proposed here, a transformation can represent a complex event chain, and not just a simple text replacement.

Dynamic code structuring [18], [19], [20] is based on explicit representation of possibly overlapping concerns in code for providing different perspectives. In the approach proposed here, dynamic code structuring can be applied to the code that defines transformations. However, dynamic structuring is potentially applicable to feature models themselves. In its essence, featural software decomposition is a decomposition by concerns with features representing the concerns, including the crosscutting ones [12]. Feature models with different organization of features in the feature diagrams can be equivalent [17]. Moreover, a feature can have alternative decompositions into subfeatures, including not being decomposed at all. The different representations of the same feature model may suit different stakeholders or situations and the transformation code attached to it can be presented in different ways accordingly. For large feature model presentation, design pattern detection techniques [21] may be of interest. Feature models can be represented as grammars [22], in which case grammar refactoring could be applied [23] to obtain different views.

## IX. Conclusions and Further Work

This paper proposes an approach of feature model driven generation of software artifacts, in which features are viewed as solution space transformations that may be executed during the generative process conducted by the feature model configuration. The approach has been evaluated on a study of the family of simple web sites comprising all the possibilities that may arise with features and transformations.

The main advantages of this approach are is that the system of transformations is basically self-contained and does not require additional modeling techniques except for the enhanced feature model. The code within the transformations is not limited with respect to its effects on the resulting software system behavior, i.e., anything that can be achieved by manual code writing can be achieved by appropriate transformations. In large part, the flexibility of the proposed approach lies in the concept of metatransformation. Metatransformations are the transformations that represent the impact of crosscutting features by modifying the common transformations before they are executed by changing their input parameters or by modifying their event chains.

A practical adoption of the approach proposed in this paper could be significantly supported by providing directly reusable transformations, transformation templates (i.e., parameterized transformations), or even just transformation schemes or examples to be adapted manually to the application context.

Actual feature models are huge and therefore are more effectively presented by individual concepts [12]. In general, a concept is an understanding of a class or category of elements in a domain [11]. Syntactically, in feature modeling, the root node of a feature diagram represents a concept [17]. Thus, raising a feature to the level of a concept is a matter of choice. Of course, this has to reflect the needs and objectives of the particular case of modeling. Having a feature model decomposed into a set of feature diagrams, rather than a single tree, involves having references between the trees (i.e., concept references [11]). Exploring how this affects feature model driven generation of software artifacts represents a research challenge.

## Acknowledgments

The work reported here was supported by the Scientific Grant Agency of Slovak Republic (VEGA) under the grant No. VG 1/1221/12.

This contribution/publication is also a partial result of the Research & Development Operational Programme for the project Research of Methods for Acquisition, Analysis and Personalized Conveying of Information and Knowledge, ITMS 26240220039, co-funded by the ERDF.

## References

[1] K. C. Kang, S. G. Cohen, J. A. Hess, W. E. Novak, and A. S. Peterson, "Feature-oriented domain analysis (FODA): A feasibility study," Software Engineering Institute, Carnegie Mellon University, Pittsburgh, USA, Tech. Rep. CMU/SEI-90-TR-21, Nov. 1990.

[2] J. O. Coplien, *Multi-Paradigm Design for C++*. Addison-Wesley, 1999.

[3] S. Apel, C. Kastner, and C. Lengauer, "FEATUREHOUSE: Language-independent, automated software composition," in *2009 IEEE 31st International Conference on Software Engineering, ICSE 2009*. Vancouver, BC, Canada: IEEE, May 2009. doi: 10.1109/ICSE.2009.5070523 pp. 221–231.

[4] K. Czarnecki and M. Antkiewicz, "Mapping features to models: A template approach based on superimposed variants," in *Proceedings of 4th International Conference on Generative Programming and Component Engineering, GPCE 2005*, ser. LNCS 3676, R. Glück and M. R. Lowry, Eds. Tallinn, Estonia: Springer, Oct. 2005. doi: 10.1007/11561347_28 pp. 422–437.

[5] K. Czarnecki, S. Helsen, and U. Eisenecker, "Staged configuration through specialization and multi-level configuration of feature models," *Software Process: Improvement and Practice*, vol. 10, pp. 143–169, Apr./Jun. 2005.

[6] Software Product Line Group, Programming Group, Univeristät Passau, "FeatureHouse: Language-independent, automated software composition," http://www.infosun.fim.uni-passau.de/spl/apel/fh/.

[7] pure-systems GmbH, "pure::variants: Variant management," http://www.pure-systems.com/pure˙variants.49.0.html.

[8] pure systems, "pure::variants user guide," 2015, http://www.pure-systems.com/fileadmin/downloads/pure-variants/doc/pv-user-manual.pdf.

[9] M. Shaw, "What can we specify? issues in the domains of software specification," in *Proceedings of 3rd International Workshop on Software Specification and Design*. IEEE CS, 1985, pp. 214–215.

[10] P. Sochos, M. Riebisch, and I. Philippow, "The feature-architecture mapping (FArM) method for feature-oriented development of software product lines," in *13th Annual IEEE International Symposium and Workshop on Engineering of Computer Based Systems, 2006, ECBS 2006*. Potsdam, Germany: IEEE, 2006. doi: 10.1109/ECBS.2006.69 pp. 308–318.

[11] V. Vranić, "Reconciling feature modeling: A feature modeling metamodel," in *Proceedings of 5th Annual International Conference on Object-Oriented and Internet-Based Technologies, Concepts, and Applications for a Networked World, Net.ObjectDays 2004*, ser. LNCS 3263, M. Weske and P. Liggsmeyer, Eds. Erfurt, Germany: Springer, Sep. 2004. doi: 10.1007/978-3-540-30196-7_10 pp. 122–137.

[12] ——, "Multi-paradigm design with feature modeling," *Computer Science and Information Systems Journal (ComSIS)*, vol. 2, no. 1, pp. 79–102, Jun. 2005.

[13] V. Chakravarthy, J. Regehr, and E. Eide, "Edicts: Implementing features with flexible binding times," in *Proceedings of 7th International Conference on Aspect-Oriented Software Development, AOSD '08*. Brussels, Belgium: ACM, 2008, pp. 108–119.

[14] V. Vranić and M. Šípka, "Binding time based concept instantiation in feature modeling," in *Proceedings of 9th International Conference on Software Reuse, ICSR 2006*, ser. LNCS 4039, M. Morisio, Ed. Turin, Italy: Springer, Jun. 2006. doi: 10.1007/11763864_34 pp. 407–410.

[15] V. Tzeremes and H. Gomaa, "A software product line approach for end user development of smart spaces," in *Proceedings of 5th International Workshop on Product LinE Approaches in Software Engineering, PLEASE 2015*. IEEE, 2015. doi: 10.1109/PLEASE.2015.14 pp. 23–26.

[16] K. Czarnecki and M. Antkiewicz, "Mapping features to models: A template approach based on superimposed variants," in *Proceedings of 4th International Conference on Generative Programming and Component Engineering, GPCE 2005*, ser. LNCS 3676, 2005. doi: 10.1007/11561347_28 pp. 422–437.

[17] K. Czarnecki and U. Eisenecker, *Generative Programming: Methods, Tools, and Applications*. Addison-Wesley, 2000.

[18] M. Nosáľ and J. Porubän, "Supporting multiple configuration sources using abstraction," *Central European Journal of Computer Science*, vol. 2, no. 3, pp. 283–299, 2012. doi: 10.2478/s13537-012-0015-7

[19] M. Nosáľ, J. Porubän, and M. Nosáľ, "Concern-oriented source code projections," in *Proceedings of 2013 Federated Conference on Computer Science and Information Systems, FedCSIS 2013*. Kraków, Poland: IEEE, 2013, pp. 1541–1544.

[20] J. Porubän and M. Nosáľ, "Leveraging program comprehension with concern-oriented source code projections," in *Proceedings of Slate'14, 3rd Symposium on Languages, Applications and Technologies*, Bragança, Portugal, 2014. doi: 10.4230/OASIcs.SLATE.2014.35 pp. 35–50.

[21] I. Polášek, P. Líška, J. Kelemen, and J. Lang, "On extended similarity scoring and bit-vector algorithms for design smell detection," in *Proceedings of 2012 IEEE 16th International Conference on Intelligent Engineering Systems, INES 2012*. Lisbon, Portugal: IEEE, 2012. doi: 10.1109/INES.2012.6249814 pp. 115–120.

[22] K. Czarnecki, S. Helsen, and U. Eisenecker, "Formalizing cardinality-based feature models and their specialization," *Software Process: Improvement and Practice*, vol. 10, no. 1, pp. 7–29, 2005. doi: 10.1002/spip.213

[23] J. Kollár, I. Halupka, S. Chodarev, and E. Pietriková, "pLERO: Language for grammar refactoring patterns," in *Proceedings of 2013 Federated Conference on Computer Science and Information Systems, FedCSIS 2013*. Kraków, Poland: IEEE, 2013, pp. 1491–1498.

# Education, Curricula & Research Methods

ECRM is a FedCSIS conference area aiming at interchange of information, ideas, new viewpoints and research undertakings related to university education and curricula as well as recommended methods of doing research in all computing disciplines, i.e. computer science, computer engineering, software engineering, information technology, and information systems. This area spans typical FedCSIS events (conferences, workshops, etc.) with rigorous paper submissions and review processes as well as panels, PhD and research consortia, summer schools, etc. Events that constitute ECRM are:

- DS-RAIT'15 – 2$^{nd}$ Doctoral Symposium on Recent Advances in Information Technology

# 2$^{nd}$ Doctoral Symposium on Recent Advances in Information Technology

THE second international Doctoral Symposium on Recent Advances in Information Technology (DS-RAIT 2015) will be held in Lodz (Poland) on September 13-16, 2015 as a satellite event of the *Federated Conference on Computer Science and Information Systems* (FedCSIS 2015) and *Education, Curricula & Research Methods* (ECRM 2015) conference.

The aim of this meeting is to provide a platform for exchange of ideas between early-stage researchers, in Computer Science, PhD students in particular. Furthermore, the symposium will provide all participants an opportunity to get feedback on their studies from experienced members of the IT research community invited to chair all DS-RAIT thematic sessions. Therefore, submission of research proposals with limited preliminary results is strongly encouraged.

Besides receiving specific advice for their contributions all participants will be invited to attend plenary lectures on conducting high-quality research studies, excellence in scientific writing and issues related to intellectual property in IT research. Authors of the two most outstanding submissions will have a possibility to present their papers in a form of short plenary lecture.

## Topics

DS-RAIT 2015 invites the submission of papers on all aspects of Information Technology including, but not limited to:

- Automatic Control and Robotics
- Bioinformatics
- Cloud, GPU and Parallel Computing
- Cognitive Science
- Computer Networks
- Computational Intelligence
- Cryptography
- Data Mining and Data Visualization
- Database Management Systems
- Expert Systems
- Image Processing and Computer Animation
- Information Theory
- Machine Learning
- Natural Language Processing
- Numerical Analysis
- Operating Systems
- Pattern Recognition
- Scientific Computing
- Software Engineering

## Event Chairs

**Gołuńska, Dominika,** Cracow University of Technology, Poland

**Kowalski, Piotr Andrzej,** Systems Research Institute, Polish Academy of Sciences; AGH University of Science and Technology, Poland

**Lukasik, Szymon,** Systems Research Institute, Polish Academy of Sciences, AGH University of Science and Technology, Poland

## Program Committee

**Arabas, Jaroslaw,** Warsaw University of Technology, Poland

**Atanassov, Krassimir T.,** Bulgarian Academy of Sciences, Bulgaria

**Balazs, Krisztian,** Budapest University of Technology and Economics, Hungary

**Bronselaer, Antoon**

**Castrillon-Santana,** Modesto, University of Las Palmas de Gran Canaria, Spain

**Charytanowicz, Malgorzata,** Catholic University of Lublin, Poland

**Corpetti, Thomas,** University of Rennes, France

**Courty, Nicolas,** University of Bretagne Sud, France

**De Tré, Guy**

**Fournier-Viger, Philippe,** University of Moncton, Canada

**Gil, David,** University of Alicante, Spain

**Herrera Viedma,** Enrique, University of Granada, Spain

**Hu, Bao-Gang,** Institute of Automation, Chinese Academy of Sciences, China

**Koczy, Laszlo,** Szechenyi Istvan University, Hungary

**Kokosinski, Zbigniew,** Cracow University of Technology, Poland

**Krawiec, Krzysztof,** Poznan University of Technology, Poland

**Kulczycki, Piotr,** Systems Research Institute, Polish Academy of Sciences, Poland

**Lilik, Ferenc,** Szechenyi Istvan University, Hungary

**Lovassy, Rita,** Obuda University, Hungary

**Mesiar, Radko,** Slovak University of Technology, Slovakia

**Noguera i Clofent, Carles,** Institute of Information Theory and Automation (UTIA), Academy of Sciences of the Czech Republic, Czech Republic

**Petrik, Milan,** Masaryk University, Czech Republic

**Sachenko, Anatoly,** Ternopil State Economic University, Ukraine

**Samotyj, Volodymyr,** Lviv Polytechnic National University

**Szafran, Bartlomiej,** Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Poland

**Tormasi, Alex,** Szechenyi Istvan University, Hungary

**Wei, Wei,** School of Computer science and engineering, Xi'an University of Technology, China

**Wysocki, Marian,** Rzeszow University of Technology, Poland

**Yang, Yujiu,** Tsinghua University, China

**Zadrozny, Slawomir,** Systems Research Institute, Poland

**Zajac, Mieczyslaw,** Cracow University of Technology, Poland

DOI: 10.15439/2015F244
ACSIS, Vol. 5

# Research Proposal: Barriers to new user and new domain adoption of the XSEDE Cyberinfrastructure

Richard Knepper
School of Informatics and Computing / Pervasive Technology Institute
Indiana University
2709 E. 10th St, Bloomington, Indiana, USA
Email: rich@iu.edu

***Abstract*—This research proposal proposes the examination of user attitudes about the Extreme Science and Engineering Discovery Environment (XSEDE). The XSEDE project supports basic research with a common system for making use of national cyberinfrastructure. The systems and infrastructure that make XSEDE useful for researchers are part of an actor network: these systems are socially constructed and they play their own part in the work of XSEDE, and in turn have an effect on the progress of basic research. I have completed previous work on the user relationships in the predecessor to XSEDE, the TeraGrid, and currently carry out participant observation with the management groups of the XSEDE project. By understanding the barriers to adoption of XSEDE by new researchers and new scientific domains, I hope to explore the linkage between resources (in this case computational resources) and scientific outputs.**

## I. Introduction

STARTING in 2001, the United States National Science Foundation (NSF) funded a distributed, high-performance computing project known as the TeraGrid, with the goal of supporting scientific research at the frontiers of current research. With over $430M invested in the TeraGrid grants to date and more projects open for solicitation, the TeraGrid has created an ecosystem of interacting researchers, technologists, and administrators [1]. The TeraGrid was superseded in 2011 by XSEDE, which continues the implementation of the architecture started in the TeraGrid project with additional work towards making these large scale systems more flexible, usable, and open to a broader range of research agendas [2]. XSEDE has been awarded over $64M to date for its management structure, with sites deploying resources based on NSF awards of as much as $77M for the National Institute for Computational Science's Kraken system, and $55M for the Texas Advanced Computing Center's Stampede system, to name the largest investments. The creation of national cyberinfrastructure in support of basic research is a resource-intensive effort: not only do the initiatives involved in XSEDE require substantial material investment, the centers that support these systems employ large numbers of staff to manage the systems, provide user support, adapt code to ever-larger systems and broader architectures, and reach out to the scientific community in order to bring in new researchers.

### A. XSEDE and outreach to new domains and users

The NSF mandates that the XSEDE project reach out to new disciplines and new institutions, including Minority Serving Institutions and Historically Black Colleges and Universities. Despite efforts to broaden the number of disciplines and institutions served in XSEDE, surveys of researchers indicate that only 30% of respondents feel that there are sufficient cyberinfrastructure resources available to meet their research needs. XSEDE's quarterly resource allocations process is oversubscribed by roughly 3 times. In order for XSEDE to meet NSF's recommendations to broaden its user-base, XSEDE needs to change the strategies used to recruit and develop users. Despite this charge from the NSF, non-traditional users of XSEDE still experience difficulties adapting and transitioning to the XSEDE project.

It is critical in the estimation of the NSF to provide a broad-use cyberinfrastructure framework for the support of basic science, and this has been emphasized in reports from the XSEDE project[3], as well as prominent researchers such as Richard Tapia and others. XSEDE includes in its management structure units for reaching out to new communities and has an emphasis on highlighting novel uses of XSEDE, but the current user base and user profile remains quite narrow. New systems on XSEDE advertise "Computing for the 99%", but this paradigm has yet to be fleshed out, not to mention actual usage information.

### B. Motivation and research goals

This investment in infrastructure for research represents considerable outlay for the government in real terms. While not large compared to other government expenditures, these services provide the long-term innovative capacity for the country, and they influence the nation's competitiveness and future technological development. XSEDE represents a system in which researchers have access to considerable resources, but only by adopting technologies as provided, and with access to resources governed by a peer-reviewed allocations process. In some ways this is like other resource-distribution mechanisms in science such as publications and grants, but the technological requirements to use and specificity of available

978-83-60810-66-8/$25.00©2015, IEEE

analyses on XSEDE resources means that different groups are more able to take advantage of resources.

The basic research questions in this study surround the impact that resources make on scientific work. XSEDE is a project intended to provide computational resources to scientists, no matter what their local resources may be. This research project is intended to reveal what these barriers to utilization of XSEDE resources are from the users' point of view, and to examine what these potential users of XSEDE say about their experiences, but also to analyze and understand the training and documentation materials provided for potential users, as well as observe some of the activities and NSF strategies for user recruitment. These findings will be examined along with the NSF's initiatives to generate user participation, with the goal of creating requirements for the next phase of the XSEDE project that can be incorporated into XSEDE's architectural processes: use cases and quality attributes. The products of architectural processes will include programmatic interfaces, training and documentation materials, and broader educational materials.

By understanding how researchers use resources, and how they compete for them, I intend to get more information about appropriate cyberinfrastructure and allocations methods to best serve the needs of the nation for basic research. By better understanding the users of XSEDE and their needs and relationships within XSEDE, successor infrastructures to this project can be built in order to do a better job supporting scientific research. My own previous research has focused on the management of XSEDE and relationships between supercomputing centers, and to understand the functioning of the project it is imperative to also see the project from the perspective of its client-constituents. By understanding linkages between resources and scientific outputs, it is possible to understand what kind of information is needed to drive science policy decisions as governments attempt to manage limited resources and still remain competitive.

## II. Literature on resources and performance

In order to explore the connection between resources available to scientists and their productivity, it is important to understand the more general background of resources for scientific activities. Most commonly these are competitively awarded grants that provide scientists with the financial means to accomplish research goals.

While it may be argued that funding for cyberinfrastructure for research is qualitatively different from research practices, I draw upon the stream of research that examines performance management for research, noting a few major correspondences between research infrastructure and research itself. First and foremost, the large scale cyberinfrastructure that characterizes the TeraGrid and XSEDE projects is in itself treated as a computer science research project. Creating software that makes use of large scale systems efficiently and effectively, providing services that allow users to make use of cyberinfrastructure with ease, and creating links between large scale systems that allow them to be used in concert with each other easily constitutes as significant challenge for computer science researchers. Secondly, the practice of providing research cyberinfrastructure for research is largely modeled after the process for obtaining grants for research itself. Proposals are created with broader impacts and scientific merits in mind, subjected to peer review, and projects are evaluated based upon the publications and reports they generate, as well services provided. Finally, there is limited evidence to suggest that infrastructure in support of research behaves and can be treated similarly to other infrastructure provided by other government agencies, thus our understanding of research funding is the closest analogue to assist in our understanding of dynamics driving cyberinfrastructure funding.

Performance management for grants is perhaps the most important difference between grant funded work and contracting work, and monitoring of scientific progress has singular difficulties. Partha and David [4] catalog the difficulties of economic evaluation of research: economic returns may come quickly or may take decades to realize, rights to intellectual properties are difficult to extract economic rents from (in fact restricting access to research may hamper further returns on initial investments), fundamental research progress may have dramatic and far-reaching impacts that are difficult to capture, and it is especially hard to forecast the success of any one particular research project. Hanson's [5] appraisal of focusing on effort as opposed to results has marked the transition from measuring research outputs towards measuring research processes. The standard operating procedure for evaluating both inputs (proposals) and outputs (scientific work) of grant-funded projects remains the peer review process. Garcia and Sanz-Menendez [6] discuss the context of peer review as metric of scientific research quite fully in their evaluation of competition in research initiatives. The authors begin tracing the path of peer review with the assertion that individual reputation and credit are central to the creation of the social structure of science, and that recognition by ones' peers is the foundation of legitimacy and leadership in a given field. Garcia and Sanz-Menendez note that the measurement of scientific production has long been based in volume and quality of scientific publications, but that these metrics cannot be separated from peer review. Peer review, despite some of its flaws outlined below, is not only the basic mechanism for ensuring quality of research, but also a critical factor in monitoring the efficiency of government investment in science. Peer review provides legitimacy to governmental bodies, and scientific work which has passed peer review has greater esteem in its scientific surroundings [6]. However, with the advent of new initiatives in government for assessing and monitoring of performance, peer review has had mixed fortunes as an evaluative tool for officials in charge of awarding research grants.

Shapira and Kulhman [7] describe the growth in requirements evaluation of research projects as governments attempt to control costs and derive better benefits from programs, noting that there are significant issues to measuring performance in this area. Impacts of these programs tend to be diffuse, as do costs, leading to difficulties in capturing all of the inputs and outputs. As research programs grow in complexity, including more disciplines and addressing broader problems, the evaluation of these programs must similarly become more complex. Government demands for continuous monitoring and program learning initiatives for research have led toward inclusion of subsidiarity, socioeconomic effects, and broader impacts into research evaluations. The increasing frequency of public-private partnerships for research also increases the complexity of program evaluation [7]. Partha and David [4] describe the the new attitudes towards measurement of research projects as a new economics of science, in which the previous free-market scientific workplace, characterized by scientists competing in the peer review process in order to gain funds and recognition is supplanted by a more interventionist government hand that is in the process of turning science toward more applied tasks. Government demands for better program evaluation both in the US and in Europe, as well as budgetary constraints from the recent economic crisis, have resulted in a call for more scrupulous examination of research performance.

The response to this call for increased evaluation and measurement of performance, has been variable at best. Cozzens [8], providing the context of evaluation in US research funding, describes the clash between the traditional evaluation tools for research, peer review and the journal selection process, and the new requirements based on the Government Performance and Results Act (GPRA) and increased requirements for management performance from the OMB. Peer review as the status quo for evaluation of science works in what Cozzens describes as an "autonomy-for-prosperity" model. Agencies support research activities in order to solve specific problems in an indeterminate amount of time, with limited oversight from Congress or agencies. Emphasis in evaluation is placed on the input end of the process, based on the quality and relevance of research proposals, and most importantly the accountability of this evaluation is placed on the research community, who is responsible for fairly making decisions, rather than on the researchers themselves to produce results to the general public. Guston [9] notes that peer review makes up a substantial amount of the selection process for research: \$37.7 billion or 86% of the reported total funding for research is merit reviewed. Applied research agencies, in contrast, have review processes based in personnel evaluation and budgeting that determine quality, although Cozzens, Bozeman, and Brown [10] note that there is a shift towards the competitive model of peer review even for these agencies. Peer reviewed grants are a feature of new federal research funding programs in the Department of Agriculture, the Environmental Protection Agency, and the Advanced Technology Program [9]. Peer review for research projects can happen both prospectively in the proposal selection process as well as retrospectively in evaluation [11], and have also been used as inputs in evaluating information for drafting regulation, creating state policies, and in evaluating courtroom decisions [9].

The peer review process conflicts directly with GPRA requirements for monitoring outputs of research, which explicitly focus on planning and achieving strategic objectives, rather than a culture of fairness in evaluation of proposals. As Cozzens [8] states, this "clashes with the traditional notion that the benefits that flow from research cannot be predicted in timing or content, but rather are visible only retrospectively". Response to the new evaluation requirements has been met by providing measures that are generic and qualitative: outcomes such as "advances in knowledge" or "ideas, peoples, and tools", or the NSF's frequently sought-after "science nugget", used to provide Congress with information about program success in a brief, easily-digestible package. As a result, such weak measures of evaluation lead to reinforcement of existing political forces, especially when coupled with another popular new metric of stakeholder input, which gives greater voice to those parties already engaged in the selection process [8]. Another approach to evaluation is to provide broad indicators of research progress: publications, funded research, and patents. Campbell [12] directly contrasts the peer review and indicator approaches finding that peer review results in complex but subjective evaluations of research work, while indicators are objective and easily quantified, but tend to be superficial in nature. Hagstrom [13] notes that peer reviewers frequently are able to identify the authors they are reviewing, or at least make educated guesses based on prior research and citation patterns. There is some evidence that researchers understand the peer review process and anticipate elements of it when drafting proposals. Knorr-Cetina [14] found in comparing proposals submitted for peer review to those without peer review that the style of the proposals changed rather than the content of the science inside. Furthermore, peer review is frequently conducted by established researchers, which leads to a problem in the assessment of new and innovative research directions, and relationships of mutual dependency that create self-reinforcing factions within scientific communities [12]. The world of the peer- reviewed scientist may be viewed as one mired in competition with other researchers first to get research proposals approved in order to get funding, and then to get the results of that research published.

Competition is in many ways the coordinating feature of scientific progress just as it is in economic activities. Hagstrom [13] describes the competition that takes place between scientists as specifically occurring when a scientist finds that her research in a particular area, on a problem not previously published, has been beaten to publication by another researcher. This form of competition may be extended to include being passed over in favor of another researcher in the grant selection process. Latour and Woolgar [15] established research funding as a vital part of researchers' credibility and reputation with other scientists. Garcia and Sanz-Menendez [6] sum the idea of competition up well: "Thus, competition for funds is an essential mechanism in the cognitive functioning of research, articulated in the credibility cycle, and a vehicle for relationships between science and government". Competition between researchers has a number of valuable features that aid scientific development. Competitive publication practices mean that additional researchers may be working on the same problems, which ensures an abundant supply of possible investigatory techniques and results. Competition drives hard work on the part of the competitors to outdo each other. Finally, competition reduces the risk of dilatory publication, and it encourages differentiation and innovation as scientists attempt to identify new problems to explore [13]. While competition should promote the best quality research, issues have been identified with competitive processes for publication and funding that may slow the progress of science. Competition thus has a complex relationship with peer review. Laudel [16] notes that the competitive process may have impacts to the course of science as scientist averse to risk select other research topics in order to avoid competition and increase favor in peer review, promote mainstream or existing research techniques in order to be more competitive with particular review boards. Reports from leaders in grant-funded research centers find that competitive resubmissions for funding has a disruptive effect on getting the work of the center done [17].

With the understanding that research is driven by competitive processes, as scientists attempt to establish a track record that allows them to build capital to drive further research, and reputation that allows them to secure that research, a number of questions arise prompting further investigation into the workings of grant funded science. Firstly, what role does scale play in the competitive process? A number of authors have investigated the shift from individually-centered projects to larger labs and research centers. If scientists are engaged in competition for resources and prestige on the individual level, what forces do these interactions exert on the organizations that they work within? Does individual competition influence the structure and performance of these organizations? Secondly, what is the interplay between competition and collaboration in scientific projects? Federal agencies in the US and abroad have invested in funding projects to take on "Grand Challenge" projects such as the Large Hadron Collider, the caBIG project for cancer research, and the Laser Interferometer Gravitational-Wave Observatory (LIGO), which incorporate researchers from multiple institutions and are multidisciplinary investigations of research questions. Collaboration is an essential element of these large-scale scientific projects. However, Edwards notes that collaborative work is frequently hindered by "friction" of various types, categorizing "data friction" as issues that keep researchers from being able to easily exchange and manipulate data and "computational" friction which keeps scientists from easily making use of supercomputers or transferring work between different supercomputing sites [18], and later positing a "collaboration friction" which is the effort required for scientists in multiple disciplines and or differing backgrounds to work together in order to achieve collaborative success [19].

Finally, how can government agencies structure their programs and investments to take advantage of these factors? The NSF and to some lesser extent, the NIH, are under considerable pressure to reduce funding outlays and to provide rigorous performance management indicators for the scientific research done under their auspices. Are there features of collaborative and competitive research that allow the grantors to get better results and better science faster?

## III. Previous research efforts

As a student of social informatics and of public management, with an interest in science policy, I have followed the development of the Supercomputing Centers program through the TeraGrid project and the XSEDE project. As an administrative manager in Indiana University's Pervasive Technology Institute, my duties include a management role within the XSEDE project, and this combination of theoretical, empirical research and access to the internal workings of the organization represent an opportunity for participatory research that is an extremely compelling case. In my efforts to get a better understanding of the TeraGrid user community, I conducted a social network analysis of the TeraGrid users and project allocations [20], noting the prevalence of traditional "hard science" disciplines, but also finding evidence of collaborations that span multiple fields of science. Since the end of the TeraGrid project and the advent of the XSEDE project, I have engaged in a case study [21], which characterizes the XSEDE project and similar advanced cyberinfrastructures as "living infrastructure" with qualitatively different attributes to traditional scientific instruments. These cyberinfrastructures have their own intents and goals that may or may not be aligned with their users' and the relationship between large cyberinfrastructure organizations and users becomes a collaborative arrangement in its own right.

Currently I am also working on a project with Katy Börner in the Department of Library and Information Sciences at Indiana University to characterize resource utilization on the XSEDE project as it relates to publication, in an effort to understand a general utilization-to-publication ratio, but also to examine whether different supercomputing centers and systems represent different research output profiles as well as the possibility of different fields of science having different resource/output functions.

## IV. Hypotheses

I hypothesize that there will be a number of barriers that are already well-described in documents such as the NSF's Advisory Committee on Cyberinfrastructure's Campus Bridging Task Force report, such as: difficulties in moving large amounts of data from its place of collection to computational systems for analysis difficulties in getting analyses to run on different resources unfamiliarity with national cyberinfrastructure frameworks and utilities I also hypothesize that in discussion with users, it will become evident that there is a lack of fit between the resources available via XSEDE and the new domains of interest. The XSEDE project and its predecessor, the TeraGrid, are rooted in the highly computational "hard science" disciplines such as high-energy physics, astronomy, and molecular dynamics. The new domains of interest to the NSF are slowly building computational emphases, but these techniques are often reliant on significantly large amounts of data, and analysis tools are developed on personal computers, rather than the codes developed in the era of centralized computing facilities. These analyses include ”Big Data” techniques–computational social sciences, for example–but also a significant number of life science analyses.

## V. Methods of analysis

A number of methods appear to make sense for investigating user benefits and difficulties taking advantage of XSEDE. XSEDE management already commits significant effort to improving practices within the project, from an in-depth architectural process to external evaluators who conduct general surveys of users and employees, as well as interviews in order to get more in-depth understanding, and case studies of individual initiatives within XSEDE. Leveraging the work of the external evaluation team can provide a considerable amount of information and indicate areas that require further inquiry.

### A. Document Analysis

In order to understand the current underlying body of knowledge, I propose to examine existing reports carried out by the external evaluation team, which has produced from 12-20 reports each project year [3]. This document analysis will be focused on user-centered evaluations (items such as the staff climate report will be excluded). Specific areas of interest will be the identification of mechanisms that support successful research activities as well as resource allocation schemes that have either facilitated projects or made them more difficult to carry out.

### B. Survey and Interview Activities

With some initial understanding of the existing data collected on XSEDE user activities, making use of the materials already present, I intend to conduct a survey of users with specific focus on sharing and utilization of resources, in order to understand how resources are obtained and allocated within projects. In order to understand the issues of new domains and new user populations–which may not easily conform to XSEDE’s practices for obtaining access to and using resources, I will elicit proposals from the XSEDE allocations committee which were not approved, looking for proposals which were not turned down for lack of preparation or scientific merit, but for lack of fit to the project or for other issues. I will interview these researchers in order to understand how they planned to make use of XSEDE and what issues they felt kept their proposal from receiving an allocation of resources. These interviews will be conducted by teleconference or videoconference by one to one arrangement.

I also plan to conduct close interviews with members of the XSEDE “Campus Champion” program, who are local users who volunteer to assist others in making use of XSEDE by training and facilitating research activities. These individuals frequently help new users address problems getting started conducting analyses on XSEDE, and they often have some research component of their own to pursue.

### C. Participant Observation

In concert with the survey and interview activities described above, I intend to participate in training activities for new users, webinars and targeted outreach events that address user needs, and to engage in other activities, in hopes of observing users in action as they learn, examine, test, and adapt to the XSEDE environment. Events for participant observation of users will include the upcoming Linux Clusters Institute Workshop[1], the XSEDE annual conference meeting, and online training and webinar events held by XSEDE, Ohio Supercomputing Center’s HPC University, and Cornell University’s Virtual Workshop program.

## VI. Potential Benefits of Research

The current climate for science policy in the United States, but also in other countries is a problematic one. Support for basic research is being eroded: not long ago it was common to hear the phrase “flat is the new doubling” when speaking to research center leadership, this has changed to “minus five percent is the new flat”. If the US and other nations are to ensure competitiveness and continued innovation, judicious use of resources and a strong understanding of the relationship between resources supporting research and scientific outputs is required. Scientific productivity is a problematic area to measure and there will be many different pieces of research that inform a larger stream, these will include studies of scientometrics and bibliometrics, but there is also significant work on large technical systems [22], [23] and the social element of technological systems [24], [25] that will contribute to our understanding of the social elements of scientific research.

This study will benefit our understanding of the linkage between access to resources and scientific outputs. In order for organizations such as the NSF to successfully support basic research, a better understanding of the effect of resources on research outcomes is required, especially in the case of shared resources such as the XSEDE project, which provides a long-lived scientific cyberinfrastructure service. XSEDE and its service provider units represent significant investment on the part of the National Science Foundation and making sure that the most effective distribution of resources can be made is critical to the longevity of the NSF and its mission. In a broader sense, understanding how centralized resources for research are best implemented in a general sense will provide better strategies for makers of science policy to present effective and equitable distribution of resources that ensures scientific progress and competitiveness in general.

## References

[1] NSF, “Nsf 07-28 cyberinfrastructure vision for 21st century discovery,” 2007.

[2] J. Towns, “Xsede: extreme science and engineering discovery environment,” 2011.

[3] N. S. F. OCI-1053575, “Xsede program years 1-3 comprehensive report,” tech. rep., XSEDE, 2014.

[1] http://linuxclustersinstitute.org

[4] D. Partha and P. A. David, “Toward a new economics of science,” *Research Policy,* vol. 23, no. 5, pp. 487 – 521, 1994. Special Issue in Honor of Nathan Rosenberg

[5] R. Hanson, “Patterns of patronage: Why grants won over prizes in science,” University of California, Berkeley, p. 11, 1998.

[6] C. Garcia and L. Sanz-Menendez, “Competition for funding as an indicator of research competitiveness,” *Scientometrics,* vol. 64, no. 3, pp. 271–300, 2005.

[7] P. Shapira and S. Kuhlman, eds., *Learning from Science and Technology Policy Evaluation.* Cheltenham, UK: Edward Elgar, 2003.

[8] S. Cozzens, *Frameworks for Evaluating S&T Policy in the United States,* ch. 4, pp. 54–64. Edward Elgar, 2003.

[9] D. Guston, *The Expanding Role of Peer Review Processes in the United States,* ch. 6, pp. 81–97. Edward Elgar, 2003.

[10] B. B.-B. E. Cozzens, S., “Measuring and ensuring excellence in government laboratories: Practices in the united states,” tech. rep., *Canadian Council of Science and Technology Advisors,* 2001.

[11] N. Science and T. C. (NSTC), “Assessing fundamental research,” 1996.

[12] D. F. Campbell, *The Evaluation of University Research in the United Kingdom and the Netherlands, Germany and Austria,* ch. 7, pp. 98–131. Edward Elgar, 2003.

[13] W. O. Hagstrom, “Competition in science,” *American Sociological Review,* vol. 39, no. 1, pp. pp. 1–18, 1974.

[14] K. Knorr-Cetina, *The Manufacture of Knowledge: an Essay on the Constructivist and Contextual Nature of Science.* Oxford: Pergamon Press, 1981.

[15] B. Latour and S. Woolgar, *Laboratory Life. The Construction of Scientific Facts.* London: Sage Publications, Ltd., 1979.

[16] G. Laudel, “The art of getting funded: how scientists adapt to their funding conditions.,” *Science and Public Policy,* vol. 33, pp. 489–504, August 2006.

[17] L. Smarr, “Hpcwire: The good, the bad and the ugly: Reflections on the nsf supercomputer center program.,” 2010.

[18] P. N. Edwards, *A vast machine: Computer models, climate data, and the politics of global warming.* MIT Press, 2010.

[19] P. Edwards, M. S. Mayernik, A. Batcheller, G. Bowker, and C. Borgman, “Science friction: Data, metadata, and collaboration,” *Social Studies of Science,* p. 0306312711413314, 2011.

[20] R. Knepper, “The shape of the teragrid: analysis of teragrid users and projects as an affiliation network,” in *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery,* p. 54, ACM, 2011.

[21] R. Knepper, “The xsede project: A living cyberinfrastructure.” In Press, 2014.

[22] T. P. Hughes, *Rescuing Prometheus: Four monumental projects that changed our world.* Vintage, 2011.

[23] T. P. Hughes, *Networks of power: electrification in Western society,* 1880-1930. JHU Press, 1993.

[24] W. Bijker, T. Hughes, and T. Pinch, eds., *The Social construction of technological systems: new directions in the sociology and history of technology.* MIT Press, 1987.

[25] D. Mackenzie and J. Wajcman, eds., *The social shaping of Technology.* Open University Press, 1999.

DOI: 10.15439/2015F135
ACSIS, Vol. 5

# Promoting the understanding of electronic components and circuit parameters by using didactic dynamic pictures - qualitative analysis of research results

Krzysztof Krupa
University of Rzeszow, 35-959
Rzeszow, al. Rejtana 16C, Poland
Email: kkrupa@ur.edu.pl

*Abstract* — **This article describes the qualitative analysis of the research results of the effectiveness of promoting an understanding of the parameters of electronic components and circuits by using dynamic circuit and block diagrams as well as the dynamic characteristics integrated with them. The introduction highlights the characteristic way of understanding the parameters of electronic components and systems and an outline of my research conducted in this area. The second chapter presents a typology of the dynamic pictures used and a description of sample animations. The third chapter contains the explication of the concepts of understanding electronic components and circuits. The fourth chapter presents the description of the research, while in chapter five the qualitative analysis of the research results is shown, from which the conclusions are contained in chapter six.**

## I. Introduction

IN MOST mechanical structures, moving parts play a major role. They are integral to sensory cognition. The phenomena taking place inside electronic structures are not available to sensory cognition, which is why images play a leading role in the teaching of electronics. Since these phenomena exhibit variations in time, the didactic dynamic picture, visualizing the time dependence, may affect the effectiveness of teaching electronics. Part of my research on the effectiveness of using dynamic pictures was spent teaching the meanings of the parameters of electronic components and circuits. They represent a part of the test results on the effectiveness of teaching using didactic dynamic pictures that were included in the dissertation [2].

## II. Didactic Dynamic Pictures In Shaping The Understanding Of Electronic Components And Circuit Parameters

Didactic dynamic pictures used in the teaching-learning of electronics were divided into four groups. They perform the following didactic functions: illustrating the operation of electronic components, functioning of electronic components in basic work systems and illustrate the construction and operation of electronic circuits. In order to visualize the electronic components, dynamic area models are utilized. The operations of electronic components in basic systems, can be presented by the analogy of hydro-mechanical and dynamic schematics. The functioning of electronic circuits can be presented by using dynamic circuit diagrams and composite structures using dynamic flowcharts. Dynamic characteristics may be used for a detailed visualization of the parameters [4].

In order to shape the understanding of electronic components and systems parameters, the most appropriate images are those under which direct visualization of parameters occurs. These include dynamic circuit diagrams and dynamic flowcharts and corresponding dynamic characteristics.

Figure 1 shows an application, visualizing the operation of a basic amplifying circuit, constructed using a bipolar transistor. The application consists of two modules. On the left is a dynamic scheme, on the right the dynamic characteristic, which was operably linked to the diagram. The value of the input voltage to the system can be changed by using a slider. According to the characteristics of the bipolar transistor, the input voltage's $U_{BE}$ changes are reflected in the value of the current flowing through the base of the transistor - $I_B$. According to the characteristics of transition, changes in the intensity of the base current of the transistor causes proportional changes in the current of the collector - $I_C$. Changes in the collector's current, which is also the current of the resistor R, causes changes in voltage across the resistor, working together with the transistor in the voltage divider, whereby it also changes the output voltage of the system.

The scheme uses a series of dynamic elements. For example, numeric values of parameters such as base-emitter voltage ($U_{BE}$), the base current ($I_B$), the collector current ($I_C$) and the voltage at the collector's resistor ($U_R$) and the output voltage ($U_{CE}$). The magnitudes of these parameters are further displayed, in graphic form, with the help of dynamic

978-83-60810-66-8/$25.00©2015, IEEE

arrows whose size varies depending on the value visualized by the arrow’s signal.

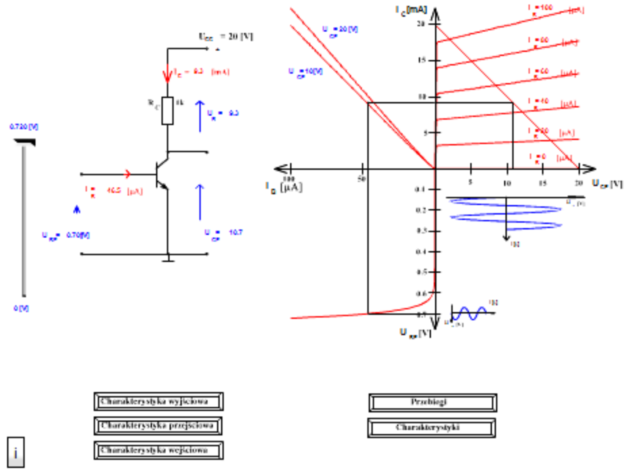


Figure 1. Didactic dynamic scheme with dynamic characteristics. The application shows the OE transistor amplifier [4].

Placed next to the schema, is a collection of dynamic characteristics of the static characteristics of the bipolar transistor in a common emitter’s circuit. Includes the input characteristics $I_B = f(U_{BE})$, transient characteristics $I_C = f(I_B)$ and the output characteristics $I_C = f(U_{CE})$. The output’s characteristics also draws the static characteristics of the resistor R. The main element in the dynamic characteristics are the lines indicating the individual parameters of the static characteristics. Furthermore, applied were dynamics, plotted out on the screen of timeline charts, of sinusoidal signals present at the input and the output of the system. All elements of the dynamic characteristics can be freely switched on and off using the appropriate buttons.

Figure 2 shows a different kind of teaching dynamic picture - a dynamic block diagram of an AC adapter for continuous operation. Despite the fact that all the blocks and timing diagrams of the system were presented, the application allows for their stepwise actuation by pointing the cursor over their names and blocks.

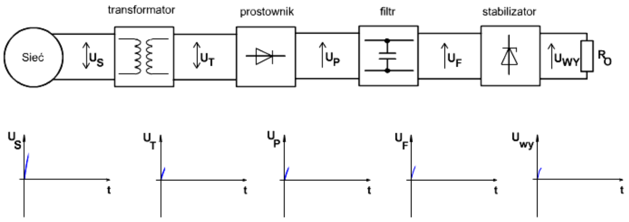


Figure 2 is a block diagram showing the dynamic power supply for continuous operation [4].

The application makes it feasible to present the essence of the shaping signals in particular functional blocks of the system. It is the basis from which to present, to the students, the construction and operation of consecutive blocks, such as the transformer, rectifying circuit, filter and stabilizing system.

The article shows only two examples of the teaching dynamic pictures. For empirical research purposes, much more of these aids were developed. These encompassed ten issues of analog electronics. The effectiveness of the application presented in the article has been confirmed [3]. Also undertaken were a series of works, aimed at creating a didactic dynamic picture visualization of digital systems. They have been applied in teaching materials used for teaching electronics over the Internet. These are applications that act as simulators of basic digital electronic systems.

## III. UNDERSTANDING THE PARAMETERS OF ELECTRONIC COMPONENTS AND CIRCUITS

This article assumes that understanding is the discovery of such content in objects, as their significance, meaning, structure, function, role, origin, durability, usefulness, look, what it does, how it arises, what it interacts with, the category to which it belongs and what specific characteristics does it possess [1]. One of the areas of understanding electronic engineering is to understand the parameters of electronic components and systems [3], which the student exhibits by their behavior relative to the described, in any way (image, symbol, text), element or electronic system [3]: assigns characteristic parameters to the element, describes the main characteristics of electronic components and systems, determines the unit of a given parameter, estimates the value of the given parameter, assigns characteristics to the stated element or electronic circuit, outlines the characteristic parameters of electronic components and systems, identifies, explains, indicates the essential points of the characteristics of electronic components and devices, explains the essence of the parameters of electronic components and devices, solves problems (including computing) using the parameters of electronic components and systems and selects a parameter or characteristic element or system to solve the design problem.

The set of these manifestations of understanding was used to construct the test tasks for the purpose of measuring the effectiveness of dynamic pictures in the development of understanding the parameters of electronic components and systems [3].

## IV. DESCRIPTION OF THE COURSE OF RESEARCH

Research of the efficiency of shaping the understanding of the parameters of electronic components and systems was carried out as part of a study involving the effectiveness of teaching electronics using didactic dynamic pictures on the example of the education process of students of technical and information technology. A didactic experiment, as the research method, was adapted and ran based on the parallel group technique. The study included students of technical-informatics education, educated under the three-year undergraduate studies and the five-year study mode.

The research used 120 didactic dynamic pictures presented to students of the experimental group consisting of ten lectures on analog electronics. The control group was given only static pictures were. After each lecture, students in both groups were subjected to a 20 task test. Four weeks after the last lecture, a test was given containing sixty tasks from ten sections of analog electronics.

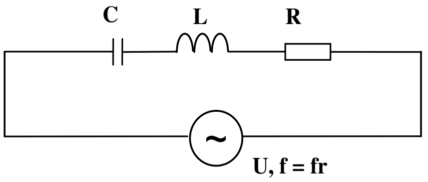
Quantitative analysis allowed for the confirmation of the hypothesis, in which it was assumed that the use of didactic dynamic pictures contributes to the efficiency of promoting the understanding of parameters of electronic components and systems. In order to fully present the results of the test, qualitative analysis was used [3].

## V.QUALITY ANALYSIS OF TEST RESULTS

Qualitative analysis allows for the identification of the kinds of difficulties faced by students in the experimental and the control groups while solving the tasks of understanding the parameters of electronic components and systems.

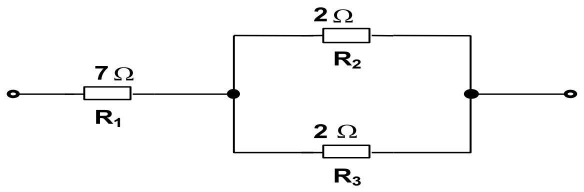
The first task was to indicate whether the impedance is at its lowest when the frequency of the current flowing through the serial RLC circuit is equal to the resonant frequency.

1. When the frequency of the current flowing through the system of equations is analyzed, resonant frequency of the circuit impedance is the lowest. true / false - underline the right answer.



RLC circuits in serial configurations, parallel and mixed are widely used in electronics, especially in the input circuit of the radio. In the case of this task, the experimental group received 83% correct responses relative to the control group that received a score of 41%.

2. Please calculate the equivalent resistance of the circuit. - Select the correct answer with a cross in the table.



| | | |
|---|---|---|
| 1 | 8 Ω | |
| 2 | 5 Ω | |
| 3 | 4 Ω | |
| 4 | 16 Ω | |

The solution to the 2nd task was to calculate the supplementary resistance of a circuit consisting of resistors connected in series and in parallel.

The experimental group received a score of 78% correct responses, and control group 48%. In the experimental group, 24% of students made the wrong choice as a result of not doing any calculations on the test sheet. In the control group, however, this situation occurred in 41%.

In view of the fact that didactic dynamic pictures do not support directly the skills of solving mathematical equations, it may indicate that there is another, deeper reason for the significant difference in the results of the groups at the expense of the control group. The didactic work with students, in the experimental group, noted that they often rely on lectures where dynamic hydraulic analogies are used. Facilitating the understanding of the presented phenomena, thus making it possible to avoid erroneous algorithms during problem solving, which were observed among those students who never dealt with analogies.

In a further, third task, a time constant of an RC circuit with specified parameters had to be calculated.

3. Calculate the time constant of the RC circuit with parameters: R = 100kΩ, C = 10ΩF.

The solution to this task depended on the reproduction of the correct mathematical expression, substitution of the given values into the formula, appropriate transformation of data prefixes and the performing of calculations. The experimental group received a score of 17% correct answers, while the control group 34%. The grading of the solution to the problem was based on the proper conversion of units.

In the experimental group, 73% of students did not attempted to solve this task, whereas in the control group, 62% reported omissions. This shows a complete inability to solve the task. A similar conclusion can be drawn on the basis of an incorrect formula, which was reported in 7% of the responses of the experimental group and in 3% of the responses of the control group of students.

The ability to apply the message in typical situations with regard to the understanding of the electronic parameters was measured with the use of task 4, in which the current-voltage characteristics of a semi-conductive diode should be drawn. On this characteristic, the forward voltage should also have been marked.

4. Please draw a semi-conductive diode's characteristics and highlight the forward voltage.

This task was solved correctly by 54% of the students in the experimental group and 34% of the control group. In the experimental group, a reported 12% had partial responses containing omissions in the description of the axis on the graph, and 17% had incorrect charts. In the control group, 31% of such completely incorrect responses were reported. In 21% of the tasks, of the control group, students did not attempt to find a solution. In the control group, in 31% of cases, the answer was completely random, for example, a student drew a sinusoidal waveform. In the experimental group, this phenomenon occurred in 17% of the cases. These experimental lectures exposed the dynamic characteristics of diodes, which illustrated the relevant parameters of their application in electronic circuits.

Obtaining a solution to task no. 5, it was crucial to indicate which of the expressions describes the amplification of current of a bipolar transistor.

5. Which of the expressions describes bipolar transistor current amplification of the OE circuit?................. – **check the right answer**

$A: \frac{I_C}{I_B};$ $B: \frac{U_{BE}}{I_B};$ $C: \frac{U_{BE}}{U_{CE}};$ $D: \frac{I_E}{I_B};$

The experimental group, for this task, received a score of 22% correct answers, while 48% was attained by the control group. Bipolar transistor current gain was mistakenly identified, by specifying the ratio of the emitter current to base current, 7% of the students in the experimental group and 31% of the control group. This demonstrates a better understanding of this parameter by persons participating in lectures, in which no didactic dynamic pictures were used. 46% of the responses in the experimental group and 17% in the control group indicated that the current gain is calculated by using the expression containing the voltage ratio value UBE to UCE. In this case, a separate dynamic picture was not applied, from which, in a direct manner, the issue of the current gain can be assimilated.

In task no. 6 it was necessary to describe the "minimum holding current". In the control group there was no single correct answer, while in the experimental group, the number of correct responses was 37% of the total number of results. This data suggest a beneficial impact of the experimental agent on the effectiveness of understanding the parameters.

6. Please describe the concept of "minimum holding current".

Majority of the students (69%) in the control group did not specify the electrodes of the thyristor, in which the parameter, included in the job, is expected to be. One student from the control group mistook the holding current with the conduction current (4%). The experimental group reported a 15% omissions of tasks, while 14% of the respondents in the control group did not attempt to solve the task.

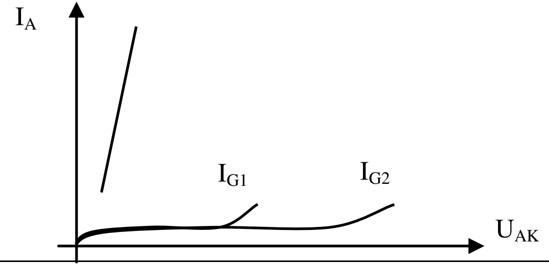
Further, task no. 7 referred to the skills used in typical situations in terms of understanding the parameters of semi-conductive switching elements. Solving this task required plotting out the current-voltage characteristics of the triac. 66% of the students in the experimental group and 38% of students in the control group solved the task correctly.

7. Please draw the characteristics of the triac. Pay attention to the description of the axis.

Basic mistakes committed during problem solving consisted of drawing unidirectional triode thyristor (9% in both groups) or bidirectional diode thyristor. Such errors were committed by 15% of the students in the experimental group and 24% in the control group. In the experimental and control group, there were no omissions of the task. A significant difference in both groups' output results can testify to the high efficacy of the use of didactic dynamic pictures in shaping the understanding of parameters.

Similarly, high results were obtained in the task 8, in which the person being tested needed to indicate a relationship between the parameters of $I_{G1}$ and $I_{G2}$ on the current-voltage characteristics of the thyristor (IA = f ($U_{AK}$)).

8. What is the relationship of the size of the current of the switching device's gate with characteristic as shown in the figure? IG1 ................... IG2- between these parameters,correctly enter the sign of the majority, minority or equality.



In this case, the experimental group received a score of 88% correct responses, while the control group 59%. The ability of answering this task indicates skills in identification, assessment and understanding of the parameters affecting the operation of the switching elements.

The task of no. 9 was to draw a phototransistor's output characteristics.

9. Please draw a phototransistor's output characteristics. Pay attention to the description of the axes and constant parameters.

The ability to draw the characteristics showed 27% of the students in the experimental group and 14% of students in the control group. The experimental group reported 58% omissions of the job or a completely wrong realization of it, while in the control group these solutions were 72%. In the experimental group, 24% of the tasks were incompletely resolved. Characteristics properly drawn, from a graphical point of view, were not described with proper parameters nor have the axis' signatures drawn into (IC = f (UCE), E = const).

The level of understanding a photoresistor's parameters was measured with the help of task 10. 39% of the students in the experimental group correctly answered the question, while the control group attained 62% correct answers. In this case, there was no positive impact of the didactic dynamic pictures on the level of to understanding the parameters.

10. Please describe the parameter of "dark resistance"

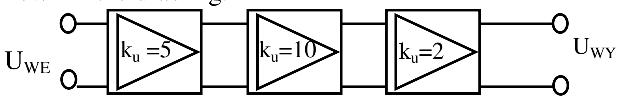
22% of students in the experimental group turned the process around by switching the cause of the phenomenon with its effect, for example, one of the responses reads: "it is a small light caused by high resistance". In the didactic dynamic picture, it was shown how a photoresistor works under conditions of complete darkness, however, the name of its resistance was not specified.

Task no.11 was dedicated to the understanding of the parameters, in which the essence of the squareness ratio needed to be explain. Here the experimental group received a result of 5%, while the control group got 10% correct answers.

11. Please explain the nature of the parameter: squareness ratio. In the explanation please use an appropriate figure and expression.

None of the dynamic pictures used in the study directly referred to this parameter, however, the issue of the steepness of the frequency characteristics of filtration systems and a tuned generator was presented with the help of the didactic dynamic pictures. However, this proved insufficient.

12. Please calculate the total voltage gain of the circuit shown in the drawing.



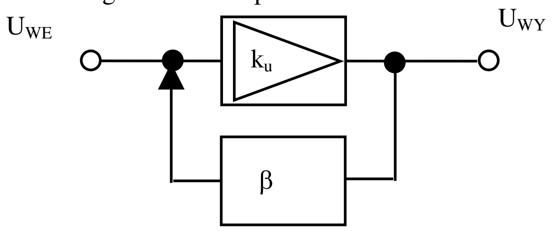
The voltage gain is:

The next task (12), relating to the understanding of the electronic circuit parameters, concerned the calculation of the voltage gain of a triple set of amplifying blocks, connected in series, with a given intensification.

In this task the experimental group received 49%, while the control group received 83% of correct answers. As is apparent from the number of correct answers, this task was relatively simple, however, in this case, no positive effect of the experimental agent was noted. Among the errors in solving this task, the most frequent were attempts the summation of individual gain values. This phenomenon occurred in 36% of the answers given by the students in the experimental group and 6% of students in the control group. The other students did not attempt to solve the task.

The next task (13) related to the ability of using the messages in problematic situations referring to the understanding of the electronic circuit parameters. In this task, the input voltage of the amplifying circuit, with feedback, had to be calculated, using the data contained in the problem.

13. Please calculate what the input voltage is given to the system if there is voltage at the output of 10V. The β value is -0.1 and the gain of the amplifier k = 100.



The experimental group received a score of 12%, while the control group achieved 3% of correct answers. The experimental as well as the control group reported the same number (24%), which acknowledges that the basic pattern of amplification of a circuit, with feedback, is well known to students and the result of this study is caused by the control group's student's inability to exploit it.

In task 14 the characteristic of a secured progressive AC adapter had to be draw. The experimental group reported 32%, while only 3% of correct answers in the control group.

14. Please draw the load carrying characteristics of a power supply with progressive protection.

In the experimental group's answers, the correct shape of characteristics were included, but did not adequately describe its axis. There were 9% of such responses. In the control group, 6% of the responses contained the mistake of shifting the graph's y-axis for the x-axis. This demonstrates the misconception of the nature of dependency on the phenomena presented with the help of the overload characteristics. It should be noted that the didactic dynamic picture was not used directly in illustrating the activities of progressive security. A higher percentage of accurate answers, in the experimental group, demonstrates the indirect effects of the didactic dynamic pictures.

The next task (15) related to the measurement of the ability to use the message in typical situations, within the meaning of the parameters of electronic systems.

15 Graph showing the following equation such as f=f($U_{CC}$) illustrates:

| | | |
|---|---|---|
| 1 | frequency response of the output signal when changing the amplitude | |
| 2 | frequency of the amplitude of the output signal when changing the supply voltage | |
| 3 | frequency response of the output signal when changing the supply voltage | |

It was necessary to explain the essence of the characteristics described for each model. The results obtained in this work are the following: the experimental group - 78% correct answers; control group - 90%. Distractor 2 was chosen by 12% of students in the experimental group and 10% of the students in the control group. Three people in the experimental group showed distractor 3 [3].

## III. Conclusion

Experimental studies have shown that the use of the didactic dynamic pictures has an important influence on the understanding of the parameters of electronic components and systems. Qualitative analysis of the results allow the following conclusions:

- students of the control group more often than the students of the experimental group did not attempt to solve the task,
- control group students more frequently than in the experimental group gave accidental or absurd answers,
- mistakes made by the students in the control group are more important than the mistakes made by the students in the experimental group.

Qualitative and quantitative analysis of the results of the studies on the effectiveness of using didactic dynamic images in the development of understanding the parameters of electronic components and systems confirm the hypothesis, in which the expected positive effect of the experimental factor.

## References

[1] E. Franus (2000), Wielkie funkcje technicznego intelektu. Kraków.

[2] K. Krupa (2013), Efektywność nauczania elektroniki z zastosowaniem dydaktycznych obrazów dynamicznych na przykładzie studentów kierunku edukacja techniczno-informatyczna, doctoral dissertation manuscript, APS, Warszawa.

[3] A. Marszałek (2001), Elektronika w edukacji technicznej dzieci i młodzieży. Rzeszów.

[4] K. Krupa (2011), Tworzenie dydaktycznych obrazów dynamicznych – przykłady realizacji struktur mechatronicznych in Wokół mechatroniki. red. W. Furmanek, L. Leniowska, Rzeszów.

DOI: 10.15439/2015F262
ACSIS, Vol. 5

# The column-oriented database partitioning optimization based on the natural computing algorithms

Artur Nowosielski[1]
[1]PhD Studies, Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: artnowo@gmail.com

Piotr A. Kowalski[2,3], Piotr Kulczycki[2,3]
[2]Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: {pakowal,kulczycki}@ibspan.waw.pl

[3]Division for Information Technology and Biometrics
Faculty of Physics and Applied Computer Science
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Cracow, Poland
Email: {pkowal,kulczycki}@agh.edu.pl

***Abstract*—This paper describes the basic components of a research project aimed at the application of natural computing metaheuristics to optimize the horizontal scaling of databases. Column oriented databases were selected for the project because of their unique properties. A mathematical model has been created in order to align the problem of horizontal scalability to the general optimization methods, such as natural computing algorithms.**

## I. Introduction

THIS article is an overview of research on column oriented databases (DB) partitioning optimization with the use of the natural computing algorithms. Column-oriented databases are believed to qualify for this purpose nicely thanks to their physical storage structure. Sometimes they are used as a NoSQL equivalent for relational database management systems because of their flexibility and partial similarity to relational model. In the class of the natural computation algorithms there are metaheuristic procedures that suite well to the problems of optimization with multiple constraints and multi-modal objective functions.

There are three main pillars of the research, described in the following subsections. Subsection I-B contains a general description of the column-oriented databases along with the most important relevant details. From the scalability options for databases, horizontal scalability gained some noticeable attention and major implementations, especially in a modern so-called “web 2.0” services. Subsection I-A describes basic features along with pros and cons of a database horizontal scaling. The natural computing algorithms described in the section I-C are an important branch of the research on computation models and methods. Their main goal is to implement heuristics inspired by the natural environment’s behaviors and processes, to solve optimization problems.

The next sections describe current state and results of the aforementioned research. They consists of the prototype implementation of a column-oriented DBMS along with its mathematical model. Selected natural computing algorithms were also implemented in order to discover and describe their features on the basis of the typical benchmarking optimization problems. Section III contains the description of the application of the algorithms to the optimization problem using a created mathematical model. Current short- and long-term plans for further research are described in order to familiarize the reader with the expectations of the research.

### A. Database partitioning

In general, two main classes of scalability solutions can be distinguished: vertical scaling and horizontal scaling. These approaches are slightly different. Vertical scalability is based on the assumption that, in order to increase system capacity, its resources should be enforced, e.g. by swapping the CPU to a faster one. In this approach, a unit of work remains assigned to only one processing unit and is therefore limited by the unit’s capabilities. Additionally, hardware capabilities are finite and it is possible to reach to the point in which it will not be possible or reasonable to deploy a more powerful hardware. Instead, this approach is easy to implement, because it preserves previously used computation model and processing methods. Especially, it does not involve any work division, which means no additional synchronisation issues are introduced. In the horizontal scalability (scaling out) approach, it is assumed the unit of work to be done exceeds the capacity of any single unit and needs to be divided. That slightly changes the processing model and introduces concurrency- and synchronisation-related issues, but in return, offers virtually infinite theoretical capacity.

Within the database domain, horizontal scalability can be split into the three classes: replication, sharding (also referred

978-83-60810-66-8/$25.00©2015, IEEE

to as a horizontal partitioning) and partitioning (also known as the vertical partitioning). Please note that the terms "vertical" and "horizontal" regarding partitioning and sharding are absolutely not related to the general scalability terms. Both are parts of the horizontal scalability class. That is, both involve some division of a data in a database into more than one database management system (DBMS) instance. Replication is out of the research scope, so it will not be described here. It has been covered widely, along with the remaining solutions in the master thesis [1].

Horizontal and vertical partitioning differ by a division plane. Horizontal partitioning duplicates database (or just single table) schema on many instances but splits the data between them. Please note that specific partitions can be distinctive (that is, every records belongs to one and only instance) as well as redundant. Partitioning function can take into consideration record's identifier, single other field or a set of fields. A value which is used to perform partitioning is so-called "partitioning key". Typical partitioning schemas include:

- by range (e.g. clients with last name starting with letters A, B, ..., I, J go to the instance#1, K, L, ..., Y, Z - #2);
- by modulo function (e.g. if record have natural numeric key they go to instance number (key % n) where n = number of instances);
- by list (if record has a field constrained by a finite set of values, instances can be assigned with its subsets; e.g. clients born on monday, tuesday and friday go to the first instance, others to the second one);
- by hash function (hash function result set must reflect instance set).

More sophisticated methods include combinations of previously enumerated ones and consistent hashing. Consistent hashing is a technique which optimizes a division for frequent remapping, reducing number of data to be migrated after every change.

Very useful concept here is the abstraction of hash function results from the physical arrangement of database engine instances. That requires additional mapping between these two tiers, but brings some significant advantages of a direct linkage. Primarily, changes in the arrangement of database servers does not involve modifications of the hash function. In other words, it satisfies well-known "single responsibility principle" coming from the software engineering domain.

Sharding can then be considered as a data division between duplicated schema instances. Vertical partitioning, in turn, can be thought as a schema division. It requires a deep analysis of the data usage and implicitly assumes that structure of every records is common if not exactly the same. Then rarely used parts of data are extracted and moved to a different instance(s) than the frequently used parts. That is not the only possible analysis schema though. It is also possible to highlight data parts which are frequently used together and divide a database by that. Obviously, in real world application both techniques can be joined and applied together in any arrangement, which leads to virtually infinite number of possible variants.

### B. Column-oriented databases

Column-oriented databases (sometimes referred to as a column-family databases, columnar databases or column stores) is a non-relational database data model. Although not strictly defined, this model brings some significant features increasing its horizontal scalability. The current section describes the model as it is understood within the conducted research. For the ease of understanding, analogies to the well known relational database model are highlighted. Nevertheless, it needs to be stated explicitly that both models are not related. Modern column-oriented data stores have been covered in [2]. Implementation details of the Apache Cassandra column-family store have been described in [3].

Fundamental terms and concepts are common to every column-oriented DBMS implementation. Keyspace is the basic storage and logical entity. It acts as a container for lower level entities and could be compared to database or schema in terms of relational databases. A column family is a named collection of records with similar or the same column set. Column families belong to the keyspaces.

The most important feature coming from the described concepts is so-called schemalessness, which means that column family's data structure is flexible. A column set for specific records within the same column family is variable. In real world solutions it is usually "defined" only by an established convention, not by a strict constraints stated during the database creation process. As opposed to the relational database tables, if a given record does not have (in a logical sense) value in given column, it does not store actual NULL value (in a storage sense). That feature brings a significant impact on the storage structure and application logic. Also, it allows to implement some use cases which are not effective or even possible in relational databases. A common use case is the time-series data store, which contains actually only one row of data, but with dozens of columns created every time a new value is saved. In that case, the value timestamp becomes a column name. Another typical appliance is a data store for a recommendation system. Such systems rely on huge matrices correlating all users with all items.

### C. Natural computing algorithms

Natural computing algorithms are a significant class of the heuristic procedures, which takes its inspiration in processes and behaviors taken from nature. According to [4], natural computing algorithms can be divided into three main categories:

- evolutionary algorithms;
- swarm intelligence algorithms;
- bacterial foraging algorithms.

They are all based on an observation, that many processes in nature are in fact non-linear, multi-modal and multi-objective optimization processes. The whole evolution process leads to survival of the fittest specimens and species. Within that process there are many constraints, most of which are not well-defined or even not well-known, in some cases they could be fuzzy. Obviously, individuals do not use any numeric, analytical methods to fit, because it is not possible. That conclusion leads to arise of metaheuristic algorithms, which follow the nature patterns and can be used to perform optimization under similar conditions. For the sake of correct understanding the algorithm's abstraction, it is very important to emphasize analogies between the biological domain and optimization domain. Evolution's goal (in nature it is simply survival and reproduction of given individual's genes) is reflected by the objective function, which is minimized or maximized during the computation. As the algorithms are iterative, iteration loop reflects the lapse of time and generations succession. Candidate solutions are equivalents of the real population members and they pass from generation to generation using rules which differ between the algorithms and applications. In contrast to nature, algorithms are initialized (that is populated with an initial population) with random solutions to provide some reasonable starting point for the algorithm execution. It is very important to note that natural computation algorithms, as they are metaheuristics, do not guarantee finding optimal solution. Although, given enough population size and generation number, they usually lead to the suboptimal solutions which are suitable enough for the most of applications.

Two algorithms for the evaluation and application were chosen in the described research: the Flower Pollination Algorithm (FPA) and the Krill Herd Algorithm (KHA).

*1) Flower Pollination Algorithm:* The first one has been proposed by Xin-She Yang of Department of Engineering, University of Cambridge in the paper [5] with further description in the paper [6]. It belongs to the class of evolutionary algorithms and it reflects the process of reproduction of flowering plants by the pollination. It is performed by pollinators, such as insects, birds or bats (biotic pollination) or by the wind or water (abiotic pollination). Pollination can take the form of a cross-pollination, in which a flower is pollinated with pollens coming from flowers of different plant, or a self-pollination, in which flowers on the same plant pollinate each other. Some pollinators stick to some species and specialize in pollinating them. This phenomenon is called flower constancy. Author highlighted four abstractions which make up the basis of the FPA:

- global pollination is an abstraction of biotic cross-pollination, pollinators move by performing Lévy flights;
- local pollination is an abstraction of abiotic self-pollination;
- flower constancy is expressed by the fact that reproduction probability increases appropriately to similarity of the two involved flowers;
- probability of occurrence of local and global pollination is defined by the switch probability p, which abstracts external factors, such as wind, physical proximity of different plants and so on.

There are some additional assumptions and rules as well:

- the best solution from the generation $g_i$ passes directly to the next generation $g_{i+1}$;

**Algorithm 1** FPA pseudocode

```
Objective min or max $f(X), X = (X_1, X_2, ..., X_d)$
Initialize a population of $n$ flowers/pollen gametes with random values
Find the best solution $g_*$ in the initial population
Define a switch probability $p \in [0, 1]$
while $t < MaxGeneration$ do
  for $i = 1..n$ do
    if $rand < p$ then
      Draw a ($d$-dimensional) step vector $L$ which obeys a Lévy distribution
      Global pollination via $X_i^{t+1} = X_i^t + L(X_i^t - g_*)$
    else
      Draw $\epsilon$ from an uniform distribution in $[0, 1]$
      Randomly choose $j$ and $k$ among all the solutions
      Local pollination via $X_i^{t+1} = X_i^t + \epsilon(X_j^t - X_k^t)$
    end if
    Evaluate new solution
    If the new solution is better, update it in population
  end for
  Find the current best solution $g_*$
end while
```

- with global pollination, candidate solutions tend to the current most fit individual;
- a random value determining pollination type is obtained for every flower in every iteration;
- newly generated solution passes to the next generation if, and only if, it is better fitted than its predecessor.

Assuming that the current best solution is represented by $g_*$, generating solution $x_i$ in step $t + 1$ in global pollination is performed by the following formula:

$$X_i^{t+1} = X_i^t + L(X_i^t - g_*) \tag{1}$$

where $L > 0$ is acquired from the Lévy distribution. A local pollination version is in turn expressed by the formula:

$$X_i^{t+1} = X_i^t + \epsilon(X_j^t - X_k^t) \tag{2}$$

where $\epsilon$ is a value obtained from the regular uniform distribution.

The FPA has been benchmarked with some typical two dimensional functions enumerated in the original FPA article, as well as custom function created especially for the sake of research. Tests were performed on the implementation created in the SciPy environment using a small custom test framework. Tunable parameters of the algorithm were changed between specific test runs, which led to the following conclusions:

- finding a solution for the unimodal problem requires much less effort (in terms of generation number and population size) than for the multimodal problem;
- probability switch parameter p does not matter significantly in unimodal problems;
- given enough iterations, every other parameter also does not really matter in unimodal problem solving;

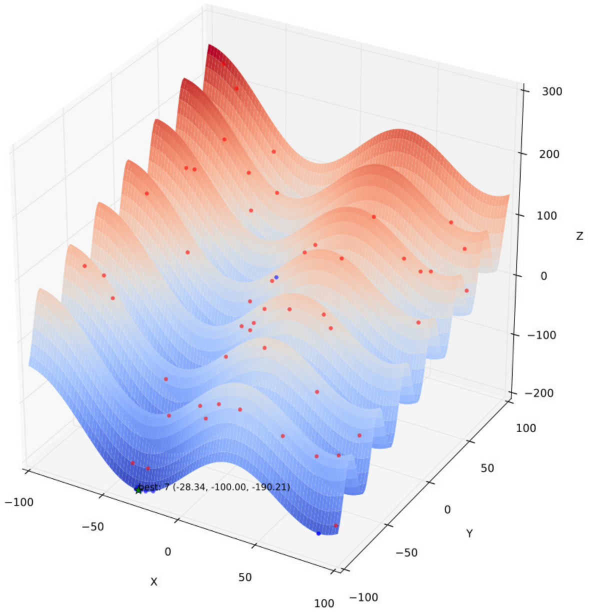


Fig. 1. Results of the application of FPA to the custom benchmark two-dimensional function.

- for multimodal problems, increasing population size gave significantly less improvement than increasing iteration number;
- in extreme cases, during the few first generations there was no improvement in comparison to the initial random population (!);
- high values of the pollination switch parameter p decreased effectiveness by sticking the computation to local minimum values in a multimodal problems.

Figure 1 presents a plot of the custom two dimensional benchmark function. Its objective function is the minimum in the domain of $[-100, 100]$ in both dimensions.

$$\begin{aligned} f(x) \quad &= \quad \left(\left(\frac{x_i}{20}\right)^2 - 2\right)\left(\frac{x_i}{20}+2\right) - \ 2 \\ &+ 50\sin\left(\frac{x_i}{40}\right) + 10\sin\left(\frac{\left(\frac{x_i}{20}\right)^2}{2}\right) \qquad (3) \end{aligned}$$

Red dots are individuals of the initial flower population, while blue dots are individuals of the population in the last generation. Green stars mark the most fit flower in the last generation along with its coordinates.

*2) Krill Herd Algorithm:* The Krill Herd Algorithm represents a slightly different category of metaheuristic nature-inspired algorithms from the FPA. It is a swarm intelligence algorithm. Originally proposed by Amir Hossein Gandomi and Amir Hossein Alavi in the paper [4], it mimics the behaviour of the individual krill specimens moving together as a herd. Such herds, or swarms, move accordingly to environmental factors, but every krill moves separately. An individual's movement is determined by three factors:

- the movement vector of the whole swarm (neighbours within the swarm);
- food foraging;
- additional random bias (diffusion).

After removal of an individual krill (caused by predator attack) from a herd, krills tend to “fix” the low-density gap while still being oriented on finding food. From this emerges a multiobjective optimization problem. Overall, generalized n-dimensional formula for difference of krill position in subsequent time units goes as follows:

$$\frac{dX_i}{dt} \quad = \quad N_i + F_i + D_i \qquad (4)$$

Enumerated aspects of individual krill moves can be described by a set of equations:

- $N_i$ - motion induced by neighbours:

$$N_i^{new} \quad = \quad N^{max}\alpha_i + \omega_n N_i^{old} \qquad (5)$$

where $N^{max}$ is the maximum possible speed that can be induced, $\omega_n$ in the range $[0, 1]$ is the inertia weight of an individual krill. $N_i^{old}$ is the motion induced in the previous turn and

$$\alpha_i \quad = \quad \alpha_i^{local} + \alpha_i^{target} \qquad (6)$$

$\alpha_i^{local}$ is the local influence of the neighbours on an individual krill, while $\alpha_i^{target}$ is the target direction. Target is determined by the position and movement of the best individual in a swarm.

$$\alpha_i^{local} \quad = \quad \sum_{j=1}^{NN} \hat{K}_{ij}\hat{X}_{ij} \qquad (7)$$

$$\hat{X}_{ij} \quad = \quad \frac{X_j - X_i}{\|X_j - X_i\| + \epsilon} \qquad (8)$$

$$\hat{K}_{ij} \quad = \quad \frac{K_i - K_j}{K^{worst} - K^{best}} \qquad (9)$$

where $K$ in general is a fitness value of a given krill, so $K^{worst}$ and $K^{best}$ are the worst and the best fitness degrees achieved so far by any individuals. $NN$ is a number of reachable krill neighbours and $\epsilon$ is an immaterial positive number introduced to avoid singularities in the formula.

The $NN$ value depends on a stated sensing scope of krill individuals. It can be defined in a static way, e.g. individuals always take into consideration a constant number of closest krills, regardless of their distance. The other way is to determine neighbour sets using the heuristic in every iteration:

$$d_{s,i} \quad = \quad \frac{1}{5N}\sum_{j=1}^{N} \|X_i - X_j\| \qquad (10)$$

Every krill individual has its target vector defined as follows:

$$\alpha_i^{target} \quad = \quad C^{best}\hat{K}_{i,best}\hat{X}_{i,best} \qquad (11)$$

where

$$C^{best} = 2\Big(rand + \frac{I}{I_{max}}\Big) \tag{12}$$

$I$, $I_{max}$ is the current iteration number and a maximum number of iterations. $rand$ is a random value between 0 and 1.

- $F_i$ - food foraging:

$$F_i = V_f \beta_i + \omega_f F_i^{old} \tag{13}$$

where $V_f$ is the food foraging speed and $\omega_f$, as previously, is the inertia of the movement. Food fitness of the individual is defined as follows:

$$\beta_i = \beta_i^{food} + \beta_i^{best} \tag{14}$$

And food attraction for the krill individual is:

$$\beta_i^{food} = C^{food} \hat{K}_{i,food} \hat{X}_{i,food} \tag{15}$$

The food coefficient, expressing global attraction of the food center, is:

$$C^{food} = 2\Big(1 - \frac{I}{I_{max}}\Big) \tag{16}$$

$$\beta_i^{best} = \hat{K}_{i,best} \hat{X}_{i,best} \tag{17}$$

where $K_{i,best}$ is the best fit achieved by given krill individual so far.

- $D_i$ - random physical diffusion. In the simplest case it can be fully random. A general rule concerning the diffusion states that the better the krill's position is, the less its random motion diffusion is. The following formula defines a diffusion on the basis of that rule and the assumption that krills' positions improve by the time:

$$D_i = D^{max}\Big(1 - \frac{I}{I_{max}}\Big)\delta \tag{18}$$

where $D^{max}$ is the maximum possible diffusion, $\delta$ is the random directional vector.

The distinctive feature of the KH algorithm is a implementation of two basic evolutionary operators, crossover and mutation, despite it is not really an evolutionary algorithm. The crossover operator is inspired by the genetic algorithms. Its result is an individual with some features of both "parents". Mutation, in turn, is a random change in an individual's features. Paper [7] contains a description of the process of incorporating mutation scheme into the KH algorithm. A stop condition for the algorithm could be a time limit, reaching a desired fitness level or the combination of these two. High level pseudocode of the KHA is presented as the Algorithm 2. Applications and studies on parameters tuning of the KHA have been described in publications: [8], [9], [10] and [11]. Papers [12] and [13] contain other proposed modification of the algorithm.

**Algorithm 2** High-level KHA pseudocode

```
Define and populate algorithm's data structures
Initialize random initial population
while stop condition is reached do
    Evaluate fitness of each krill individual on the basis of its position
    Calculate motion of each krill individual
    Perform genetic operations
    Update each krill position in the search space according to calculated motion and eventual genetic operations results
end while
```

## II. CODB - PROTOTYPE IMPLEMENTATION OF A COLUMN-ORIENTED DB

One of the main goals of the research is the implementation of a column-oriented database management system (the term CODB will be used). A typical contemporary Java SE development stack has been chosen as an implementation environment. It consists of the standard Java Development Kit 8 along with supplementary libraries Google Guava (general purpose library), Logback (logging facility), jUnit (unit testing framework) and Mockito (mocking facility for unit tests). This implementation is used as a foundation for the further research.

Implementation objectives were stated as follows:

- possibility to be used as embedded database in Java and other JVM-based programming languages, but enabling future use with custom connectivity protocol and drivers for other languages, as well as a REST service;
- usage of the memory-mapped data files, thus requiring 64-bit OS for sufficient performance;
- UUID v1 as an objects' identifiers;
- custom binary storage file structure (please see the description below);
- all values stored as an UTF-8 encoded strings;
- full in-memory indexing;
- static storage garbage collection;
- full unit test coverage of logic and storage code (except so called boilerplate code, such as field accessor methods);
- lack of any access rights/database user management facility;
- object-oriented API;
- partitioning and/or sharding support.

### A. Storage structure

The CODB supports one keyspace in one running instance, i.e. one instance serves for only one keyspace. Keyspace is simply a directory in the operating system's file system tree. It does not have any metadata except a name. The keyspace directory contains subdirectories representing column families.

The column family directory contains binary column datafiles, one file per one column. Binary data files contain a sequence of value entries. Every value entry consists of the actual UTF-8 encoded value, a set of UUIDs of records which contains a given value and length values necessary to calculate offsets. UUIDs are stored as a pair of long (64-bit) values. Every length value is just a long 64-bit value. Please note that every specific value is stored only once, notwithstanding the number of records which contain it. Table I shows a structure of a single entry in the column datafile.

The current implementation offers full in-memory indexing. A full datafiles scan is performed during startup in order to build the indices. Planned materialized indices or eventual column metadata would be stored in the same directory as the data. The indexing facility in the CODB engine has two aspects:

- indexing physical location (offset) of column values in the datafile;
- indexing offset of spare space fragments in the datafile (see description below).

Because of performance reasons record entries (that is pairs (value, UUID)) are not actually updated, but removed and created again with new values. When a value which belongs to a given record is changed, the record's identifier is removed from the UUID set assigned to the previous value and will be added into UUID set of a new value. If there is no entry of a new value, it is appended at the end of the datafile. If the identifier set of the new value does not have enough space to be extended directly in its place, it is removed from its current place and appended at the end of the datafile. When a value is removed from the last record which possessed it, the record's identifier is removed from the set but the value remains in the datafile. That feature raises a need of some garbage collecting mechanism. A prototype implementation offers a static garbage collection facility, that is a separate application which is intended to be run when the actual DBMS engine is not running. The garbage collector performs a full data scan to build the full index and then squashes the datafile by placing values one directly after another and throwing out the values without any corresponding records. That way, spacings introduced while updating values, as well as unused value entries, are removed from the datafiles.

## III. Mathematical model of a column-oriented DB and application

In order to apply selected natural computation algorithms to the CODB project, one needs to translate fundamental concepts of the domain into the mathematical model which reflects the nature domain. Correct translation between these three fundamentally different worlds is the key to obtaining satisfying results. Natural computation algorithms are a general heuristic mean of optimization problem solving. They are not tailored to any particular problem domain, but the problem itself should provide a mathematical model instead. Such models consist of a few basic components:

- objective function;
- input as a scalar or vector value or n-dimensional matrix;
- output as a scalar or vector value or n-dimensional matrix.

The objective function for the considered problem must take several criteria into consideration when calculating the score for the candidate partitioning solution, these include:

- partitioning scheme application cost, that is cost of moving data back and forth between database engine instances;
- estimated cost of the data division introduced by the selected partitioning scheme.

Input for the objective function is a candidate solution. Its output is the fitness degree, a value which determines a given candidate's quality. Enumerated objectives can be formulated on the basis of a query log and other data usage statistics, which must be analysed for mutual co-appearance of different data parts. The superior goal here is placing data which often occur together in the same instance.

The input of the algorithm should consider at least the following aspects:

- relationship between different data parts;
- current state of the database;
- cost of moving the data item between instances (usually size).

For every application, such a list can differ fundamentally, but probably always it will be an n-dimensional matrix, in which one of the dimensions will represent all records currently stored in the considered database.

The output of the algorithm is the most fit model of partitioning the database. In the simplest, non-optimized case, it will be at least a 2-dimensional matrix, in which one dimension covers all the records.

## IV. Summary

This paper summarises the overall perspective on the research about horizontal scalability of column-oriented databases with use of natural computing algorithms. The research has already brought some encouraging results. Currently, the biggest challenge is the creation of an appropriate objective function reflecting all the necessary aspects of distributed database operation. Correct mapping of the problem from the database domain to the mathematical model also is crucial for the effectiveness of the proposed solution. The primary goal at the current stage is careful creation of the model and elaboration of the database usage log analysis methods. Another important objective is inspection of the FPA and KHA algorithms' features and properties in details. Creation of the fully featured column oriented database management system in Java, intended as a foundation of solution application is yet another goal, although it goes beyond the scope of the research.

## References

[1] A. Nowosielski, "RDBMS horizontal scalability – architectures review and example implementation," 2013, AGH University of Science and Technology, M.Sc.-Thesis.

[2] D. Abadi, "The design and implementation of modern column-oriented database systems," *Foundations and Trends® in Databases*, vol. 5, no. 3, pp. 197–280, 2012. doi: 10.1561/1900000024. [Online]. Available: http://dx.doi.org/10.1561/1900000024

TABLE I
SINGLE ENTRY STORAGE STRUCTURE

| Data type | Size | Content |
|---|---|---|
| long | $8B$ | value length - $l$ |
| UTF-8 encoded string | $\geq lB$ | actual value |
| long | $8B$ | number of associated keys - $n$ |
| 0 or more (long,long) pairs | $n \times 16B$ | big endian encoded keys |

[3] "Apache Cassandra$^{\text{TM}}$ 2.1 Documentation," 2014, (access 28th June 2015). [Online]. Available: http://docs.datastax.com/en/cassandra/2.1/pdf/cassandra21.pdf

[4] A. H. Gandomi and A. H. Alavi, "Krill herd: A new bio-inspired optimization algorithm," *Communications in Nonlinear Science and Numerical Simulation*, vol. 17, no. 12, pp. 4831–4845, 2012. doi: 10.1016/j.cnsns.2012.05.010. [Online]. Available: http://dx.doi.org/10.1016/j.cnsns.2012.05.010

[5] X.-S. Yang, "Flower pollination algorithm for global optimization," in *Unconventional Computation and Natural Computation*. Springer Science  Business Media, 2012, pp. 240–249. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-32894-7_27

[6] X.-S. Yang, M. Karamanoglu, and X. He, "Multi-objective flower algorithm for optimization," *Procedia Computer Science*, vol. 18, pp. 861–868, 2013. doi: 10.1016/j.procs.2013.05.251. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2013.05.251

[7] G. Wang, L. Guo, H. Wang, H. Duan, L. Liu, and J. Li, "Erratum to: Incorporating mutation scheme into krill herd algorithm for global numerical optimization," *Neural Comput & Applic*, vol. 24, no. 5, pp. 1231–1231, 2013. doi: 10.1007/s00521-013-1422-y. [Online]. Available: http://dx.doi.org/10.1007/s00521-013-1422-y

[8] S. Łukasik and P. A. Kowalski, "Study of flower pollination algorithm for continuous optimization," in *Intelligent Systems'2014*. Springer Science  Business Media, 2015, pp. 451–459. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11313-5_40

[9] P. A. Kowalski and S. Łukasik, "Experimental study of selected parameters of the krill herd algorithm," in *Intelligent Systems'2014*. Springer Science  Business Media, 2015, pp. 473–485. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11313-5_42

[10] G. P. Singh and A. Singh, "Comparative study of krill herd, firefly and cuckoo search algorithms for unimodal and multimodal optimization," *IJISA*, vol. 6, no. 3, pp. 35–49, 2014. doi: 10.5815/ijisa.2014.03.04. [Online]. Available: http://dx.doi.org/10.5815/ijisa.2014.03.04

[11] P. K. Adhvaryyu, P. K. Chattopadhyay, and A. Bhattacharjya, "Application of bio-inspired krill herd algorithm to combined heat and power economic dispatch," in *2014 IEEE Innovative Smart Grid Technologies - Asia*. IEEE, 2014. doi: 10.1109/isgt-asia.2014.6873814. [Online]. Available: http://dx.doi.org/10.1109/isgt-asia.2014.6873814

[12] L. Guo, G.-G. Wang, A. H. Gandomi, A. H. Alavi, and H. Duan, "A new improved krill herd algorithm for global numerical optimization," *Neurocomputing*, vol. 138, pp. 392–402, 2014. doi: 10.1016/j.neucom.2014.01.023. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2014.01.023

[13] G.-G. Wang, A. H. Gandomi, and A. H. Alavi, "Stud krill herd algorithm," *Neurocomputing*, vol. 128, pp. 363–370, 2014. doi: 10.1016/j.neucom.2013.08.031. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2013.08.031

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1043–1047

DOI: 10.15439/2015F62
ACSIS, Vol. 5

# Evaluation of Methods to Combine Different Speech Recognizers

Tomas Rasymas
Vilnius University
Muitinės St. 8, Kaunas, Lithuania
Email: tomas.rasymas@khf.vu.lt

Vytautas Rudžionis
Vilnius University
Muitinės St. 8, Kaunas, Lithuanian
Email: rudzionis@vukhf.lt

***Abstract*— The paper deals with the problem of improving speech recognition by combining outputs of several different recognizers. We are presenting our results obtained by experimenting with different classification methods which are suitable to combine outputs of different speech recognizers. Methods which were evaluated are: k-Nearest neighbors (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR) and maximum likelihood (ML). Results showed, that highest accuracy (98.16 %) was obtained when k-Nearest neighbors method was used with 15 nearest neighbors. In this case accuracy was increased by 7.78 % compared with best single recognizer result. In our experiments we tried to combine one native (Lithuanian language) and few foreign speech recognizers: Russian, English and two German recognizers. For the adaptation of foreign language speech recognizers we used text transcribing method which is based on formal rules. Our experiments proved, that recognition accuracy improves when few speech recognizers are combined.**

## I. Introduction

Speech recognition applications could be subdivided into two broad classes: the applications using large vocabulary continuous speech recognition and applications using the recognition of voice commands from a predefined set of voice commands. It may seem that the first type of applications has the wider area of possible applications. But it is more complicated task to ensure the necessary recognition accuracy when using large vocabulary continuous speech recognition. At the same time there are a lot of potential applications when high accuracy of voice commands from a predefined set of allowable voice commands (may be even very big set of voice commands) is enough to achieve users satisfaction. The area of similar applications is big and such applications could be developed more rapidly than applications based on continuous speech recognition. The areas of voice commands based applications could be transport, logistic, medical and other information systems, various personal assistants, etc. It should be noted that for widely used languages (English, Spanish, German, etc.) voice recognition based applications became everyday reality and could be found in a various situations and areas. Among the well known examples we can mention set of tools distributed by Google or Nuance.

The development of large vocabulary speech recognition systems requires enormous resources: both material and human resources. It is difficult to find such resources in the countries where relatively not widely spoken languages are used as a primary mean of communication. This could be illustrated by the fact, that companies such as Microsoft, Apple, Nuance aren't particularly interested in developing Lithuanian speech recognition systems, because Lithuanian language is not so widely used as some others and don't have significant market potential. Among the possible solutions for the problem might be to try to create own speech recognition engine, or to adapt the ones created for other languages. The proprietary recognizer has bigger potential and is more flexible solution, but this is also the more costly solution. At the same time it has been shown that proper adaptation of existing foreign language acoustic models could speed up the development of recognizer and lead to the acceptable recognition level in that language [1]–[4], [6], [7]. Some previous studies have shown that speech recognition systems of languages such as English, Spanish or Russian can be quite well adapted for Lithuanian speech recognition [1], [3], [4]. However, the recognition results are not always as good as necessary and depend on many factors. So, it is natural to try to create hybrid systems, which are based on combination of different speech recognition systems and consequently try to achieve better recognition accuracy. The essence of hybrid recognition is a parallel use of several different recognizers with the hope, that at least one of the recognizers will give the correct result and it will be possible to detect the correct answer [4]. Hybrid approach is one of the ways to achieve higher recognition accuracy in speech recognition systems. This implies combination of hypotheses provided by different recognition engines in order to get higher recognition accuracy.

The idea of creating hybrid speech recognizer and adapting other languages acoustic models is not new. These kinds of researches are especially important for all under resourced languages. There were successful attempts to estimate acoustic models for new target language using speech data from varied source languages, but only limited data from the target language [10]. Also, Google researchers show very promising results in transformation of English to other languages such as Lithuanian, French and so on. What is more, researchers are experimenting with different acoustic models adaptation methods in order to maximize the recognition performance with small amount of non-native data available [11]. Statistical algorithms for combining different acoustic models are used quite often and produces promising results [1], [3], [4], [6], [11], [12]. These researches shows, that in many cases it is possible to achieve high enough recognition accuracy by using hybrid systems with adapted acoustic models.

The paper presents our activities to adapt several foreign

978-83-60810-66-8/$25.00©2015, IEEE

language (English, German, Russian) speech recognizers for the recognition of limited Lithuanian vocabulary and evaluate some methods (k-Nearest neighbors, Linear Discriminant Analysis, Quadratic discriminant Analysis, Logistic Regression, and maximum likelihood), used for different speech recognizers combination.

Further paper is organized as follows. In Chapters I and III we are presenting method and tools used for adaptation of foreign language recognizers. In Chapter IV there is presented prototype system used in experimental evaluation experiments. Chapter V briefly summarizes the speech corpus used in recognition experiments. Finally in Chapter VI there are presented and discussed the results of experiments. In Chapter VII several conclusions are presented and discussed.

## II. Foreign Language Recognizers Adaptation

For the evaluation purposes we decided to use one native[1] (Lithuanian) and several foreign language recognizers. Among foreign language recognizers we used Russian[2], English[3] and two German[4] language open source speech recognizers. The adaptation procedure will be described as follows. First of all foreign speech recognizers were adapted to recognize Lithuanian commands. Adaptation was done by using formal rules method [5]. All Lithuanian commands, that were collected in this corpus, where transcribed by using foreign language phonemes. By using formal rules method a set of transcription rules were created. The structure of rules was as follows: left context; current letter; right context and list of phonetic units. This list represents foreign language sound that best matches current letter with left and right contexts. If left or right context of the rule can be any, then symbol '*' was used. In this way the new written form of Lithuanian voice command was obtained. Some of the transcribing rules are listed in Table I.

Table I.
Some examples of transcribing rules

| Transcribing rules | | | |
|---|---|---|---|
| **English (voxforge)** | **Russian** | **German** | **German (voxforge)** |
| *;A;I;AY,AA IY | *;A;I;ay | *;A;I;ai | *;A;I;AY |
| *;E;I;EH IY | *;E;I;e ii | *;E;I;ei | *;E;I;EH IIH |
| *;O;I;OY | *;O;I;oo ii | *;O;I;oy | *;O;I;OY |
| *;U;I;UW IY | *;U;I;uu ii | *;U;I;ui | *;U;I;UU IIH |
| *;A;U;AW | *;A;U;aa uu | *;A;U;au | *;A;U;AW |
| *;E;U;EH W | *;E;U;ae uu | *;E;U;ee uu | *;E;U;EH UUH |
| *;O;U;OW | *;O;U;oo uu | *;O;U;oo uu | *;O;U;OOH UUH |
| *;U;O;UW AO | *;U;O;uu oo | *;U;O;uu oo | *;U;O;Y OOH |
| *;I;E;IY AE | *;I;E;i ae | *;I;E;ii ee: | *;I;E;IIH EEH |
| *;I;AI;EY | *;I;AI;i ay | *;I;AI;ii ai | *;I;AI;IH AY |

## III. Methods Used For Evaluation

We proposed a method to combine different speech recognition engines by using neural networks algorithms [4]. Results in earlier studies showed, that this method increased speech recognition accuracy by almost 5% compared with the best results of single recognizer. As the next step we decided to evaluate other methods and to see how efficient they could be for combining different speech recognizers. We selected five methods which we think are quite good for this task: k-Nearest neighbors (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR) and maximum likelihood (ML). These methods were selected because of their efficiency and well studied properties.

## IV. Hybrid Speech Recognition Prototype

For evaluation of the selected methods hybrid speech recognition system prototype was developed. Python programming language was used for its development. Block diagram of such system is showed in Fig. 1.

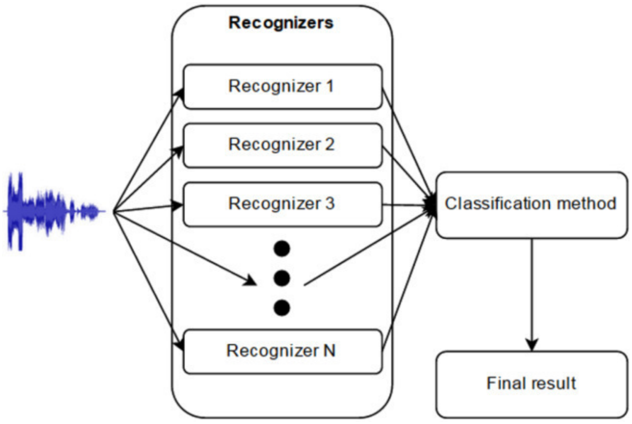


Fig. 1. Block diagram of hybrid speech recognition system.

As could be seen in the prototype, voice command is passed to all speech recognizers in parallel. After that, all recognizers produces output. Output of the recognizer is the hypothesis: score of how well audio signal matches the acoustic model [8]. This hypothesis score is passed to classification algorithm and it makes final decision.

To develop speech recognizers, PocketSphinx toolkit was used. PocketSphinx is a lightweight speech recognition engine, specifically tuned for handheld and mobile devices, though it works equally well on the desktop computers and notebooks. It is distributed under the same permissive license as Sphinx toolkit itself. Algorithmically this is hidden Markov model based speech recognition framework, which provides simple way for creating custom speech recognition systems [8].

For the quicker classification methods realization, we used scikit-learn library [9]. Scikit-learn is an open source machine learning library for the Python programming language. It realizes various classifications, regression and clustering algorithms including support vector machines, logistic regression, naive Bayes, random forests, gradient boosting, k-means and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy [9].

## V. Speech Corpus

A speech corpus of 25 drug names and 25 names of dis-

[1] Downloaded from https://github.com/mondhs/lt-pocketsphinx-tutorial/tree/master/impl/models/hmm/lt.cd_cont_200

[2] Downloaded from http://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/Russian%20Voxforge

[3] Downloaded from http://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/English%20Voxforge

[4] Downloaded from https://www.lt.informatik.tu-darmstadt.de/de/data/open-acoustic-models and http://goofy.zamia.org/voxforge/de

eases was used. Speech commands, collected in the corpus, are shown in the Table II.

TABLE II.
SPEECH CORPUS USED FOR METHODS EVALUATION

| | | |
|---|---|---|
| ANALGINAS | RADIREKSAS | ARTERIJŲ EMBOLIJA |
| BIFOVALIS | RANIGASTAS | ARTERINĖ HIPERTENZIJA |
| CYKLODOLIS | TRACHISANAS | ARTERIJŲ TROMBOZĖ |
| ENARENALIS | TRAVATANAS | ARTROZĖ |
| FERVEKSAS | TRENTALIS | ATEROSKLEROZĖ |
| GASTROVALIS | TRILEPTALIS | ATOPINIS DERMATITAS |
| HEKSORALIS | VALOKORDIN LAŠAI | BIPOLINIS AFEKTINIS SUTRIKIMAS |
| HEMATOGENAS | VERDINAS | BLAUZDOS KAULŲ LŪŽIAI |
| KETANOVAS | AIDS | BRONCHŲ ASTMA |
| KETONALIS | AKIŲ NUDEGIMAI | CELIULITAS |
| KREONAS | AKTINOMIKOZĖ | CHEMINIAI NUDEGIMAI |
| METFORALIS | ALERGIJA | CISTITAS |
| MIKARDIS | ALKOHOLIO TOKSINIS POVEIKIS | CUKRINIS DIABETAS |
| NEBIKARDAS | ANAFILAKSINIS ŠOKAS | DANTŲ DYGIMO SINDROMAS |
| PANANGINAS | ANKILOZINIS SPONDILITAS | DANTŲ DYGIMO SUTRIKIMAI |
| PREDUKTALIS | ANTRINĖ GLAUKOMA | DANTŲ VYSTYMOSI SUTRIKIMAI |
| PROPODEZAS | APELSINO ŽIEVELĖ | |

Speech corpus, used in the experiments, was gathered by recording speech of 12 people (5 female and 7 male). Each of these speakers pronounced each command name 20 times at sampling rate 16 kHz in a single session. So, every command was pronounced for 240 times. Vocabulary of all commands used in this experiment is listed in Table II.

It should be noted, that the corpus, used in these experiments, is the part of the bigger medical terms Lithuanian speech corpus. The selection of this particular set of voice commands was based on the fact, that 25 commands were those voice commands, which resulted in the highest number of recognition errors using proprietary Lithuanian speech recognizer, while the additional 25 commands were selected randomly.

## VI. EXPERIMENTAL EVALUATION OF DIFFERENT SPEECH RECOGNIZERS COMBINATION METHODS

For the evaluation of methods, we used the developed prototype and described speech corpus. All acoustic models used in the recognition experiments were derived without the use of the speech corpus presented in Chapter V. So the recognition experiments were performed in speaker independent mode. Default PocketSphinx configuration was used for evaluation.

First of all, single recognizers were tested using obtained recordings. Recognition results are shown in Table III.

TABLE III.
SINGLE RECOGNIZERS ACCURACY

| Recognizers | Accuracy, % |
|---|---|
| Lithuanian | 89.26 |
| Russian | 81.32 |
| English (voxforge) | 88.30 |
| German | 81.38 |
| German (voxforge) | 90.38 |

Best results were obtained using German recognizer from voxforge repository. Other recognizers, such as Lithuanian and English (voxforge), showed similar recognition accuracy too. Accuracy of other recognizers was above 80 %, but lower than above mentioned recognizers.

Before the experiments, we thought that Russian recognizer will be one of the best, because Russian language and Lithuanian language have a lot similar sounds, but as results shows, our guess failed.

Later all the selected speech recognizers combination methods were trained using obtained recordings. 168 recordings were used for training and 72 recordings for testing. After training, selected methods accuracy was evaluated. The obtained results are presented in the Table IV.

TABLE IV.
ACCURACY OF COMBINED SPEECH RECOGNIZERS

| Combination method | Accuracy, % |
|---|---|
| k-Nearest neighbors (11) | 89.70 |
| k-Nearest neighbors (15) | 98.16 |
| k-Nearest neighbors (21) | 89.70 |
| Linear Discriminant Analysis | 93.16 |
| Quadratic Discriminant Analysis | 98.05 |
| Logistic Regression | 93.60 |
| Maximum likelihood | 89.70 |

Results shows, that three methods (k-Nearest neighbors (11), k-Nearest neighbors (21) and maximum likelihood) can't be used for speech recognition engine combination, because obtained accuracy is lower than best single recognizer. Other methods are suitable for speech recognizers combination. Best results (98.16 %) were acquired, when k-Nearest neighbors (15) method was used. It is very interesting, that such a simple classifier as k-Nearest neighbors generated the best results. We think that it is because of data used to evaluate selected classification methods. As we know, k-Nearest neighbors classifier requires a small amount of training data to estimate the necessary parameters. We are planning to increase number of data used for classification methods evaluation and repeat experiments to see if our guess is right. Detailed commands recognition accuracy is displayed in Table V (results were rounded to fine integer values).

TABLE V.
RECOGNITION ACCURACY % OF EVERY COMMAND

| Command | k-Nearest neighbors (11) | k-Nearest neighbors (15) | k-Nearest neighbors (21) | Linear Discriminant Analysis | Quadratic Discriminant Analysis | Logistic Regression | Max hypothesis |
|---|---|---|---|---|---|---|---|
| ANALGINAS | 69 | 86 | 69 | 78 | 82 | 79 | 69 |
| BIFOVALIS | 85 | 99 | 85 | 96 | 99 | 96 | 85 |
| CYKLODOLIS | 97 | 100 | 97 | 99 | 100 | 99 | 97 |

| Command | k-Nearest neighbors (11) | k-Nearest neighbors (15) | k-Nearest neighbors (21) | Linear Discriminant Analysis | Quadratic Discriminant Analysis | Logistic Regression | Max hypothesis |
|---|---|---|---|---|---|---|---|
| ENARENALIS | 100 | 100 | 100 | 97 | 100 | 99 | 100 |
| FERVEKSAS | 99 | 99 | 99 | 100 | 99 | 99 | 99 |
| GASTROVALIS | 99 | 99 | 99 | 99 | 99 | 99 | 99 |
| HEKSORALIS | 99 | 100 | 99 | 99 | 99 | 99 | 99 |
| HEMATOGENAS | 97 | 97 | 97 | 97 | 100 | 97 | 97 |
| KETANOVAS | 99 | 100 | 99 | 99 | 100 | 99 | 99 |
| KETONALIS | 92 | 99 | 92 | 93 | 96 | 94 | 92 |
| KREONAS | 82 | 89 | 82 | 83 | 92 | 83 | 82 |
| METFORALIS | 71 | 99 | 71 | 96 | 99 | 97 | 71 |
| MIKARDIS | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| NEBIKARDAS | 96 | 100 | 96 | 100 | 100 | 100 | 96 |
| PANANGINAS | 92 | 92 | 92 | 89 | 92 | 90 | 92 |
| PREDUKTALIS | 97 | 99 | 97 | 96 | 97 | 96 | 97 |
| PROPODEZAS | 97 | 97 | 97 | 97 | 97 | 97 | 97 |
| RADIREKSAS | 96 | 97 | 96 | 88 | 99 | 88 | 96 |
| RANIGASTAS | 94 | 99 | 94 | 97 | 99 | 97 | 94 |
| TRACHISANAS | 100 | 100 | 100 | 99 | 100 | 99 | 100 |
| TRAVATANAS | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| TRENTALIS | 94 | 96 | 94 | 93 | 94 | 93 | 94 |
| TRILEPTALIS | 93 | 96 | 93 | 94 | 96 | 96 | 93 |
| VALOKORDIN LAŠAI | 100 | 100 | 100 | 90 | 100 | 92 | 100 |
| VERDINAS | 97 | 97 | 97 | 97 | 97 | 97 | 97 |
| AIDS | 0 | 100 | 0 | 69 | 93 | 74 | 0 |
| AKIŲ NUDEGIMAI | 90 | 97 | 90 | 81 | 100 | 81 | 90 |
| AKTINOMIKOZĖ | 93 | 100 | 93 | 99 | 100 | 99 | 93 |
| ALERGIJA | 74 | 100 | 74 | 86 | 100 | 88 | 74 |
| ALKOHOLIO TOKSINIS POVEIKIS | 76 | 99 | 76 | 92 | 100 | 92 | 76 |
| ANAFILAKSINIS ŠOKAS | 86 | 100 | 86 | 89 | 100 | 89 | 86 |
| ANKILOZINIS SPONDILITAS | 84 | 100 | 84 | 96 | 100 | 96 | 84 |
| ANTRINĖ GLAUKOMA | 82 | 99 | 82 | 78 | 99 | 78 | 82 |
| APELSINO ŽIEVELĖ | 90 | 100 | 90 | 99 | 100 | 99 | 90 |
| ARTERIJŲ EMBOLIJA | 81 | 92 | 81 | 92 | 89 | 93 | 81 |
| ARTERINĖ HIPERTENZIJA | 92 | 100 | 92 | 97 | 100 | 97 | 92 |
| ARTERIJŲ TROMBOZĖ | 93 | 99 | 93 | 99 | 100 | 99 | 93 |
| ARTROZĖ | 89 | 97 | 89 | 71 | 96 | 72 | 89 |
| ATEROSKLEROZĖ | 82 | 99 | 82 | 94 | 100 | 94 | 82 |
| ATOPINIS DERMATITAS | 92 | 100 | 92 | 92 | 100 | 92 | 92 |
| BIPOLINIS AFEKTINIS SUTRIKIMAS | 100 | 100 | 100 | 97 | 100 | 97 | 100 |
| BLAUZDOS KAULŲ LŪŽIAI | 99 | 99 | 99 | 85 | 100 | 86 | 99 |
| BRONCHŲ ASTMA | 51 | 96 | 51 | 89 | 97 | 90 | 51 |
| CELIULITAS | 97 | 100 | 97 | 97 | 100 | 97 | 97 |
| CHEMINIAI NUDEGIMAI | 100 | 99 | 100 | 93 | 99 | 93 | 100 |
| CISTITAS | 96 | 100 | 96 | 94 | 100 | 97 | 96 |
| CUKRINIS DIABETAS | 99 | 100 | 99 | 100 | 100 | 100 | 99 |
| DANTŲ DYGIMO SINDROMAS | 99 | 100 | 99 | 99 | 100 | 99 | 99 |
| DANTŲ DYGIMO SUTRIKIMAI | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| DANTŲ VYSTYMOSI SUTRIKIMAI | 97 | 97 | 97 | 97 | 97 | 97 | 97 |

We calculated average accuracy of every command and results showed, that almost 58 % of all commands are recognized with 95 – 100 % accuracy, 22 % with 90 – 95 % accuracy, 14 % with 80 – 90 % accuracy and 6 % of all commands are recognized with 40 – 80 % accuracy.

## VII. CONCLUSIONS

The results of our experiments showed, that it could be reasonable to use k-Nearest neighbors (15) or Quadratic Discriminant Analysis methods to combine different speech recognizers using open source PocketSphynx based recognizers. Comparing with the best single recognizer and the best combined speech recognizers, average error was decreased by 7.78 %. In some cases, even bigger increase of recognition accuracy has been observed.

Foreign language speech recognition adaptation shows, that English, German, Russian recognizers could be quite good adapted for Lithuanian voice commands recognition.

One of the interesting areas for further research could be investigation of how different acoustic models from different language could be used to recognize the same Lithuanian voice command.

In the future, we are planning to increase recognition accuracy by finding better transcriptions to recognize Lithuanian commands using foreign languages speech engines. Also, it is necessary to increase size of the vocabulary used in the experiments. Especially important is to increase the variety of the phonetic elements used in the adaptation process.

## REFERENCES

[1] R. Maskeliūnas, A. Rudžionis, K. Ratkevičius, V. Rudžionis, “Investigation of Foreign Languages Models for Lithuanian Speech Recognition”, *Electronics and Electrical Engineering*, no. 3(91), pp. 15–20, 2009.

[2] V. Rudžionis, G. Raškinis, A. Rudžionis, K. Ratkevičius, “Comparative Analysis of Adapted Foreign Language and Native Lithuanian Speech Recognizers for Voice User Interface“, *Electronics and Electrical Engineering*, vol. 19, no. 7, pp. 90–93, 2013.

[3] V. Rudžionis, G. Raškinis, A. Rudžionis, K. Ratkevičius, G. Bartišiūtė, “Web Services Based Hybrid Recognizer of Lithuanian Voice Commands“, *Electronics and Electrical Engineering*, vol. 20, no. 9, pp. 50–53, 2014.

[4] T. Rasymas, V. Rudžionis, “Combining Multiple Foreign Language Speech Recognizers by using Neural Networks”, *Human Language*

*Technologies – The Baltic Perspective,* IOS Press, doi:10.3233/978-1-61499-442-8-33, pp. 33–39, 2014.

[5] P. Kasparaitis, “Transcribing of the Lithuanian Text Using Formal Rules”, *Informatica*, vol. 10, no. 4, pp. 367–376, 1999.

[6] P. Kasparaitis, “Lithuanian Speech Recognition Using the English Recognizer”, *Informatica*, vol. 19, no. 4, pp. 505–516, 2008.

[7] V. Rudžionis, K. Ratkevičius, A. Rudžionis, G. Raškinis, R. Maskeliūnas, “Recognition of Voice Commands Using Hybrid Approach”, *ICIST2013, CCIS 403,* Springer-Verlag Berlin, pp. 249–260, 2013.

[8] D. Huggins-Daines, M. Kumar, A. Chan, A. W Block, M. Ravishankar, A. I. Rudnicky, “Pocketsphinx: a free, real-time continuous speech recognition system for hand-held devices”, *IEEE ICASSP 2006 Proceedings,* vol. 1, pp. 185–188, 2006.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Duchesnay, “Scikit-learn: Machine Learning in Python”, *The Journal of Machine Learning Research,* vol. 12, pp. 2825–2830, 2011.

[10] T. Schultz, A. Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition”, *Speech Communication* 35 (1), 31–52, 2001.

[11] Z. Wang, T. Schultz, A. Waibel, “Comparison of Acoustic Model Adaptation Techniques on Non-Native Speech”, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* pp. 540–543, 2003.

[12] H. Meneido, J. Neto, “Combination of acoustic models in continuous speech recognition hybdrid systems”, *Proceedings of the International Conference in Spoken Language Processing,* vol. 9, pp. 1000–1029, 2000.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1049–1054

DOI: 10.15439/2015F222
ACSIS, Vol. 5

# Applying fuzzy clustering method to color image segmentation

Omer Sakarya
University of Gdańsk
Institute of Informatics
osakarya@inf.ug.edu.pl

***Abstract*—The goal of this paper was to apply fuzzy clustering algorithm known as Fuzzy C-Means to color image segmentation, which is an important problem in pattern recognition and computer vision. For computational experiments, serial and parallel versions were implemented. Both were tested using various parameters and random number generator seeds. Various distance measures were used: Euclidean, Manhattan metrics and two versions of Gower coefficient similarity measure. The $F$ and $Q$ segmentation evaluation measures and output images were used to assess the result of color segmentation. Serial and parallel run times were compared.**

## I. Introduction

COLOR image segmentation is a method of assigning pixels of given image to segments which share similar color. Pixels from a segment should be similar colorwise and pixels from different segments should be distinct. The problem of color image segmentation is one of the most difficult problems in computer vision. There exist many algorithms for this particular problem, however none of them work well for all kinds of images. Photos of real world are very different in colors, shapes and noise. Usually before choosing an algorithm for color image segmentation, domain knowledge is used to assess the type of algorithm needed for particular set of photos. The goal of color image segmentation research is to find an universal algorithm that would not require domain knowledge prior use and would provide good results for all kinds of photos. Color image segmentation is an important part of various computer vision problems, including pattern recognition. It is a step performed before pattern recognition, so if the color segmentation is poor, the pattern recognition step may fail. The aim of this paper was to apply fuzzy clustering algorithm known as Fuzzy C-Means to color image segmentation with intention of developing a general method for various types of images.

The paper is organized as follows. Section II introduces color image segmentation problem. In section III distance measures and fuzzy clustering method are described. In section IV the implementation details are presented. Section V presents a possible way of source code parallelization for speed-up. Section VI overviews the computational experiments and achieved results. The last section contains conclusions and plans for further research.

## II. Color image segmentation

Formally image segmentation can be defined as follows [1]: If $P()$ is a homogeneity predicate defined on groups of connected pixels, then segmentation is a partition of the set $F$ into connected subsets or regions $(S_1, S_2, \ldots, S_n)$ such that:

$$\bigcup_{i=1}^{n} S_i = F \wedge \forall_{i \neq j} S_i \cap S_j = \phi \wedge \forall_i P(S_i) = true \wedge \forall_{i \neq j} P(S_i \cup S_j) = false \tag{1}$$

The two most important problems in color image segmentation is choosing the proper algorithm for given type of images and choosing the right colorspace. There are various color representations used in color image segmentation, however none of them is perfect for all kinds of images [1]. In computational experiments the RGB color space was used (see section IV).

## III. Fuzzy clustering

In data clustering, elements from data set are divided into clusters where elements in the same cluster are similar and elements from different clusters are not. There are many similarity and distance measures $dist(x, y)$ that can be used in combination with a clustering algorithm. Example distance measures are Euclidean (2) and Manhattan (3) metrics:

$$d_e(x, y) = \sqrt{(y_1 - x_1)^2 + \ldots + (y_n - x_n)^2} \tag{2}$$

$$d_m(x, y) = \sum_{k=1}^{n} |x_k - y_k| \tag{3}$$

Fuzzy clustering is one of the possible approaches to clustering. As opposed to hard clustering where data element $x$ belongs exclusively to one cluster, in fuzzy clustering element $x$ belongs to every cluster to some degree.

Fuzzy C-Means (FCM) is one of the fuzzy clustering algorithms that can be used for color image segmentation. It is an iterative algorithm that can make use of various similarity and distance measures. The FCM algorithm assigns membership values to each data element, which are inversely related to the relative distance of an element to the centroids. In FCM, the closeness of each data $x_k$ to the center of a cluster $v_i$ (centroid) is defined as the membership $(u_{ik})$ of $x_k$ to the $i$-th cluster of data set minimizing the following objective function [2]:

$$J_m(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} dist(x_k, v_i)^2 \tag{4}$$

where $X = \{x_1, \ldots, x_N\}$ a given set of unlabeled $N$ data; $V = \{v_1, \ldots, v_c\}$ are the cluster centers and $m = [1, \infty]$

978-83-60810-66-8/$25.00©2015, IEEE

is the weighting exponent which determines the fuzziness of the resulting clusters, $U = [u_{ik}]$ matrix $c$ x $n$, where $u_{ik}$ is membership of $x_k$ to the $i$-th cluster $\sum_{i=1}^{c} u_{ik} = 1, \forall k = 1, 2, \ldots, n$. The cluster centers and the memberships are computed as:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}} \tag{5}$$

$$u_{ik} = 1/\sum_{j=1}^{c} \left( \frac{dist(x_k, v_i)}{dist(x_k, v_j)} \right)^{2/(m-1)} \tag{6}$$

Algorithm 1 presents the Fuzzy C-Means method. The $initU$ function randomly initializes the membership matrix $U$ and the parameter $maxItNum$ specifies the maximum number of iterations of the algorithm. $U'$ stores the previous values of $U$ and is used in line 8 to check if the solution found so far has converged according to some $\varepsilon$ and the algorithm should stop, $dist()$ is one of the distance measures. As a result of the algorithm membership matrix $U$ and centroids $V$ are computed. In (6) it is possible that the denominator will be equal to 0, such situation should be considered in computer implementation of the algorithm to avoid errors.

**Algorithm 1** Fuzzy C-Means

```
Require: $X$ - data set, $c$ - number of clusters, $dist$ - distance
    measure, $maxItNum$ - maximum number of iterations
Ensure: $U$ - membership matrix, $V$ - centroids
 1: $U \Leftarrow initU$
 2: $i \Leftarrow 0$
 3: while $i < maxItNum$ do
 4:   compute new centroids $V$ using (5)
 5:   $U' \Leftarrow U$
 6:   compute new membership matrix $U$ using (6) and $dist$
 7:   $i \Leftarrow i + 1$
 8:   if $max|U - U'| < \varepsilon$ then
 9:     break
10:   end if
11: end while
```

## IV. Applying FCM to color image segmentation

Images in a computer are stored in various formats, however many of them use the RGB color space. This means that each pixel is represented by three numbers being the red, green and blue components. Such triples can be treated as vectors, which means that it is is possible to use the FCM algorithm to cluster such data. The advantage of using fuzzy clustering for color image segmentation is that as a result we obtain a membership matrix which may be used to generate more than one segmented image. This can be achieved by choosing for each pixel which one of the membership values we want to use as the one defining the final color of a pixel in segmented image.

To perform the computational experiments for this paper the FCM algorithm was implemented in the C programming language on GNU/Linux operating system. The portable pixmap image format was used for simplicity, because it only contains a small header and the following values represent RGB triples. Binary method of randomly initializing the membership matrix $U$ was used, where each column has value $1$ in one of the rows, and the rest contains $0$. The output membership matrix $U$ and centroid values $V$ were used to generate the segmented image file. In $U$, the membership information defines to which segments to what degree given pixel belongs. The centroids $V$ contain colors of the segments. For each column of $U$ the maximum membership value was selected, then the index of this value was used to assign color from the set of centroids $V$ to a pixel. Algorithm 2 presents the segmentation method.

**Algorithm 2** Segmentation

```
Require: $pixels$ - array of RGB triples representing pixels
    of original image, $N$ - number of pixels, $c$ - number of
    clusters, $U$ - membership matrix, $V$ - centroids
Ensure: $pixels$ - segmented image
for $k \Leftarrow 1; k \leq N; k \Leftarrow k + 1$ do
  $m \Leftarrow 0$
  for $i \Leftarrow 1; i \leq c; i \Leftarrow i + 1$ do
    if $u_{ik} > u_{mk}$ then
      $m \Leftarrow i$
    end if
  end for
  $pixels_k \Leftarrow V_m$
end for
```

## V. Parallelizing source code using OpenMP

Serial source code was parallelized using OpenMP. Independent 'for loop' iterations were identified and OpenMP pragmas were used, which resulted in speed-up caused by parallel computation. This solution is automatically scalable, which means that when the same program is executed on a CPU with more computing cores, the program will use all of them automatically and execute faster. The experiments for this paper were performed on a laptop with Intel Core 2 Duo processor.

Although the main loop in FCM algorithm is not independent, it was possible to parallelize centroids vector $V$ and membership matrix $U$ computation in each iteration. Each centroid and membership can be computed independently from others, there is no data race condition. In the implementation, for research purposes, also the computation of $J_m$ objective function and square error criterion were parallelized. It is important to note that there was no need to modify the algorithm, only the source code was parallelized using few OpenMP pragmas.

## VI. Results overview

Two versions of the program were used - serial and parallel. The difference between them was that the parallel version used OpenMP pragmas. Both versions were compiled from the same source code, where the serial version had disabled pragmas. The run time was measured for both programs executed with the same parameters, however it is the wall-clock time of the whole program, which not only computes the FCM function, but also reads, writes files and computes $J_m$, square error criterion and more. The programs would execute even faster if the additional operations and computations were removed. The time measurement is just a very general information of how OpenMP pragmas influenced the run time of a parallel program in comparison to serial program.

The experiment was performed using different parameters on the same image (photo), it was 402 pixels wide and 600 pixels high. Parameters that could be specified for the program were: photo file, number of clusters, maximum iterations number, distance measure and random number generator seed. The available distance measures were Euclidean, Manhattan metrics and two versions of Gower coefficient [3] - regular (7,9) and modified (8,9). The modified version that takes specifics of RGB vector data into account was prepared for experiments in this paper.

$$S_i(x_i, y_i) = \begin{cases} 1, & \text{if } x_i = y_i \\ 0, & \text{if } x_i \neq y_i \end{cases} \tag{7}$$

$$S_i(x_i, y_i) = \begin{cases} 1, & \text{if } x_i \in [y_i - C, y_i + C] \\ 0, & \text{if } x_i \notin [y_i - C, y_i + C] \end{cases} \tag{8}$$

$$S(x, y) = \sum_{i=1}^{n} \frac{\omega_i S_i(x_i, y_i)}{n} \tag{9}$$

Since Gower coefficient is a similarity measure, it was converted to distance measure as follows: $dist(x, y) = 1 - S(x, y)$. $C$ is some constant and $\omega$ is weight applied to similarity of the $i$-th element of data vector. For experiments, the value of $\omega$ was set to 1, which means that each color component of pixel RGB vector was equally important and various values of $C$ were tested. No satisfactory results were achieved with regular and modified Gower coefficient. The final segmented image had only one, sometimes few segments with colors not similar to colors found in original photo (see table VII). Specifying random number generator seed allowed to test both programs with the same pseudo random numbers. For each parameter setting and different seed the programs were executed 10 times.

As suggested in [4] two evaluation measures $F$ and $Q$ were used (10,11) to assess the quality of result color image segmentation. Both measures do not require any parameter or threshold value and can be used for automatic evaluation, however it is important to remember that they should not be treated as definitive evaluation of final segmented image.

$$F(I) = \frac{1}{1000 \times N} \sqrt{r} \sum_{i=1}^{r} \frac{e_i^2}{\sqrt{A_i}} \tag{10}$$

$$Q(I) = \frac{1}{1000 \times N} \sqrt{r} \times \sum_{i=1}^{r} \left[ \frac{e_i^2}{1 + \log A_i} + \left( \frac{r(A_i)}{A_i} \right)^2 \right] \tag{11}$$

$I$ is the segmented image, $N$ the image size (number of pixels), $r$ the number of regions of the segmented image, while $A_i$ and $e_i$ are, respectively, the area and the average color error of the $i$-th region; $e_i$ is defined as the sum of the Euclidean distances between RGB color vectors of the pixels of region $i$ and the color vector attributed to region $i$ in the segmented image. $r(A_i)$ represents the number of regions having an area equal to $A_i$. The smaller the $F$ and $Q$, the better the segmentation result should be. Equation (10) is composed of three terms: the first is a normalization factor that takes into account the size of the image, the second, $\sqrt{r}$, penalizes segmentations that form too many regions, the last term, the sum, penalizes segmentations having non-homogeneous regions. Since the average color error $e_i$ of the region is significantly higher for large regions than for small ones, $e_i$ has been scaled by the factor $\sqrt{A_i}$. In equation (11) the body of the sum is composed of two terms: the first is high only for non-homogeneous regions (typically, large ones), while the second term is high only for regions whose area $A_i$ is equal to the area of many other regions in the segmented image (typically, small ones) [4].

Number of result nonzero clusters was recorded, usually it was smaller than required through parameter $c$ number of clusters. The average, minimum and maximum values of $F$, $Q$, serial and parallel run times were computed. During the experiments, the value of $m$ was set to 2. Computational experiment results are presented in tables I-VI and the images are in tables VII and VIII.

## VII. Conclusions

While performing computational experiments the following observations were made.

- The final number of clusters was always smaller than required $c$
- Using OpenMP caused significant speed-up
- Euclidean and Manhattan metrics were used successfully
- In performed experiments, both regular and modified Gower coefficient similarity measures could not be used efficiently for color image segmentation

Table VIII illustrates the best and worst result color image segmentations according to $Q$ evaluation measure.

Plans for further research include applying various clustering algorithms to color image segmentation for example kernel methods and using different distance measures and color representations. The FCM algorithm may be sensitive to initial membership matrix $U$, so different initialization experiments could be performed. The source code, program output logs, input and output photos are available on-line [5].

TABLE I
EXPERIMENT 1 RESULTS: C=8, MAXITNUM=200, DIST=EUCLIDEAN, SEED=RANDOM; R - NUMBER OF OUTPUT NON-ZERO CLUSTERS

| id | seed | it. num. | r | F | Q | s. time [s] | p. time [s] |
|---|---|---|---|---|---|---|---|
| 1 | 1429546833 | 16 | 4 | 969.95 | 21937.85 | 10.323 | 5.486 |
| 2 | 1429547378 | 35 | 5 | 661.64 | 12932.12 | 22.304 | 11.546 |
| 3 | 1429547770 | 18 | 4 | 850.30 | 18145.39 | 11.587 | 6.089 |
| 4 | 1429548126 | 20 | 4 | 818.16 | 16749.37 | 12.842 | 6.758 |
| 5 | 1429548449 | 44 | 4 | 819.41 | 16973.20 | 27.952 | 14.444 |
| 6 | 1429549265 | 13 | 3 | 1292.71 | 29458.85 | 8.428 | 4.528 |
| 7 | 1429549571 | 25 | 3 | 1821.21 | 48692.53 | 15.982 | 8.351 |
| 8 | 1429549847 | 19 | 4 | 818.12 | 16748.04 | 12.217 | 6.407 |
| 9 | 1429550122 | 23 | 4 | 966.34 | 21815.84 | 15.388 | 8.059 |
| 10 | 1429550350 | 18 | 4 | 850.30 | 18145.39 | 11.585 | 6.116 |

TABLE II
EXPERIMENT 2 RESULTS: C=16, MAXITNUM=200, DIST=EUCLIDEAN, SEED=RANDOM; R - NUMBER OF OUTPUT NON-ZERO CLUSTERS

| id | seed | it. num. | r | F | Q | s. time [s] | p. time [s] |
|---|---|---|---|---|---|---|---|
| 1 | 1429554268 | 28 | 6 | 523.80 | 9184.89 | 64.51 | 32.75 |
| 2 | 1429554740 | 60 | 9 | 385.67 | 6427.85 | 137.63 | 69.58 |
| 3 | 1429555239 | 43 | 7 | 440.25 | 7504.62 | 98.89 | 50.06 |
| 4 | 1429555659 | 68 | 6 | 573.76 | 9821.39 | 155.97 | 78.81 |
| 5 | 1429556554 | 24 | 7 | 481.22 | 7418.76 | 55.26 | 28.05 |
| 6 | 1429556961 | 40 | 6 | 551.52 | 9711.68 | 91.78 | 46.49 |
| 7 | 1429557450 | 63 | 9 | 347.47 | 5299.01 | 146.93 | 74.23 |
| 8 | 1429557970 | 52 | 7 | 422.69 | 7111.55 | 122.66 | 65.5 |
| 9 | 1429558433 | 48 | 7 | 446.11 | 7535.54 | 110.04 | 55.72 |
| 10 | 1429558905 | 39 | 8 | 372.50 | 5664.38 | 89.48 | 49.73 |

TABLE III
EXPERIMENT 3 RESULTS: C=8, MAXITNUM=200, DIST=MANHATTAN, SEED=RANDOM; R - NUMBER OF OUTPUT NON-ZERO CLUSTERS

| id | seed | it. num. | r | F | Q | s. time [s] | p. time [s] |
|---|---|---|---|---|---|---|---|
| 1 | 1430606237 | 16 | 4 | 815.49 | 16742.46 | 8.57 | 4.62 |
| 2 | 1430606543 | 14 | 3 | 1268.10 | 28595.22 | 7.51 | 4.06 |
| 3 | 1430606803 | 12 | 3 | 1268.74 | 28628.85 | 6.48 | 3.53 |
| 4 | 1430607058 | 17 | 4 | 946.32 | 20943.45 | 9.12 | 4.87 |
| 5 | 1430607264 | 16 | 3 | 1205.41 | 26793.42 | 8.55 | 4.58 |
| 6 | 1430607686 | 20 | 5 | 635.21 | 11408.41 | 10.64 | 5.66 |
| 7 | 1430607926 | 26 | 6 | 616.88 | 11120.28 | 13.74 | 7.24 |
| 8 | 1430608179 | 25 | 3 | 1531.17 | 35771.00 | 13.22 | 7.29 |
| 9 | 1430608583 | 14 | 3 | 1268.10 | 28595.22 | 7.5 | 4.05 |
| 10 | 1430608741 | 20 | 4 | 827.35 | 16917.65 | 10.63 | 5.66 |

TABLE IV
EXPERIMENT 1 RESULTS: C=8, MAXITNUM=200, DIST=EUCLIDEAN, SEED=RANDOM; AVERAGE, MINIMUM AND MAXIMUM; R - NUMBER OF OUTPUT NON-ZERO CLUSTERS

| | average | minimum | maximum |
|---|---|---|---|
| it. num. | 23.1 | 13 | 44 |
| r | 3.9 | 3 | 5 |
| F | 986.82 | 661.64 | 1821.21 |
| Q | 22159.86 | 12932.12 | 48692.53 |
| s. time [s] | 14.860 | 8.428 | 27.952 |
| p. time [s] | 7.778 | 4.528 | 14.444 |

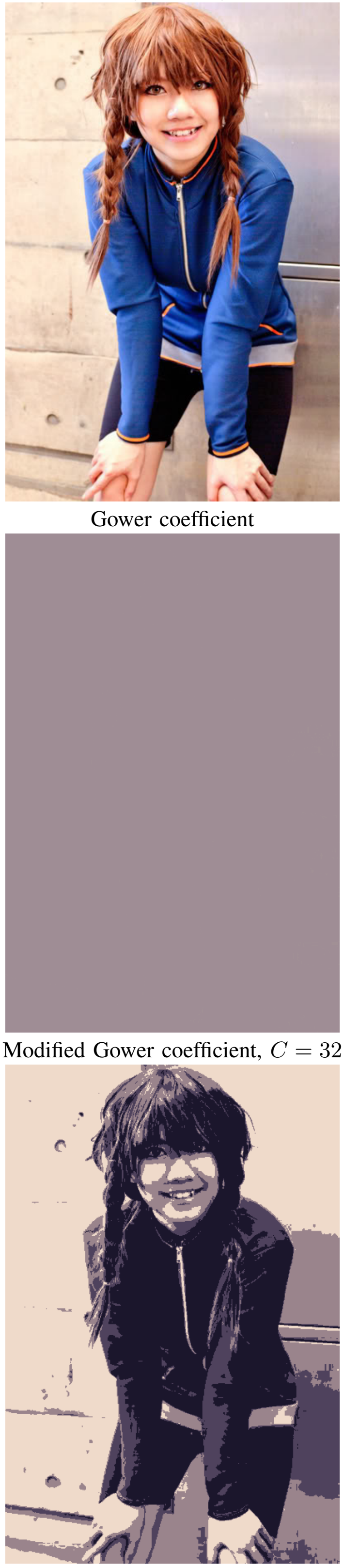
TABLE V
EXPERIMENT 2 RESULTS: C=16, MAXITNUM=200, DIST=EUCLIDEAN, SEED=RANDOM; AVERAGE, MINIMUM AND MAXIMUM; R - NUMBER OF OUTPUT NON-ZERO CLUSTERS

| | average | minimum | maximum |
|---|---|---|---|
| it. num. | 46.5 | 24 | 68 |
| r | 7.2 | 6 | 9 |
| F | 454.50 | 347.47 | 573.76 |
| Q | 7567.97 | 5299.01 | 9821.39 |
| s. time [s] | 107.315 | 55.26 | 155.97 |
| p. time [s] | 55.092 | 28.05 | 78.81 |

TABLE VI
EXPERIMENT 3 RESULTS: C=8, MAXITNUM=200, DIST=MANHATTAN, SEED=RANDOM; AVERAGE, MINIMUM AND MAXIMUM; R - NUMBER OF OUTPUT NON-ZERO CLUSTERS

| | average | minimum | maximum |
|---|---|---|---|
| it. num. | 18 | 12 | 26 |
| r | 3.8 | 3 | 6 |
| F | 1038.28 | 616.88 | 1531.17 |
| Q | 22551.60 | 11120.28 | 35771.00 |
| s. time [s] | 9.59 | 6.48 | 13.74 |
| p. time [s] | 5.156 | 3.53 | 7.29 |

TABLE VII
SEGMENTATION RESULTS - GOWER COEFFICIENT (SCALE = 0.33)

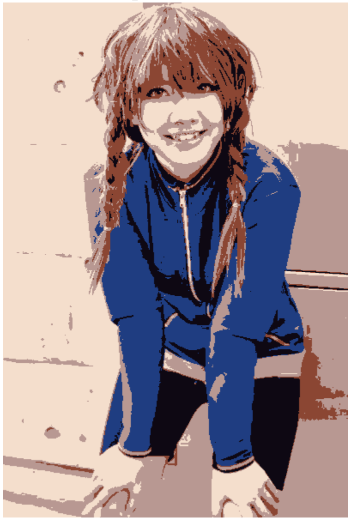
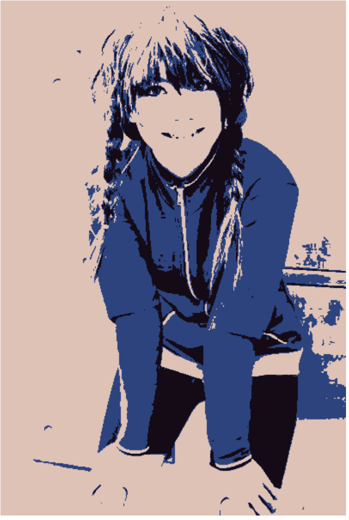
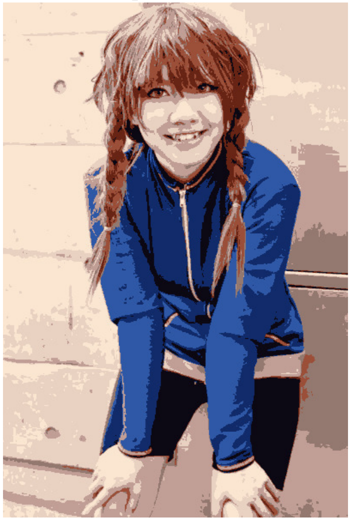
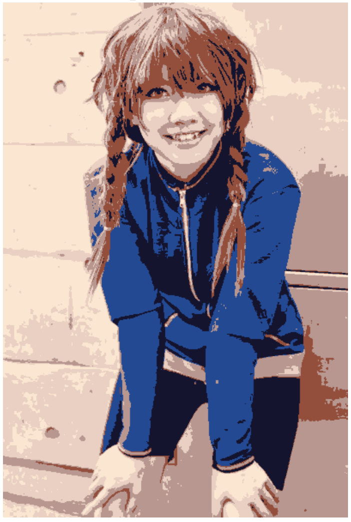
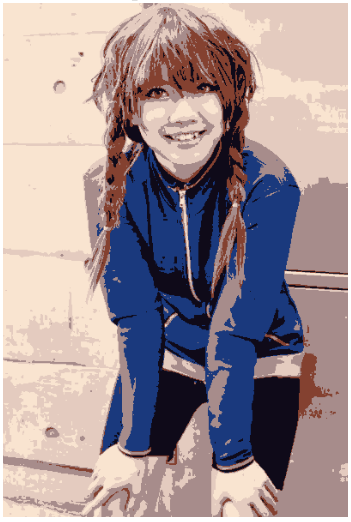
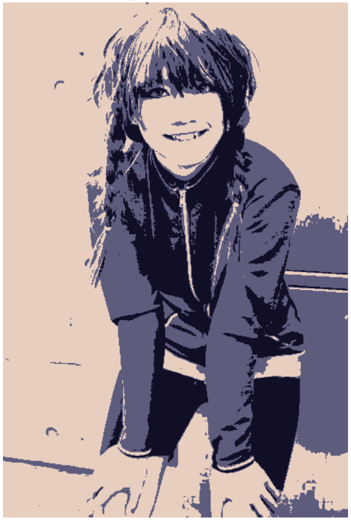


original image

Gower coefficient

Modified Gower coefficient, $C = 32$

TABLE VIII
IMAGES (SCALE = 0.33)

| original image | exp. 1 min $Q$ | exp. 1 max $Q$ |
|---|---|---|
| |  |  |
| original image | exp. 2 min $Q$ | exp. 2 max $Q$ |
| |  |  |
| original image | exp. 3 min $Q$ | exp. 3 max $Q$ |
| |  |  |

## References

[1] Cheng H.D., Jiang X.H., Sun Y., Wang Jingli. 2001. Color image segmentation: advances and prospects. Pattern Recognition, Volume 34, Issue 12:2259-2281, http://dx.doi.org/10.1016/S0031-3203(00)00149-7

[2] Correa C., Constantino V., Barreiro P., Diago M. P., Tardaguila J. 2011. A Comparison of Fuzzy Clustering Algorithms Applied to Feature Extraction on Vineyard. Inteligencia artificial revista iberoamericana de inteligencia artificial, Volume 1, Issue 1:778

[3] dos Santos T.R.L, Zarate L.E. 2015. Categorical data clustering: What similarity measure to recommend? Expert Systems with Applications, Volume 42, Issue 3:1247-1260, http://dx.doi.org/10.1016/j.eswa.2014.09.012

[4] Borsotti M., Campadelli P., Schettini R. 1998. Quantitative evaluation of color image segmentation results. Pattern Recognition Letters, Volume 19, Issue 18:741-747, http://dx.doi.org/10.1016/S0167-8655(98)00052-X

[5] http://inf.ug.edu.pl/˜osakarya

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1055–1058

DOI: 10.15439/2015F36
ACSIS, Vol. 5

# Pattern of the global syndrome for multiprocessor system of $H^4$ type for the MM model

Artur Arciuch
Military University of Technology
ul. Kaliskeigo 2, 00-908
Warszawa, Poland
Email: artur.arciuch@wat.edu.pl

***A problem of identification of faulty processors of a multiprocessor system is investigated. A method to reduce a pattern of a global syndrome for multiprocessor system which has a 4-cube topology, for the MM model, is presented. Results of the method for some topologies are also presented.***

## I. Introduction

In the paper a problem of identification of faulty processors of multiprocessor system is investigated. Identification of faulty processors is a problem which is analysed in many publications ([2]-[8],[12]-[15]). The process of identifying faulty processors in a system by analysing the outcomes of available inter-processor tests is the system level diagnosis. Faulty processors, beside fault-free processors, are involved in testing process. The basis of the system level diagnosis and original diagnostic model, namely the PMC model, were proposed by Preperata, Metze and Chien in [9]. An example of another diagnostic model is the MM model (comparison-base diagnosis model) [3], [2], [8]. PMC model and MM model assume that communication links between the processors are reliable (useable). In PMC model all tests are performed between two adjacent processors, and it was assumed that a test result is reliable (respectively, unreliable) if the processor that initiates the test is fault-free (respectively, faulty). In the MM model, the same job is assigned to a pair of processors of the network and their outputs are compared by a central observer. This central observer performs diagnosis using the outcomes of these comparisons. The comparison-based diagnosis model was extended [2] to allow comparisons carried out by processors themselves. In [7] authors proposed that comparisons have no central observer involved. The diagnosability of hypercube under the comparison-based diagnosis model were presented (i.a.) in [9].

A multiprocessor network (system), presented in the paper, belongs to class of the fault-tolerant system ([1]) and is mounted onto objects which are difficult to access. The network belongs to the class of self-diagnosable systems [7] as well. The multiprocessor system, in general, has a regular logical structure (e.g.: torus, hypercube) and is homogenous. A faulty processor in the system is not interchanged nor repaired. The faulty processor is removed from the logical structure of the network and access to it is blocked. If certain conditions are met the system with degraded structure continues to operate, and tasks of faulted processors are taken over by fault-free processors of the system (network) with degraded structure. The diagnosability of the network (system) is defined as the maximum number $t$ such that the network is self-diagnosable as long as the number of the faulty processors is not greater than $t$.

This paper focuses on selected issues connected with identifying of faulty processors of the 4-dimensional hypercube network and its node induced subgraphs ($H^4$ class for short) based on MM model and comparison diagnosis for such $t$-diagnosable system, where $1 \leq t \leq 2$. The rest of this paper is organised as follow. Section 2 gives some preliminaries; Section 3 focuses on a method of reduction of number of comparative trials. Section 4 concludes the paper.

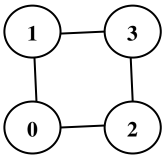
## II. Preliminaries

The processors network topology is represented by an undirected graph $G = \langle V, E\rangle$, where each node $u \in V$ denotes a processor and each edge $(u, v) \in E$ denotes a two-way link between nodes $u$ and $v$. In the paper graph $G$ is a 4-dimensional hypercube (4-cube for short) or nodes-included subgraph of 4-cube. A 4-cube ($H^4$) is such undirected graph $G = \langle V, E\rangle$, $|V| = 2^4$, $|E| = 4{\cdot}2^{4-1}$. Each node $u \in V$ is assigned an unique 4-bit binary vector (a coordinate) and each edge $(u, v) \in E$ links only those nodes whose coordinates differ in exactly one bit position (the Hamming distance between coordinates of linked nodes is equal 1).

Comparison diagnosis is based on inference of a network state on the basis of a set of results of comparative trials. Three processors are involved in the comparative trial: comparator $c \in V$, and comparative pair: $\{p_1, p_2\} \in V$: $\{p_1, p_2\} \subset V(c)$ ($V(c)$ is a set of nodes adjacent to $c$). A comparator $c$ instructs adjacent nodes $p_1, p_2$ to perform the same task and then checks to see if test results are the same. A $\psi = (c; p_1, p_2)$ is named a comparative trial.

978-83-60810-66-8/$25.00©2015, IEEE

a)



b)

| $\psi$ <br> $V^1$ | $d((0;1,2),V^1)$ | $d((1;0,3), V^1)$ | $d((2;0,3), V^1)$ | $d((3;1,2), V^1)$ |
|---|---|---|---|---|
| {∅} | 0 | 0 | 0 | 0 |
| {0} | *x* | 1 | 1 | 0 |
| {1} | 1 | *x* | 0 | *x* |
| {2} | 1 | 0 | *x* | 1 |
| {3} | 0 | 1 | 1 | *x* |

Fig. 1 A pattern of the global syndrome (b) for $H^2$ network (a)

Let $\Psi(H^4)$, $E(\psi)$, $K(\psi)$ and $P(\psi)$ denote respectively: a set of all possible comparative trials of $H^4$, a set of processors participating in the comparative trail $\psi$, a comparator of $\psi$, and a comparative pair of $\psi$.

A set $\Psi' \subseteq \Psi(H^4)$ is a comparative trial cover of set of processors if $P(\Psi') = V$. In other words a set $\Psi' \subseteq \Psi\ (H^4)$ is a comparative trial cover of set of processors if $\forall(e \in V)\ \exists(\psi \in \Psi')\ e \in P(\psi)$.

A diagnostic structure of $H^4$ based on comparative trails is such an ordered pair $\langle H^4, \Psi'\rangle$ $(\Psi' \subseteq \Psi(H^4))$, that a set $\Psi'$ is a comparative trials cover of set of processors of $H^4$.

Let $V^1$, $V^0$ and $d(\psi, V^1)$ denote respectively the set of faulty processors of the network, the set of fault-free processors of the network, and the result of comparative trial $\psi$ for set $V^1$, wherein $d(\psi, V^1) = 0$ denotes that test results of comparative trail for both processors are identical, and $d(\psi, V^1) = 1$ denotes that test results of comparative trail for both processors are different

The following rule of inference based on comparative trials is valid [2], [3].

$$[(K(\psi) \in V^0) \wedge (P(\psi) \cap V^1 = \emptyset)] \Rightarrow [d(\psi,V^1) = 0];$$
$$[(K(\psi) \in V^0) \wedge (P(\psi) \cap V^1 \neq \emptyset)] \Rightarrow [d(\psi,V^1) = 1]; \quad (1)$$
$$[(K(\psi) \in V^1)] \Rightarrow [d(\psi,V^1) = x;\ x \in \{0,1\}].$$

From formula 1 and Fig. 1 follows that if the network has faulty processors, then there exist several syndromes which faulty processors could produce. Let $\sigma(V^1)$ represents the set of syndromes which could be produced. Two distinct sets $V'$, $V'' \subset V$ are said to be indistinguishable if and only if $\sigma(V') \cap \sigma(V'') \neq \emptyset$; otherwise, $V'$, $V''$ are said to be distinguishable. Let $\sigma^*$ denotes a set of patterns of syndromes which faulty processors could produce, then from Fig.1 follows that i.e. $\sigma^*(\{1\}) \cap \sigma^*(\{2\}) \neq \emptyset$ which implies that there is no possibility to point out a faulty processor (for $t \geq 1$ ).

A processors network is one-step $t$-diagnosable by a set of comparative trials $\Psi' \subseteq \Psi(H^4)$ if each pair such sets $V'$, $V''$ $(|V'| \leq t, |V''| \leq t)$ of faulty nodes is distinguishable by at least one comparative trail $\psi \in \Psi'$.

*Theorem 1*[7]: For any $V'$, $V''$ where $V'$, $V'' \subset V$ and $V' \neq V''$ is a distinguishable pair if and only if at least one of following conditions is satisfied:

1) $(\exists(i,\ k \in V \setminus \{V' \cup V''\}) \wedge \exists\ (j \in \{(V' \setminus V'') \cup (V'' \setminus V')\})) \Rightarrow (\psi = (k;\ i, j) \wedge \psi \in \Psi')$,
2) $(\exists(i, j \in V' \setminus V'') \wedge \exists(k \in V \setminus \{V' \cup V''\})) \Rightarrow (\psi = (k\ ;\ i, j) \wedge \psi \in \Psi')$, (2)
3) $(\exists(i, j \in V'' \setminus V') \wedge \exists(k \in V \setminus \{V' \cup V''\})) \Rightarrow (\psi = (k\ ;\ i, j) \wedge \psi \in \Psi')$.

*Theorem 2*[7]: A system, is $t$-diagnosable if and only if each node has order of at least $t$ and for each distinct pair of sets $V'$, $V'' \subset V$ such that $|V'| = |V''| = t$, at least one of the conditions of theorem 1 is satisfied.

Let us note, for example, that, for sets $V' = \{0\}$, $V'' = \{3\}$ (graph Fig. 1a) and diagnosability $t = 1$ none of conditions of theorem 1 is satisfied.

For a processors network given by $H^4$ and $\langle H^4, \Psi'\rangle$ and for a set of nodes $X \subset V$, $T(X)$ denotes the set of those nodes in $V \setminus X$ which are compared to some nodes of $X$ by some nodes of $X$:

$$T(X) = \{j \in V \setminus X : \exists(\psi \in \Psi')\ \ \psi = (k\ ;\ i, j) \wedge i, k \in X\}.$$

*Theorem 3*[7]: A system with $n$ nodes is $t$-diagnosable if and only if

1) $n \geq 2t + 1$,
2) each node has order of at least $t$,
3) $\forall(0 \leq p \leq t - 1)\ \forall(X \subset V : |X| = |V| - 2t + p) : |T(X)| > p$.

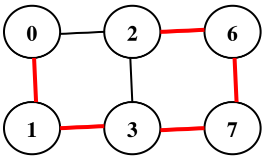
Let us note, for example, that for the graph on Fig. 1a for set $X = \{0,3\}$ and $t = 1$ theorem 3 is not satisfied.

On the basis of the above definitions and notation the problem of measurement of the multiprocessors system integrity is addressed in the next section.

## III. Reducing the number of comparative trails

The system diagnosability of $H^4$ class depends on orders of nodes and the number of nodes. It is known (*Theorem 3*) that diagnosability is not greater than the minimum order of the network node. Note that, for a processors networks having n nodes, maximum number of comparative trials is $\sum_{i=0}^{n-1}\binom{\mu(i) : \mu(i) > 1}{2}$, where $i$ is a node label and $\mu(i)$ is an order of $i$ node.

a)



b)

| $\Psi'$ / $V^1$ | (0;1,2) | (1;0,3) | (2;0,3) | (2;0,6) | (2;3,6) | (3;1,2) | (3;1,7) | (3;2,7) | (6;2,7) | (7;3,6) |
|---|---|---|---|---|---|---|---|---|---|---|
| {∅} | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| {0} | x | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| {1} | 1 | x | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| {2} | 1 | 0 | x | x | x | 1 | 0 | 1 | 1 | 0 |
| {3} | 0 | 1 | 1 | 0 | 1 | x | x | x | 0 | 1 |
| {6} | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | x | 1 |
| {7} | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | x |

Fig. 2. A pattern of global syndrome (b) computed for multiprocessor system (a)

*Corollary 1:* Path graph[11] of graph $G = \langle V,E\rangle$ which has at least 6 nodes ($|V| \geq 6$) describes an 1-diagnosable system under MM model.

Proof: We must show that $\forall(X \subset V : |X| = |V|\text{-}2) : |T(X)| > 0$ (*Theorem 3* point 3). If $(|V| \geq 6$ then $|X| \geq 4$, $|V \setminus X| = 2$ and $\exists(k; i, j) : k, i \in X \wedge j \in \mathrm{V} \setminus \mathrm{X} \wedge (k, i) \in E \wedge (k, j) \in E$.

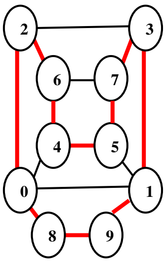
If a processors network of $H^4$ class is described by graph $G = \langle V, E\rangle$: $|V| \geq 6$ which has a Hamiltonian path[10] then the number of comparative trails for diagnosability of 1 is $|V| - 2$.

*Example 1:* Given cube $G' = \langle \mathrm{V}',E'\rangle$: $V' = \{0,1,2,3,6,7\}$, we want to find a pattern for global syndrome under MM model for diagnosability of 1. On Fig.2 is presented the entire pattern of global syndrome. Red edges (Fig. 2a.) and cells of table filled with red colour (Fig. 2b) presents the global pattern for Hamiltonian path of graph *G'* (after reduction).

*Corollary 2:* Hamiltonian graph[11] of graph $G = \langle V,E\rangle$ which has at least 10 nodes ($|V| \geq 10$) describes 2-diagnosable system under MM model.

Proof: We must show that that $(\forall(X \subset V : |X| = |V| - 4) : |T(X)| > 0) \wedge (\forall(X \subset V : |X| = |V| - 3) : |T(X)| > 1)$ (*Theorem 3* point 3). If $(p = 0$ and $|V| \geq 10$ then $|X| \geq 6$, $|V \setminus X| = 4$ and $\exists(k; i, j) : k, i \in X \wedge j \in \mathrm{V} \setminus \mathrm{X} \wedge (k, i) \in E \wedge (k, j) \in E$. If $p = 1$ and $|V| \geq 10$ then $|X| \geq 7$, $|V \setminus X| = 3$ and $\exists\Psi'':|\Psi''| >1$

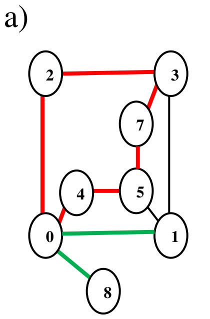
a)



b)

| $\Psi'$ / $V^1$ | (8;9,0) (9;8,1) (1;3,9) (3;1,7) (7;5,3) (5;4,6) (4;5,6) (6;4,2) (2;0,6) (0;2,8) | $\Psi'$ / $V^1$ | (8;9,0) (9;8,1) (1;3,9) (3;1,7) (7;5,3) (5;4,6) (4;5,6) (6;4,2) (2;0,6) (0;2,8) | $\Psi'$ / $V^1$ | (8;9,0) (9;8,1) (1;3,9) (3;1,7) (7;5,3) (5;4,6) (4;5,6) (6;4,2) (2;0,6) (0;2,8) |
|---|---|---|---|---|---|
| {∅} | 0000000000 | {1,2} | 001x11x100 | {4,5} | x1000001x1 |
| {0} | 100000001x | {1,3} | 10001x101x | {4,6} | x11x100001 |
| {1} | 01x1000000 | {1,4} | 01x11x1000 | {4,7} | x10001x101 |
| {2} | 00000001x1 | {1,5} | 00001x11x1 | {4,8} | x1001x1001 |
| {3} | 001x100000 | {1,6} | 001x1x1000 | {4,9} | x100001x11 |
| {4} | 000001x100 | {1,7} | 00001xx100 | {5,6} | x101x10001 |
| {5} | 00001x1000 | {1,8} | 1000001x1x | {5,7} | 1x1000001x |
| {6} | 0000001x10 | {1,9} | 01x1001x10 | {5,8} | 1xx1000000 |
| {7} | 0001x10000 | {2,3} | 0000001xx1 | {5,9} | 1x100001x1 |
| {8} | x100000001 | {2,4} | 001x101x10 | {6,7} | 1x1x100000 |
| {9} | 1x10000000 | {2,5} | 000001xx10 | {6,8} | 1x1001x100 |
| {0,1} | 11x100001x | {2,6} | 00001x1x10 | {6,9} | 1x101x1000 |
| {0,2} | 10000001xx | {2,7} | 1001x1001x | {7,8} | 1x10001x10 |
| {0,3} | 01x10001x1 | {2,8} | 01x1x10000 | {7,9} | 1x11x10000 |
| {0,4} | 101x10001x | {2,9} | 0001x101x1 | {8,9} | xx10000001 |
| {0,5} | 01xx100000 | {3,4} | 001xx10000 | | |
| {0,6} | 001x1001x1 | {3,5} | 0001x1x100 | | |
| {0,7} | 100001x11x | {3,6} | 0001xx1000 | | |
| {0,8} | 01x101x100 | {3,7} | 0001x11x10 | | |
| {0,9} | 000001x1x1 | {3,8} | x10000001x | | |
| | | {3,9} | x1x1000001 | | |

Fig. 3 A reduced pattern of global syndrome (b) computed for multiprocessor system (a)



b)

| $\Psi'$ / $V^1$ | (2;3,0) (3;2,7) (7;5,3)<br>(5;4,7) (4;5,0) (0;1,4)<br>(0;2,8) |
|---|---|
| {∅} | 0000000 |
| {0} | 10001xx |
| {1} | 0000010 |
| {2} | x100001 |
| {3} | 1x10000 |
| {4} | 0001x10 |
| {5} | 001x100 |
| {7} | 01x1000 |
| {8} | 0000001 |

Fig. 4 A reduced pattern of global syndrome (b) computed for multiprocessor system (a) which has not Hamiltonian cycle or a Hamiltonian path

$\Rightarrow(\forall(k;\, i,\, j)\in\Psi'' : k, i \in X \wedge j \in \mathrm{V} \setminus \mathrm{X} \wedge (k, i) \in E \wedge\ (k, j) \in E)$.

If a processors network of H4 class is described by graph $G = \langle V, E\rangle$: $|V| \geq 10$ is a Hamiltonian graph then number of comparative trails for diagnosability of 2 is $|V|$.

Example 2: Given cube $G' = \langle V', E'\rangle$: $V' = \{0,1,2,3,6,7,8,9\}$, we want to find a reduced pattern for global syndrome under MM model for diagnosability of 2. Fig.3b presentes the reduced pattern of global syndrome. Red edges (Fig. 3a.) present the Hamiltonian cycle of graph $G'$.

If a multiprocessor system $G'$ has no Hamiltonian path or is not a Hamiltonian graph then you should find longest path or longest graph cycle and add the missing nodes.

Example 2: Given cube $G' = \langle V', E'\rangle$: $V' = \{0,1,2,3,5,7,8\}$, we want to find a reduced pattern for global syndrome under MM model for diagnosability of 1. On Fig.4b is presented the reduced pattern of global syndrome. Red edges (Fig. 4a.) presents the longest path of graph G' and green edges present links to missing nodes.

## IV. Conclusion

The problem of reduction of number of comparative trials for MM model is complex. The paper only addresses the problem of generating a pattern of global syndrome for multiprocessor system of $H^4$ class under MM model and reduce the global pattern. Corollaries and examples presented in the paper shown the benefits of reducing a pattern of global syndrome. Other issues that should be considered are: the development of diagnostic procedures and the development of a test for a single processor system.

## References

[1] Kuhl J. G., Reddy S.M. 1980. Distributed fault-tolerance for large microprocessors systems. Proc.7th Symp. Comput. Architecturepp. 23-30, http://doi.acm.org/10.1145/800053.801905

[2] Maeng J., Malek, M. 1981. A Comparison Connection Assignment for Self-Diagnosis of Multiprocessor Systems. Digest Int'l Symp.FTC. pp. 173-175.

[3] Malek M. 1980. A Comparison Connection Assignment for Self-Diagnosis of Multiprocessors Systems, Proc. Seventh Int'l Symp. *Computer Architecture*, 1980, pp. 31-35, http://dx.doi.org/10.1145/800053.801906

[4] Friedman A. D., Simoncini L. 1980. System-level fault diagnosis, The Computer Journal, vol. 13, no. 3, pp. 47-53, http://dx.doi.org/10.1109/MC.1980.1653532

[5] Ishida Y., Adachi N., Tokumaru H. 1987. Diagnosability and distinguishability analysis and its applications, IEEE Transactions on Reliability, vol. 36, no. 5, pp. 531-538, http://dx.doi.org/10.1109/TR.1987.5222465

[6] Somani A. K., Agarwal V. K., Avis D. 1987. A generalized theory for system level diagnosis, IEEE Transactions on Computers, vol. 36, no. 5, pp. 538-546, http://dx.doi.org/10.1109/TC.1987.1676938

[7] Sengupta A., Dahbura A.T. 1982. On Self-Diagnosable Multiprocessors Systems: Diagnosis by the Comparison Approach, IEEE Trans. Comput., 41, 11, pp.1386-1396, http://dx.doi.org/10.1109/12.177309

[8] Chwa K. Y., Hakimi S. L. 1987. On fault identification in diagnosable system, IEEE Trans. Comput., C-30, 6, pp. 414-422, http://dx.doi.org/10.1109/TC.1981.1675807

[9] Gross J.T., Yellen J., 2006. Graph Theory and Its Applications, 2nd ed. Boca Raton, FL: CRC Press, http://dx.doi.org/10.1137/1.9781611970401

[10] Wilf H. S. 1994.Algorithms and Complexity, Summer, pp. 120-122, http://www.math.upenn.edu/~wilf/AlgoComp.pdf.

[11] Harary F. 1994. Graph Theory, Reading, MA: Addison-Wesley. 4-4.

[12] Zieliński Z., Kulesza R.2011.The life period of a 4-dimmensional cube processors' network diagnosed with the use of the comparison method, Biuletyn Instytutu Automatyki I Robotyki 17(30), pp. 17-32.

[13] Chudzikiewicz J.,Zieliński Z. 2010. "Reconfiguration of a processor cube-type network., Electrical Review, ISSN 0033-2097, R. 86 NR 9/2010, pp. 139-145.

[14] Chudzikiewicz J., Zieliński Z. 2014. On some resources placement schemes in the 4-dimensional soft degradable hypercube processors network, Advances in Intelligent and Soft Computing. Proc. of the Ninth Int. Conf. on Dependability and Complex Systems DepCoS-RELCOMEX (W. Zamojski at al. Eds., Series Ed.: Kacprzyk Janusz), Springer 2014, pp. 133-143.

[15] Malinowski T., Arciuch A.2014. The procedure for monitoring and maintaining a network of distributed resources, Annals of Computer Science and Information Systems, Volume 2, Proceedings of the FedCSIS 2014, pp. 947–954, http://dx.doi.org/10.15439/2014F159.

DOI: 10.15439/2015F144
ACSIS, Vol. 5

# Securing transmissions between nodes of WSN using TPM

Janusz Furtak
Military University of Technology
ul. Kaliskiego 2,
00-908 Warszawa, Poland
Email: jfurtak@wat.edu.pl

Jan Chudzikiewicz
Military University of Technology
ul. Kaliskiego 2,
00-908 Warszawa, Poland
Email: jchudzikiewicz@wat.edu.pl

*Abstract* — **Nowadays, Wireless Sensors Networks (WSN) are the most important components in the booming Internet of Things (IoT). Given the use of WSN in systems that are part of the critical infrastructure of a country, the primary task is continuous authentication of WSN nodes. This paper describes how to use the Trusted Platform Module (TPM) to authenticate sensors which create a sensors' domain in WSN. A model of wireless sensor network as well as operations associated with authentication in the sensors domain are presented. Additionally, an implementation of selected operations in the sensors domain is described; this includes: the master node initialization, slave nodes registration, and data transfer between them. Testing environment including the construction of nodes equipped with the TPM is described. The solution developed by the authors of the paper is only a partial realization of a broader concept of authentication in WSNs supported by the TPM.**

## I. Introduction

In the age of common electronic communications security plays an increasingly important role. This applies to a wide range of aspects of everyday life starting from household to complex control systems. Mainly results used for communication generally available network in order to minimize costs, increase the efficiency of data processing, as well as reducing access time to data. In this issue also part of Wireless Sensor Networks WSNs networks that have a wide application.

A WSN can be defined as a group of independent nodes, communicating wirelessly over limited frequency and bandwidth [1]. Execution of the tasks by the WSNs compared to typical sensor networks depends on dense deployment and coordination of sensors. Only the level of technology and human imagination are a limitation in applying of WSNs in any field of life. In a certain implementations of WSNs (e.g. in military areas) an ensuring of adequate level of security is required. In various implementations of WSNs confidentiality and reliability play an important role. A suitable level of confidentiality and reliability of data as well as security level against attacks can be achieved by applying a data encryption and an authentication of the nodes.

An attempt to design a secure WSN requires that the security components should be included into each node in the system. Any component of a network implemented without any security could easily become a point of attack. This means that security must permeate every aspect of design applications of wireless sensor networks that collect or disseminate sensitive information. Such solutions require a high level of safety.

Considering the military, police, emergency services or others, secrecy is part of their nature, so the data (sensed/disseminated/stored) are required to remain confidential. This is critical to the successful operation of a military, police, emergency applications. Enemy/threat tracking and targeting are among the most useful applications of wireless sensor networking.

Usually, the wireless sensor networks (WSNs) consist of large number of ultra-small, low-power and inexpensive wireless sensor nodes with sensing, computing and communication capabilities [2], [3]. It is assumed that such sensors must operate unattended for long periods of time such as several months or even years. In military applications, where the most important element is the safety, the times of maintenance-free operation are not most important, more that, often the life of the sensor will be limited to, for example, a few hours or days. In consequence, the power consumption may not be a critical parameter.

Security mechanisms deployed in WSNs should involve collaborations among the nodes due to the decentralized nature of the networks and absence of any infrastructure. The situation becomes critical when the nodes are equipped with cryptographic materials such as keys and other important data in the sensor nodes. Moreover, enemies/ adversaries can introduce fake nodes similar to the nodes available in the network which further leave the sensor nodes as un-trusted entities.

A characteristic property of WSN are limited resources of nodes creating the network. Attempting to implement an additional functions in such network is always a big challenge for designers of such a network. Introduction to WSN of any security mechanism is also the subject to this rule. Therefore, scientists have tried to offer various network security solutions tailored to the limited resources WSN. These proposals includes: secure and efficient routing protocols [1], [4], secure data aggregation protocols [5], [6],

978-83-60810-66-8/$25.00©2015, IEEE

[7], [8] and additional security mechanisms supported by Trusted Platform Module (TPM) [9], [10], [11], [12], [13], [14], [15]

A characteristic property of WSN are limited resources of nodes creating the network. Attempting to implement an additional functions in such network is always a big challenge for designers of such a network. Introduction to WSN of any security mechanism is also the subject to this rule. Therefore, scientists have tried to offer various network security solutions tailored to the limited resources WSN. These proposals includes: secure and efficient routing protocols [4], [1], secure data aggregation protocols [5], [6], [7], [8] and additional security mechanisms supported by Trusted Platform Module (TPM) [9], [10], [11], [12], [13], [14], [15].

Usually, the sensors used in military applications should be capable of being used for a relatively short period of time (e.g. several hours or days rather than months or years).The time is limited to the execution of a single task. In such situations typically, there is no restriction on energy consumption by sensor. Examples of such applications are shown in Fig. 1.

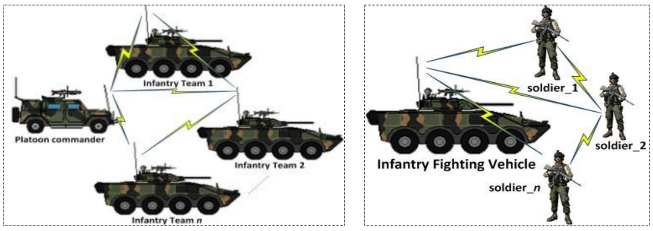


Fig. 1 WSN in military applications

Considering the above, the secure method of transmitting and storing data in WSNs is proposed in the paper. The Trusted Platform Module (TPM) is the basis of the presented method. A TPM is used for secure storing the necessary data to authenticate the nodes, and generate symmetric keys, and asymmetric keys (private/public). The solution presented in the paper uses the concept of authentication in WSNs using TPM developed by the authors of the paper and described in detail in [13].

In the second section proposed architecture of WSNs, and basic definitions are presented. The basic data of each node are stored in every nodes (type and scope of stored data depends on the role they played in the network e.g. domain master (node M), and slave (node S)). In the section the basic data structures used in the nodes are also defined. The third section shortly describes procedures to ensure proper authentication of sensors in domain and correct data transfer between sensors and in detail describes a certain operations in sensors' domain. In the fourth section a few experiments with selected operations in sensors domain and obtained results are presented. Finally, a few concluding remarks are presented.

## II. The model of Wireless sensor network with authentication[1]

In the domain of sensors there are two authorities. The first is the node (Data Collector) which is the recipient of the data emitted by the domain sensors. The node which manages the Root of trust is the second authority. The Root of trust is to be used to authenticate all sensors involved in the exchange of data between elements of the domain of sensors. The second authority is to act as a master of domain and will be called the node M. The presented concept assumes that both the role of the recipient of data from the sensors (i.e. Data Collector) and the role of the master of domain plays the same node.

In the sensors' domain is exactly the one node that acts as the domain master (node M). To this domain belong sensors of type slave (nodes S), which are registered by the node M. Nodes S are the source of data. Node S is initiated and authenticated by node M of domain. Node M stores the root trust of sensors' domain. The sensors' domain structure is shown in Fig. 2.

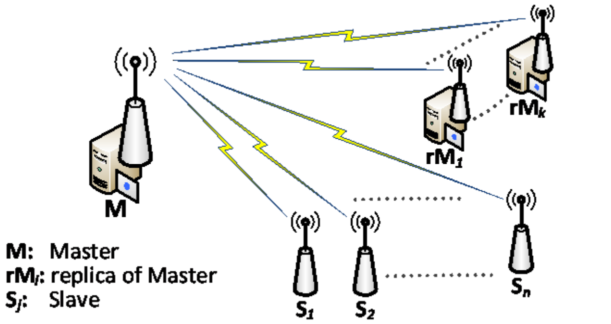


Fig. 2 The structure of sensors' domain

In the domain may be designated nodes acting as backups masters (replicas of master - rM). Such a node may be a S node after the establishment the role rM for him, on condition that its hardware and software resources provide this capability. In the domain may be no node type rM (this is not recommended), but there may be a few such nodes. The task of node rM is to store a copy of the root of trust from the node M of domain.

From the viewpoint of authentication procedures nodes M and rM for nodes S are the same. Node rM can become a new node M of domain after changing its role, due to proven inactivity of old node M. In this case the node, which has so far acted as a node M, becomes a node rM, or node S, or is removed.

When the sensor does not function, is turned off or damaged, it is assumed that this node is in a *non-active* state, and when the sensor is functioning, then the node is in the *active* state.

Sensor, which acts as a node M receives data from S nodes.

[1] The model of WSN with authentication and concept of authentication in such WSN was presented on Federated Conference on Computer Science and Information Systems, 2014 [13].

Minimum requirements for a sensor type S are as follows:

- sensor must be equipped with a TPM (see the next section);
- sensor must have an interface that allows direct connection to the node M (e.g. via USB or Serial) in the registration procedure of the node in the domain;
- the ability to send sensor data (i.e. measurement data) to M node using only wireless connection.

In order to enable automatic authentication procedure of the node and regeneration procedure for S node credentials, S node should be able to receive data transmitted by node M via a wireless connection. Otherwise, the node authentication procedure is not possible and change of credentials of this node will be possible only after the re-registration of the node. Nodes that are designed to play the role of M or rM must be able to bi-directional communication with other nodes, and should also have adequate resources in terms of power, processing capability and storage capacity.

### A. *Trusted Platform Module*

In the presented model for authentication sensors are used mechanisms offered by the Trusted Platform Module (TPM). It is assumed that each element of the domain of sensors is equipped with TPM.

TPM is an implementation of a standard developed by the Trusted Computing Group [16]. This module is designed to support the cryptographic procedures and protocols that can be used for securing data [17]. Trusted Platform Module provides the following functions:

- generating an asymmetric key pair,
- secure storage of keys,
- generating an electronic signatures,
- encryption and decryption,
- implementation of an operation defined by the standard PKCS #11.

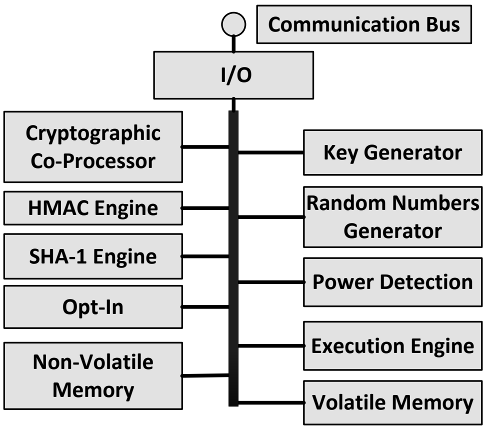


Fig. 3 TPM Component Architecture (based on [16])

The following algorithms are typically implemented in TPM [18]: RSA, SHA-1, HMAC and AES[2] [19]. In addition, each TPM chip stores a unique serial number and its RSA private key that is never available to read. TPM components are shown in Fig. 3. In laboratory stand was used TPM (AT97SC3205) developed by Atmel, which was designed in accordance with the security requirements for cryptographic modules (FIPS 140-2) Level 1 [20]. Used module additionally meets the requirements described in Security Policy for Atmel TPM [21], which says that authentication mechanisms meet the strength requirements of FIPS 140-2, Level 2.

### B. *Resources of sensors*

Each sensor is equipped with a TPM. The necessary data to authenticate the node in domain are stored in non-volatile memory of TPM. Access to the memory is protected by Endorsment Key of the module. The data structure of the node acting as the S is shown on Fig. 4[3]. Sensors, which are to play the role of M or rM must be equipped with additional memory, which is intended to store the description of the domain and descriptions of remaining domain nodes.

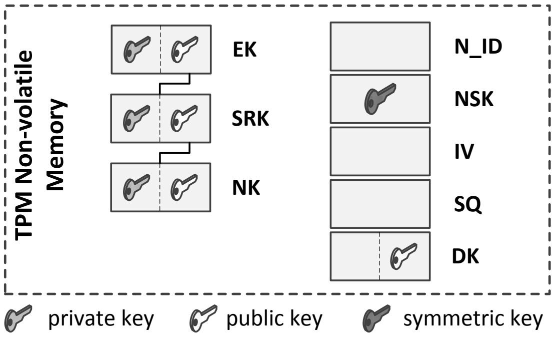


Fig. 4 The data stored on S node

Content of credentials stored in non-volatile memory of the TPM, which are used by a node S (Slave data):

- EK (*Endorsment Key*) - key pair (private/public) generated in the development phase of the TPM – the private part of the key never leaves the module and it is not possible to read this part of the key;
- SRK (*Storage Root Key*) - key pair (private/public) generated during the process of initiating the TPM in the procedure for registering a S node in the domain of sensors; private part of the key is wrapped by public part of EK, and access to the key is protected by secret of module owner;
- NK (*Node Key*) - key pair (private/public) of node; generated during the procedure for registering a S node in the domain of sensors; private part of the key is bound by public part of SRK;
- N_ID (*Node ID*) – ID of the sensor;
- NSK (*Node Symmetric Key*) – symmetric key to encrypt the data sent from this node to M node and to decrypt the data received from M node; obtained during the procedure for registering the node

[2] TPM uses a symmetric algorithm AES to protect the confidentiality of the session in which it participates. However, symmetric encryption functions are not normally accessible outside the TPM.

[3] The data shown on Fig. 5, Fig. 6 and further have been partially modified during the implementation of the method to those described in [13].

in the domain and renovated in the regeneration procedure of S node credentials;

- IV – initiating vector for encryption using NSK key in Cipher Block Chaining mode;
- SQ - the sequence number of the last sent frame (modified after each message);
- DK (*Domain Key*) – public part of the key of sensors' domain to which the node belongs; obtained during the procedure for registering the node in the domain.

Access to keys stored in non-volatile RAM is protected by the secret of the TPM module owner.

Credentials stored by the node M (the structure of the data is shown on Fig. 5) consist of three resources: Node description, Domain description and Description of domain nodes. The content of these resources is as the following:

- **Node description** - it is similar to the description of the node S but, instead the public part of the key DK, are stored both parts (i.e. public and private) of the DK key in Root of trust of M node and additionally the DK key is bound by NK key of M node;

  DK (*Domain Key*) – key pair (private/public) of sensors' domain; generated in the process of creating the domain of sensors and establishing the role of the "master" in the domain for the first node;

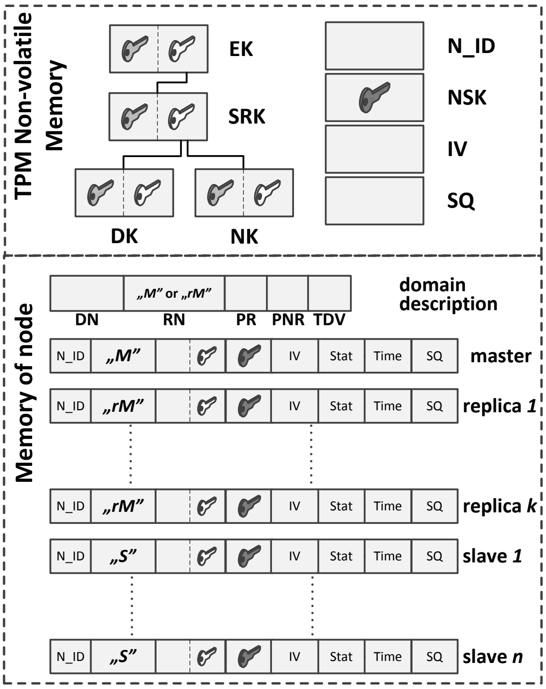


Fig. 5 The data stored on M node or rM node

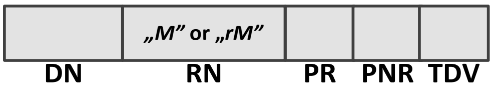


Fig. 6 The data structure describing a domain

- **Domain description** (the structure of the data is showed on Fig. 6):
  - DN (*Domain Name*) – the name of domain (up to 20 chars length);
  - RN (*Role of Node*) – determine whether below data are the resource of master node or the resource of replica of master; it is synonymous with the role it plays in the domain; may have one of values: M or rM;
  - PR (*Period of Replication*) – the time (in *msec*) after which the rM node is required to establish communication with the node M and refresh the domain data;
  - PNR (*Period of Non-success Replication*) - the time (in *msec*) after which the node rM is obliged to repeat the attempt to establish communication with the M node if the previous attempt refreshing the domain data was not successful;
  - TDV (*Time of data validity*) – after this time (in *msec*) and the inability to refresh, the domain data are invalid and node becomes a node S.

  All the data of domain description are encrypted using the NSK key and IV vector of M node.

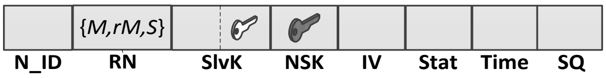


Fig. 7 The data structure describing a node

- **Description of domain nodes**. Description of each node contains the following data (the structure of the data is showed on Fig. 7):
  - N_ID (*Node ID*) – ID of the node (4 bytes length);
  - RN (*Role of Node*) – the role filled by the node in the domain; it can take values from the set {M, rM, S};
  - SlvK - public part of an asymmetric key N_ID node of sensors' domain;
  - NSK - symmetric key to encrypt the data sent from this node to M node and to decrypt the data received from M node; obtained during the procedure for registering the node in the domain and renovated in the procedure for the regeneration of S node credentials;
  - IV – initiating vector for encryption using NSK key in Cipher Block Chaining mode;
  - Stat - status of the node; it can take one of the values: *non-active(-1)*, *active(0)*, *active non-confirmed (n)*, where *n* is the number of consecutive unsuccessful attempts to establish communication with the node
  - Time (in *msec*) - moment of the last and the effective transmission [4];

[4] It was assumed that *Time* field is modified each time the field SQ is modified. In order not to complicate the understanding of the procedures outlined in the following sections, this field has not been included in these procedures.

- SQ - the sequence number of the last sent frame (modified after each message).

All node description data except N_ID field are encrypted using the NSK key and IV vector of M node.

Key EK, SRK, NK and DK form a root of trust node M. Access to keys from a root of trust and access to other data in the non-volatile memory is protected by the secret of the TPM module owner.

Because the description of the domain and descriptions of nodes are encrypted, they can be stored outside the TPM module non-volatile memory, for example in SD memory.

## III. Operations in the wireless sensor network with authentication

In [19] was presented the concept of authentication in WSNs using TPM. Ensuring proper authentication of sensors in domain and correct data transfer between sensors were taken into account in the concept. The concept consist the following procedures:

1. Procedure for initiating M node.
2. Procedure for registering the S node in the domain of sensors.
3. Procedure for removing rM or S node from the sensors' domain.
4. Authentication procedure of the node.
5. Integration test of nodes in sensors' domain.
6. Procedure for the regeneration of S node credentials.
7. Procedure of sending data from S node to M node.
8. Procedure of reading data on M node which were received from S node.
9. Procedure for giving role rM in the domain for S node.
10. Procedure for updating resources of rM node based on resources of M node.
11. Procedure for changing the node role from rM role to role M;
12. Procedure for determining the "new" node M after the failure of the "old" node M.
13. Integration test of resources of M and rM nodes.

In this study in the following sections the procedures listed in paragraphs 1, 2, 7 and 8 are comprehensively described. The procedures implementation details are described in the next section.

### A. *The procedure for initiating M node*

This procedure is intended to create the domain of sensors and to initiate the node that will play the master of the domain role.

Input data:

- M node owner secret;
- NK usage secret;
- DN - sensors' domain name;
- N_ID - node identifier;
- time periods (i.e. PR, PNR and TDV) associated with the operation of nodes rM..

The procedure for initiating M node comprises the following steps:

1. Take ownership of the TPM and SRK key generation.
2. Generate asymmetric key NK (NK attributes: Binding, Non-Migratable, Authority_always), SRK is a parent of NK);
3. Put NK into the root of trust stored in the TPM of "M" node.
4. Generate the data for M node:
   - generate asymmetric key DK for sensors' domain and put it into the root of trust stored in the TPM of "M" node (DK attributes: Storage, Migratable, Authority_always), SRK is a parent of DK; later public part of DK will be used by "S" node to bind the data which will be sent from "S" node to "M" node);
   - generate symmetric key (NSK – size 32 bytes) and initialization vector (IV – size 16 bytes) for AES cryptography;
   - generate sequential number SQ for M node;
   - put M node data into non-volatile memory of the TPM of S node.
5. Prepare of the domain description, which includes the fields DN, RN, PR, PNR, TDV and then encrypt this description using the NSK key and IV vector. The RN field should have a content of "M".
6. Prepare of the M node description and then encrypt this description using the NSK key, and IV. The fields of the description should have the following values:
   - N_ID = input data N_ID (the field is not encrypted);
   - RN = „M";
   - SlvK = public part of the node NK key;
   - NSK = the node NSK key;
   - IV = initiating vector for NSK key;
   - Stat = 0;
   - Time = current time;
   - SQ = random number from the range <0; 65535>.
7. Save the M node description in M node resources.

### B. *The procedure for registering the S node in the domain of sensors*

In the procedure of registration S node in the domain is assumed that during this procedure S node is connected to the node M via the Serial interface[5].

Input data:

- N_ID - node identifier;
- public part of the DK key.

After installing S node in serial port of M node the procedure for registering S node in the domain comprises the following steps:

1. On S node take ownership of the TPM and SRK key generation.

[5] If it was not possible to use the USB interface, in order to ensure the safety of the registration procedure, is required to develop additional ways of mutual authentication of both parties involved in the registration.

2. Generate asymmetric key NK of S node (NK attributes: Binding, Non-Migratable, Authority_always); SRK is a parent of NK).
3. Put NK to TPM resources of S node.
4. Generate the data for S node:
   - obtain the public part of the DK key from non-volatile memory of the TPM of M node; send a *dom_pub_key_req* packet from S node to M node through the serial line:

     ***dom_pub_key_req***

     | **code** | |
     |---|---|

     where:
     **code** = 101 for *dom_pub_key_req* packet;
     and get from M node a dom_pub_key_ans packet:

     ***dom_pub_key_ans***

     | **code** | **DK** |
     |---|---|

     where:
     **code** = 102 for *dom_pub_key_ans* packet;
     **DK** public part of Domain Key of sensors' domain;
   - generate symmetric key (NSK – size 32 bytes), initialization vector (IV – size 16 bytes) for AES cryptography;
   - put S node data into TPM non-volatile memory of S node.
5. Prepare S *node_description_req* packet

   ***node_description_req***

   | **code** | **N_ID** | **NSK** | **IV** | **NK** |
   |---|---|---|---|---|

   where:
   **code** = 103 for *node_description_req* packet;
   N_ID, NSK, IV and public part of NK key (the first three fields are bound using public part of DK key).
6. Transfer the blob to M node and then unbind it using the private part of DK key.
7. On M node prepare the S node description and then encrypt this description using NSK key and IV vector of M node. The fields of the S node description should have the following values:
   - N_ID = input data N_ID (the field is not encrypted)
   - RN = „S";
   - SlvK = public part of the S node NK key which be registered;
   - NSK = the NSK key of node which be registered;
   - IV = initiating vector for NSK key;
   - Stat = 0;
   - Time = current time;
   - SQ = random number from the range <0; 65535>.

   Save the S node description in M node resources.
8. Send a confirmation of registration to the node S. The confirmation should contain N_ID, Time and SQ and be encrypted using NSK key and IV vector of node S.

   ***node_description_ans***

   | **code** | **N_ID** | **Time** | **SQ** |
   |---|---|---|---|

   where:
   **code** = 104 for *node_description_ans* packet;
   N_ID, Time and SQ (the fields are encrypted using NSK key of S node).
9. Put SQ into TPM non-volatile memory of S node.
10. Uninstall the S node from serial port of M node

### C. *The procedure of sending data from S node to M node*

Input data:

- N_ID – identifier of node;
- SD – sensor's data;
- NSK – symmetric key of S node;
- IV = initiating vector for NSK key;
- DK – public part of domain key.

***sensor packet***

| **code** | **N_ID** | **SD** | **SQ** | **Hash** |
|---|---|---|---|---|

The structure of the frame containing the sensor data is showed above. It includes the following fields:

- code = 7 for sensor packet;
- N_ID = input data N_ID;
- SD = Sensor's Data encrypted using the NSK key and IV vector;
- SQ = current SQ incremented by 1;
- Hash = the value of the hash function determined on the basis of fields N_ID, SD and SQ.

The procedure of sending data from S node to M node comprises the following steps:

1. Preparing *sensor packet* containing the sensor data, as shown above
2. Binding the frame using the public part of DK.
3. Sending the frame to M node by XBee link.
4. Incrementing SQ field in resources of S node.

### D. *The procedure of reading data on M node which were received from S node.*

Input data:

- Received frame from S node;
- Resources of M node.

The procedure of receiving data on M node from S node comprises the following steps:

1. Receiving of the frame, as shown on Fig. 8.
2. Unbinding of the frame using the private part of DK.
3. Searching the description of N_ID node in description of domain nodes stored in protected resources of M node. If not a success, the N_ID node is unrecognized.
4. Comparing SQ field from received frame and SQ field from node description. If not equal, the SQ is incorrect.
5. Updating the description of N_ID node:
   - **stat** = 0;
   - **Time** = current time;
   - **SQ** = **SQ**+1.

6. Decrypting of the SD field using the NSK and IV acquired from description of domain nodes of “slave 1” node.

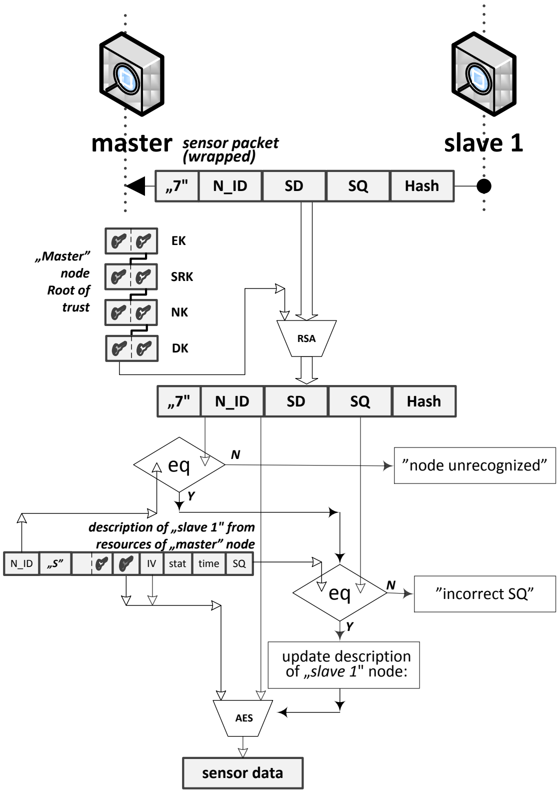


Fig. 8 The procedure of reading data on “master” node which were received from “slave 1” node

## IV. THE TESTBED TO EXAMINE AUTHENTICATION PROCEDURES IN WSN

The laboratory stand to examine the authentication procedures in WSN utilizing TPM was developed. The laboratory stand includes a few sensors equipped with TPM and several workstations to perform research. Block diagram of the sensor is showed on Fig. 9 and view of an exemplary sensor used in the experiments is showed on Fig. 10.

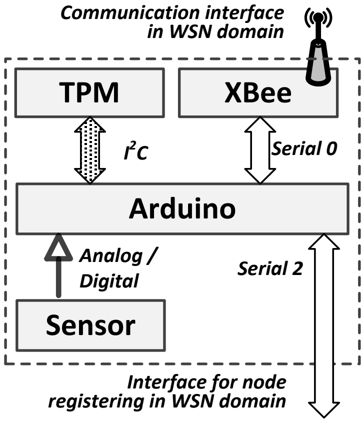

Communication interface in WSN domain
TPM
XBee
I²C
Serial 0
Arduino
Analog / Digital
Serial 2
Sensor
Interface for node registering in WSN domain

Fig. 9 Block diagram of the sensor

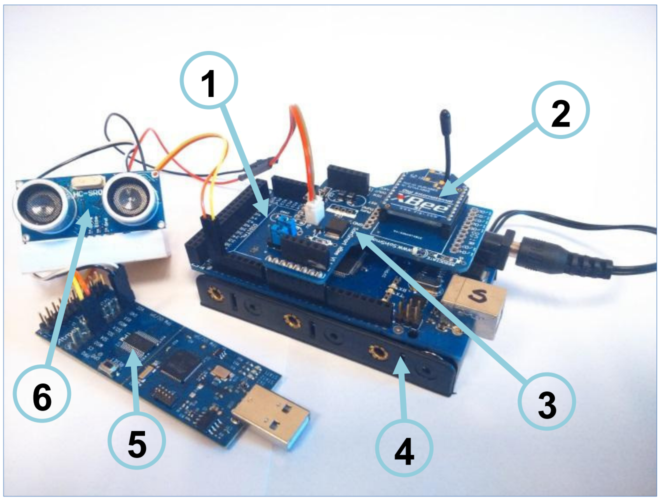


Fig. 10 View of an exemplary sensor

Sensor (showed in Fig. 10) used in the experiments was built with the following components:

- Arduino Mega2560R3 (in Fig. 10, indicated by 1) – based on microcontroller ATmega2560 (clock speed 16 MHz, 256 KB of flash memory for storing code (of which 8 KB is used for the bootloader), 8 KB of SRAM and 4 KB of EEPROM). The board has: 54 digital input/output pins (of which 15 can be used as PWM outputs), 16 analog inputs, 4 UARTs (hardware serial ports).
- XBee 1mW Wire Antenna Series 1 (indicated by 2) – wireless communication module with other wireless modules (compatible with the 802.15.4 standard). The module is connected to the Arduino Mega by adapter XBee Shield (indicated by 3) and communicates with Arduino by Serial 0.
- Ultrasonic distance sensor (indicated by 6) includes ultrasonic transmitters, receiver and control circuit. Provides 2cm - 400cm non-contact measurement function, the ranging accuracy can reach to 3mm.
- TPM (indicated by 5) – detachable part of hardware component of Atmel I$^2$C/SPI Demonstration Kit connected to Arduino through the I$^2$C Interface.
- Power bank (indicated by 4) – 9V power supply.

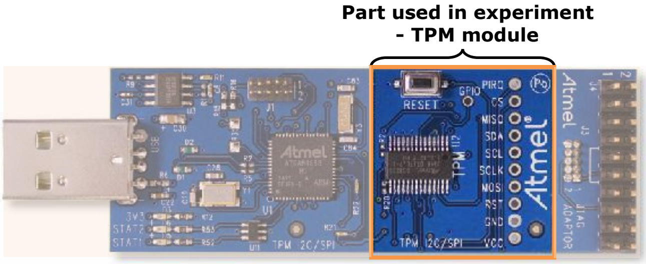


Fig. 11 Atmel I2C/SPI Demonstration Kit

In this laboratory stand was realized an experiment consisting of the following stages:

1. Initiating M node.
2. Registering the S node in the domain of sensors.
3. Transferring data from S node to M node:

a) sending a first frame (its structure is shown in subsection III.C) from node S to M as a plain text i.e. without encryption field SD and without binding frame with the public part of DK.
b) sending a second frame (its structure is the same) from S to M node containing encrypted SD field with NSK, but without binding the frame with the public part of DK;
c) sending a third bound frame with the public part of DK from S to M node containing encrypted SD field with NSK.

**STAGE 1.** The entire first stage is initiated and implemented autonomously on the node that will act as the Master. After this step this step the TPM is initiated and node ownership is acquired. The description of M node is written in non-volatile memory of TPM. Moreover, encrypted[6] description of sensors domain, in which is registered one node (i.e. Master), is created. Exemplary, encrypted description of sensors domain which was created as a result of this step for M node (node ID is 0xCC CC CC CC) is shown in Fig. 12.

```
--- Domain description
60 1C 38 44 45 82 63 85 51 5D 54 B4 1F 53 32 AC
0D 0E 20 DC 64 ED C5 C5 07 75 56 C9 27 62 D6 90
FD 69 21 98 B8 3B 2A CD 4D 48 AC FB 14 55 DD 5C

--- Descriptions of nodes - type: MASTER
--- Node ID    CC CC CC CC
--- Node description:
41 84 F7 D4 1A 69 FF 4B 0C 42 ED 0E 13 3D F7 76
72 4C 7C 8A 23 B2 52 F4 7E 4F AF D5 8A C9 9A 90
14 E0 2F DD 0E B1 70 9D F5 F3 4C 7F 9D F3 15 0C
AA D9 77 0D 64 7F 6C 23 F4 D6 3F F5 34 B5 1E 9D
6B 67 BB A7 33 C9 D8 7C 6C 27 5D 96 A7 06 83 F9
23 15 49 B5 A1 86 08 6C 06 ED 46 8A 73 5B B6 1B
11 BC 18 D8 FA E1 EF 21 6A A1 64 93 B4 08 03 DE
FC 9E 85 88 DF 71 56 52 B8 27 65 D3 89 44 DD 9E
D8 D5 96 3B 91 BE 52 B7 DB EE 40 F8 F4 19 55 A8
0E 6A 99 81 9A AB 2A 41 E0 07 A7 89 2C E6 01 C8
CA C2 3B 25 63 48 9A 97 6E 6F 46 88 E6 A9 54 F6
98 88 7F F3 4A A4 68 C0 E1 C3 05 F4 01 38 A7 5E
B5 4E 25 DF A5 8B 61 45 A1 1F 0B 3F B9 36 E2 67
07 17 2A EB F3 3A C3 2E D5 F5 38 B6 A5 E2 D7 00
52 0C 47 6A 5B 69 D6 E2 14 FC 55 DB 53 1A 1E 1D
9E 0C 18 2E 4D FE 69 BD 08 B2 7F E6 20 96 A9 0E
EC 66 C5 67 30 8F AD E3 71 B9 93 91 67 53 B8 83
91 09 2B 12 4E DD F6 4F FD 93 6C C8 A6 9B 2C 9F
D2 42 FA 4A C0 95 98 BB C8 F6 55 4D A2 B9 E2 61
58 58 FB F9 89 C8 51 DF 76 59 EF 6C F9 27 49 39
ED C6 92 B8 76 81 BF 6F F6 DA 16 0C 22 AB A0 D7
54 00 CD F6 5E A8 83 75 09 F3 AB 76 DC 37 C2 C6

--- Descriptions of nodes - type: REPL. of MASTER
    No nodes
--- Descriptions of nodes - type: SLAVE
    No nodes
```

Fig. 12 Description of sensors domain after M node initiating procedure

**STAGE 2.**

Before the start of the second stage S node should be connected to the node M over a Serial. The data shown have been partially modified during the implementation of the method to those described in [13], as shown in Fig. 13.

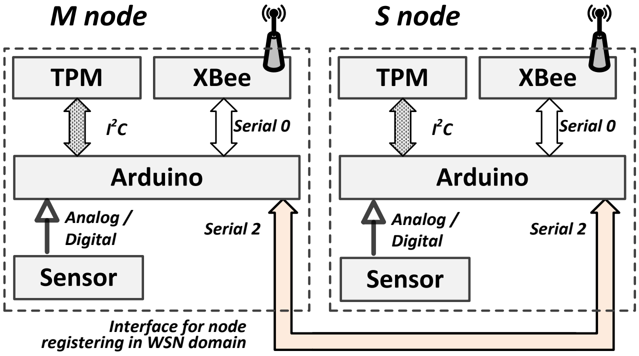


Fig. 13 Block diagram of M node and S node during the procedure of S node registering

```
--- Domain description
60 1C 38 44 45 82 63 85 51 5D 54 B4 1F 53 32 AC
0D 0E 20 DC 64 ED C5 C5 07 75 56 C9 27 62 D6 90
FD 69 21 98 B8 3B 2A CD 4D 48 AC FB 14 55 DD 5C

--- Descriptions of nodes - type: MASTER
--- Node ID    CC CC CC CC
--- Node description:
41 84 F7 D4 1A 69 FF 4B 0C 42 ED 0E 13 3D F7 76
72 4C 7C 8A 23 B2 52 F4 7E 4F AF D5 8A C9 9A 90
    |            |            |
    |            |            |
    |            |            |
ED C6 92 B8 76 81 BF 6F F6 DA 16 0C 22 AB A0 D7
54 00 CD F6 5E A8 83 75 09 F3 AB 76 DC 37 C2 C6

--- Descriptions of nodes - type: REPL. of MASTER

    No nodes

--- Descriptions of nodes - type: SLAVE
--- Node ID    01 01 01 01
--- Node description:
14 CE E3 D8 D1 E7 C0 6B 5B 19 D0 D6 20 87 57 88
CD DC EB 97 17 08 E8 BF 0F 00 4B D1 E7 6E 27 0D
    |            |            |
    |            |            |
    |            |            |
13 36 1F E6 9A 56 0B B6 3F EF A1 D9 89 98 13 9B
E2 5C 7E 9E 46 0B 37 C0 C2 2D AB 9C 25 C3 69 D9
```

Fig. 14 Description of sensors domain after S node registering procedure

In the first three steps of the stage TPM of S node is initiated, node ownership is acquired and the root of trust on S node is created. Then direct connection to M by Serial interface node is needed to transfer public part of the DK. DK is transferred as a plain text. In next step NSK and IV is randomly generated and put into non-volatile memory of S node. Then N_ID, NSK, IV and public part of NK key are bound using public part of DK key and transferred to M node through the Serial interface. On the basis of these data M node prepares a description of the node S and attach it

[6] In description of sensors domain all fields are encrypted with the exception of node IDs

to the sensors' domain description. Now domain description. After it the domain description might look like on the Fig. 14.

In the last step confirmation of S node registering (encrypted using NSK key of S node) is sent to S node. In this moment S node is registered and should be disconnected from Serial interface connecting it with M node.

### STAGE 3.

The S node is ready to transfer its sensor data by XBee interface – Serial line used in stage 2 is disconnected. In experiment takes part, in addition to S and M node, Observer station equipped with Xbee interface as shown on Fig. 15 and Fig. 16.This node is designed to interception the data transmission between nodes S and M.

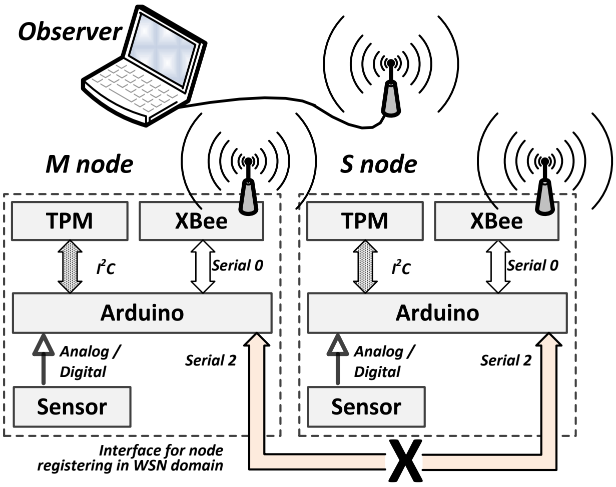


Fig. 15 Block diagram of M node, S node and observer during transferring data between S node and M node

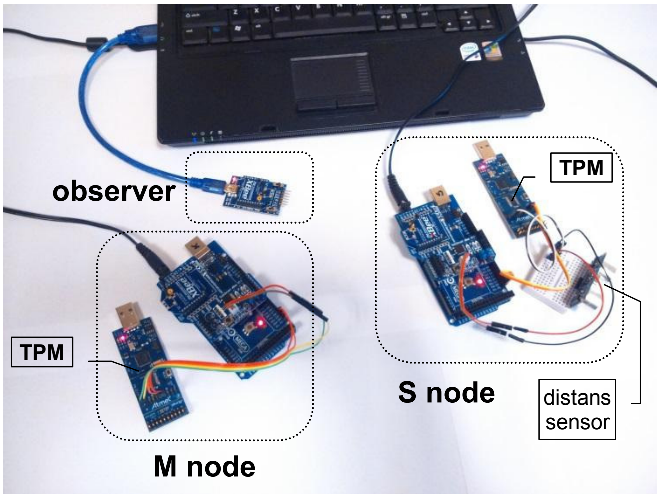
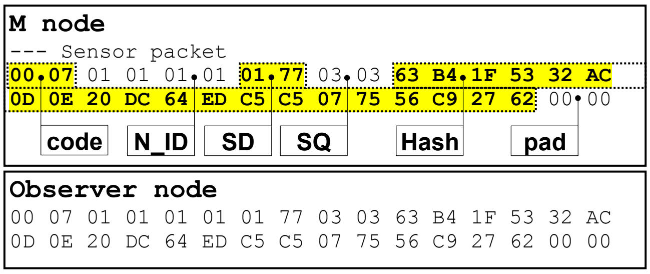


Fig. 16 View of the testbed during transferring data between S node and M node

Data received in step a) by nodes M and Observer should be the same - an example is shown in Fig. 17.

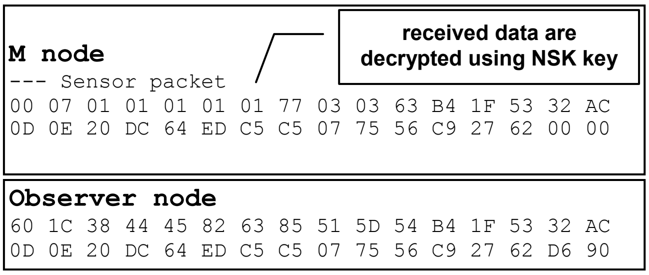


```
M node
--- Sensor packet
00 07 01 01 01 01 01 77 03 03 63 B4 1F 53 32 AC
0D 0E 20 DC 64 ED C5 C5 07 75 56 C9 27 62 00 00
```

```
Observer node
00 07 01 01 01 01 01 77 03 03 63 B4 1F 53 32 AC
0D 0E 20 DC 64 ED C5 C5 07 75 56 C9 27 62 00 00
```

Fig. 17 Data received on M node and on Observer node in step a)

Data received in step b) by nodes M and Observer are also the same, but for M node NSK key of S node is known and it can decrypt the SD field from received frame. The result is showed on Fig. 18.



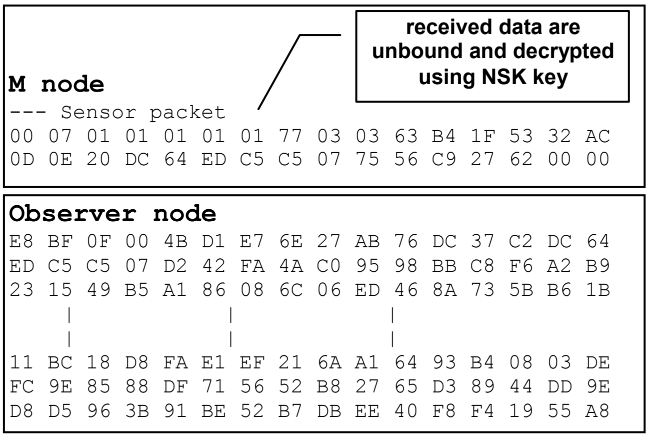
received data are decrypted using NSK key

```
M node
--- Sensor packet
00 07 01 01 01 01 01 77 03 03 63 B4 1F 53 32 AC
0D 0E 20 DC 64 ED C5 C5 07 75 56 C9 27 62 00 00
```

```
Observer node
60 1C 38 44 45 82 63 85 51 5D 54 B4 1F 53 32 AC
0D 0E 20 DC 64 ED C5 C5 07 75 56 C9 27 62 D6 90
```

Fig. 18 Data after step b)

Data received in step c) by nodes M and Observer are also the same, but M node knows NSK key of S node and private part of DK and M node can first unbind received frame and then decrypt the SD field from the frame. The result is showed on Fig. 19.



received data are unbound and decrypted using NSK key

```
M node
--- Sensor packet
00 07 01 01 01 01 01 77 03 03 63 B4 1F 53 32 AC
0D 0E 20 DC 64 ED C5 C5 07 75 56 C9 27 62 00 00
```

```
Observer node
E8 BF 0F 00 4B D1 E7 6E 27 AB 76 DC 37 C2 DC 64
ED C5 C5 07 D2 42 FA 4A C0 95 98 BB C8 F6 A2 B9
23 15 49 B5 A1 86 08 6C 06 ED 46 8A 73 5B B6 1B
    |               |               |
    |               |               |
11 BC 18 D8 FA E1 EF 21 6A A1 64 93 B4 08 03 DE
FC 9E 85 88 DF 71 56 52 B8 27 65 D3 89 44 DD 9E
D8 D5 96 3B 91 BE 52 B7 DB EE 40 F8 F4 19 55 A8
```

Fig. 19 Data after step c).

The experiment shows that the data transferred between nodes S and M are secured. Unauthorized nodes that are not registered in the domain of sensors, even if they are able to receive the data, they are not able to use them.

## V.CONCLUSION

This paper presents the model, concept of authentication in sensors' domain and implementation of securing transmissions between nodes of WSN. For this purpose, the mechanisms provided by the TPM are used. In paper was presented only the most important operations in sensors domain: nodes initiating and transfer data between the nodes. Particular attention was paid to secure the transmission and to secure the nodes of network. In all procedures hardware support provided by the TPM was used. If you apply all the requirements specified in the security Requirements for cryptographic modules (FIPS 140-2), the securing data is very strong. The effect is, however, come at a price relatively high power consumption and requires the use of modules that have more computing power and more resources of RAM. The biggest problem during the implementation was the shortage of sufficient RAM in used Arduino modules. For this reason, in further work we anticipate to use the EEPROM and/or SDRAM memory.

## References

[1] Boyle D., „Securing Wireless Sensor Networks: Security Architectures", Journal Of Networks, Vol. 3, No. 1, January 2008, pp.65-77.

[2] K. Sohraby, D. Minoli, T. Znati, „Wireless Sensor Networks Technology, Protocols, and Applications", Wiley, New Jersey 2007, DOI: 10.1002/047011276X.

[3] R. Faludi, "Building Wireless Sensor Networks", O'Reilly, 2011.

[4] A. Perrig et al., "SPINS: Security Protocols for Sensor Networks", Wireless Networks, vol. 8, no. 5, Sept. 2002, pp. 521–34, DOI: 10.1023/A:1016598314198.

[5] A. Al-Dhelaan, "Pairwise Key Establishment Scheme for Hypercube-based Wireless Sensor Networks", Recent Researches in Computer Science.

[6] Y Mohd Yussoff, H. Hashim, M. Dani Baba, "Identity-based Trusted Authentication in Wireless Sensor Network", International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, May 2012.

[7] L. Hu and D. Evans, "Secure Aggregation for Wireless Networks," Wksp. Security and Assurance in Ad Hoc Networks, 2003.

[8] B. Przydatek, D. Song, and A. Perrig, "SIA: Secure Information Aggregation in Sensor Networks," SenSys '03: Proc. 1st Int'l. Conf. Embedded Networked Sensor Systems, New York: ACM Press, 2003, pp. 255–65, DOI: 10.1145/958491.958521.

[9] W. Hu, H. Tan, P. Corke, W. Chan Shih, S. Jha, "Toward Trusted Wireless Sensor Networks", ACM Transactions on Sensor Networks, Vol. 7, No. 1, Article 5, August 2010, DOI: 10.1145/1806895.1806900.

[10] C. Krauß, F. Stumpf, C. Eckert, "Detecting Node Compromise in Hybrid Wireless Sensor Networks Using Attestation Techniques", Lecture Notes in Computer Science Volume 4572, Springer-Verlag Berlin Heidelberg 2007, pp. 203–217, DOI: 10.1007/978-3-540-73275-4_15.

[11] J. Furtak, T. Pałys, J. Chudzikiewicz, "How to use the TPM in the method of secure data exchange using Flash RAM media", Proceedings of the Federated Conference on Computer Science and Information Systems, 2013, pp. 831–838.

[12] Hu W., Corke P., Chan Shih W., Overs L., „secFleck: A Public Key Technology Platform for Wireless Sensor Networks", Wireless Sensor Networks, Lecture Notes in Computer Science Volume 5432, 2009, pp 296-311, DOI: 10.1007/978-3-642-00224-3_19.

[13] J. Furtak, T. Pałys, J. Chudzikiewicz, "The concept of authentication in WSNs using TPM", Position Papers of the Federated Conference on Computer Science and Information Systems, 2014, pp. 183-190, DOI: 10.15439/2014F176.

[14] Y. Wang, G. Attebury, B. Ramamurthy, "A survey of security issues in Wireless sensor networks", IEEE Communications Surveys & Tutorials, , Volume 8, No. 2, 2ND Quarter 2006, DOI: 10.1109/COMST.2006.315852.

[15] J. Sen, "A Survey on Wireless Sensor Network Security", International Journal of Communication Networks and Information Security, Vol. 1, No. 2, August 2009.

[16] *TPM Main Part 1 Design Principles*. Specification Version 1.2. Revision 116, Trusted Computing Group, Incorporated, 2011

[17] *TCG Software Stack (TSS)* Specification Version 1.2 Part1: Commands and Structures (http://www.trustedcomputinggroup.org /files/resource_files/6479CD77-1D09-3519-AD89EAD1BC8C97F0/TSS_1_2_Errata_A-final.pdf).

[18] S. Kinney, "Trusted platform module basics: using TPM in embedded systems", Embedded Technology Series,Elsevier Inc., 2006

[19] W. Stallings "Cryptography and network security principles and practice fifth edition", Prentice Hall, 2011, ISBN 13: 978-0-13-609704-4.

[20] Security Requirements For Cryptographic Modules. Federal Information Processing Standard (FIPS 140-2), National Institut of Standard and Techology, 2002-12-03. Retrieved 2013-05-18.

[21] Atmel Trusted Paltform Module AT97SC3204/ AT97SC3205 Security Policy FIPS 140-2, Level 1, Colorado Sprins, April 03, 2014.

[22] Q. A. Al-Haija, M. Al Tarayrah, H. Al-Qadeeb and A. Al-Lwaimi, "A Tiny RSA Cryptosystem Based On Arduino Microcontroller Useful For Small Scale Networks", International Symposium on Emerging Inter-networks, Communication and Mobility (EICM 2014), Procedia Computer Science 34 (2014) pp. 639 – 646, doi:10.1016/ j.procs.2014.07.091.

DOI: 10.15439/2015F160
ACSIS, Vol. 5

# Object Oriented Internet

Mariusz Postol
Technical University of Lodz
Institute of Computer Science
ul. Stefanowskiego 18/22, budynek
A14, 90-924 Łodź, Poland
Email: postol@zsk.p.lodz.pl

***Abstract*—The widespread use of the HTTP and hypertext makes it possible to freely publish new information and expose it in the context of its description. Unfortunately, this is a human-centric environment that cannot easily be adapted to an application-centric approach, which is required to provide distributed enterprise management and real-time process control. In this article new architecture is presented that can provide a generic solution for publishing and updating information in the context that can be used to describe and discover it. It is proposed to distribute the publisher (server) tasks to three classes: (a) information context management using the object oriented programming paradigm, (b) a predefined fixed set of services to access data and meta-data, and (c) a pluggable custom process data binding mechanism. It is also proposed to implement this architecture using the OPC Unified Architecture - a new emerging industrial integration standard.**

## I. Introduction

BEFORE commencing discussion about how to use the Internet the question “What is it?” must be addressed. Usually we can hear a definition like that: “the public worldwide computer network system that carries a vast array of information resources and services”. It seems to be too broad for further discussion. From the definition, it is networking technology providing access to information resources and services. “To get access”, the information transfer must be carried out between the Internet users, i.e. a resource or service provider and its consumers. To enable this, the following assertions must be made:

- Communicating parties must use the Internet protocol (IP) [1];
- All access points to the Internet communication infrastructure must have globally unique addresses;
- Users must be attached to those access points.

As long as the above rules are obeyed the communicating parties use Internet communication – colloquially, they are connected to the Internet. Reversing this sentence, we can say that the Internet is an infrastructure connecting any entities following the above principles. From the user point of view that is all, but to traverse information over the network the Internet infrastructure must be smart enough to locate the access points.

From any ICT solution we would expect information processing rather than having only interconnection between the Internet access points. To meet this requirement the Internet users must be processing engines rather than hosting applications responsible for this work. Application to application connectivity is, therefore, required that can be provided by an additional transport protocol. Examples of such protocols are TCP (connection oriented) and UDP (connectionless). It is worth stressing that we can use a variety of protocols as the transport protocol and the above assertions still hold true.

On the other hand, any user expecting information processing from the communicating party is interested in selecting appropriate functionality, but not a particular application instance. Therefore, we can distinguish three meanings of the transport protocol address, called a port:

- Functionality selector – for a consumer interested in utilizing a particulate information resource or service;
- Functionality publication end point – for servers offering resources or services;
- Address to identify sending and receiving application end-points – for the protocol stack.

The TCP and UDP protocols share the same address space with a capacity of 64k end points. Even today, having so many applications hosted on any network node is impractical and hard to manage. Unfortunately, mapping between functionalities and their identifiers is static, which means that the majority of available port numbers are globally unique functionality identifiers governed by the Internet Assigned Numbers Authority (IANA). The others called dynamic/private ones are not assigned and used to identify sending and receiving application end-points only.

Using a global dictionary instead of a description, discovery and integration mechanism results in a fully exhaustion of transport protocol address space and the registration of new functionality becomes very difficult. Therefore, to communicate over the Internet, users need to select one from the 49152 existing options.

The selection should be based on well-defined requirements, but how to define the requirements having only the general assumption that we expect access to the

978-83-60810-66-8/$25.00©2015, IEEE

information resources or services. Many aspects may be taken into consideration. For the above assumptions a solution that allows server to freely publish new resources and services is needed. A globally acceptable discovery and integration mechanism is a possible option. Alternatively, another protocol must be selected on the following assumptions:

- The publishing server is responsible for managing the address space;
- The protocol provides an infinite address space capacity;
- The protocol is transparent for the payload transported.

The above assumption make the Internet a publication platform containing countless resources, but, to be useful, consumers must be allowed to find the appropriate ones using a description and discovery mechanism. It requires that publisher must provide additional information (meta-data), which describes the resources to allow the selection. Additionally, the descriptions must be coupled with addresses to selectively access them. For human-centric solutions a graphical interface is an appropriate mechanism. HTTP [1] as a protocol and HTML [2] (more general a hypertext) as a description language are the big winners selected by millions of people and they have led to the establishment of the World Wide Web.

Unfortunately, for application to application connectivity a programming interface (API) is required. Because community acceptance and reuse of the existing solutions is so important for the Internet evolution, a new solution – called web services - atop HTTP has been developed by the World Wide Web Consortium (W3C) [2]. This specifications suit is commonly referred to as WS-* and contains:

- Simple Object Access Protocol (SOAP) to use the services;
- Web Services Description Language (WSDL) to describe the services;
- Universal Description, Discovery and Integration (UDDI) to get access to the services description.

To obtain applications interoperability all clients consuming services provided by a server offering them must conform to a WSDL specification prepared in advance that defines a contract between them. Hence, this process requires software development and, because WSDL cannot provide complete semantics of the service, the process is usually manual and requires conformance testing. There are no good global scope solutions of this issue. Today solution is for the server publisher organization to provide complimentary compliant client applications. In this approach, typical problems like operating system dependence, software updating and versioning must be solved. Finally, it leads to a static solution where functionality is exposed as a fixed set of services.

It seems that the next level of abstraction is needed to meet the above mentioned goal and allow the server to freely publish resources and services. Generally speaking, all ICT systems are expected to provide information processing capabilities. Information is an abstract knowledge; it cannot, therefore, be directly processed by physical machines. To make information capable of being processed, it must be represented as computer-centric binary data. To propose a solution that meets those requirements two questions should be addressed:

- How to get access to (transport) the process data?
- How to represent (model) the information?

To answer the first question we need a globally accepted, platform-neutral (assuring that the above stated assertions hold true) communication standard that allows also addressing the second question, i.e. designing of an appropriate Information Model.

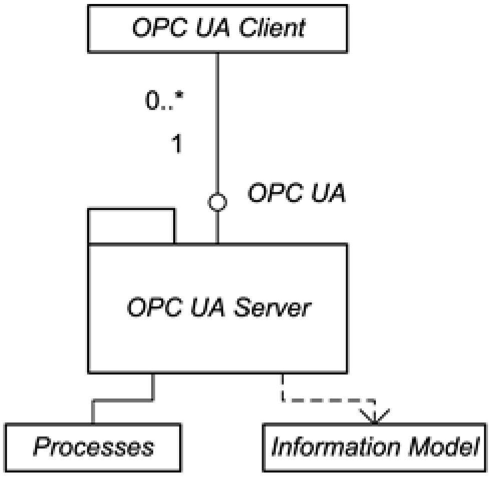


Fig. 1 OPC Unified Architecture archetype

The OPC Unified Architecture (UA) (Fig. 1) technology [3], [4], [5], [6] meets all the requirements, because:

- It is Internet based technology;
- It is a platform neutral standard allowing easy implementation on any system including embedded systems;
- It is designed to support complex data types and object models;
- It is designed to achieve high speed data transfers using efficient binary protocols;
- It is scalable from embedded applications up to the process control and enterprise management/operation systems;
- It has broad industry support and is being used in support of other industry standards such as ISA S95, ISA S88, EDDL, MIMOSA, OAGiS, etc. [7].

It is a broad class of application domains where business IT and control systems are converged in a global scope to make a large whole with the aim to improve performance as the result of the macro optimization and synergy effect. One of the main requirements of the Industrial ICT is to provide a consistent mechanism for the integration of the vast varieties of systems. This requirement can be met as the result of employing the OPC Unified Architecture (UA) as the mechanism for the integration. It is assumed that it should be robust and the implementation should be platform independent. Fig. 1 illustrates the architecture of the proposed solution. In this approach three elements are distinguishable from the typical client server archetype:

- *OPC UA*: an interface representing invariable Service Model [8] responsible for providing client/server connectivity;
- *Information Model*: application domain unique description of a context the process data is made accessible in to the clients.
- *Processes*: source of exposed information resources and services hereinafter referred to as process data.

To make systems interoperable, the data transfer mechanism must be associated with a consistent information representation model. OPC UA uses an object as a fundamental notion to represent data and activity of underlying processes (see Sec. II.B). The objects are placeholders of variables, properties, events and methods and are interconnected by references. This concept is similar to well-known object oriented programming (OOP) that is a programming paradigm using "objects" – data structures consisting of fields, events and methods – and their interactions to design computer programs [9]. The OPC UA Information Model [10], [11] provides features such as data abstraction, encapsulation, polymorphism, and inheritance.

The OPC UA object model allows servers to provide type definitions for objects and their components. Having defined types in advance, clients may provide dedicated functionality, for example: displaying the information in the context of specific graphics.

The OPC UA information modeling concept (Sect. III) is based on layers, which step by step expand the basic model provided by the OPC UA Specification [19]. The Information Model is abstract and hence, in a real environment, it must be implemented in terms of bit streams (to make information transferable) and addresses (to make the data selectively available). To meet this requirement, OPC UA introduces a node notion as an atomic addressable entity that consists of attributes (value-holders) and references (address-holders of coupled nodes). The set of nodes that an OPC UA server makes available to clients is referred to as its Address Space [4], [5], [12], which enables representation of both underlying processes environment and its behavior. The Address Space exposed by the server makes up a context the process data is made available in to the clients (Fig. 1). Creation of this context (Sec. IV) depends on an application domain unique Information Model.

*Processes* in Fig. 1 represents a class of functions responsible for getting access to business or industrial processes source of exposed information resources and services hereinafter referred to as process data.

Basing on the defined typical enterprise systems structure and requirements [25], a new architecture is proposed (Sect. IV) where an intermediate component called Process Observer is proposed. The model allows for significant reduction of the solution complexity, but the implementation of this model proves that the architecture additionally could increase robustness by adding redundancy. The example (sect. V.B), where the presented model has been used, shows that the approach can be a platform for multi-enterprise collaboration to benefit from synergy effect and macro optimization.

Sect. III, III.C and IV describe novel architectural proposal and corresponding communication algorithms allowing building robust real time distributed systems. A case study where the presented solutions have successfully been implemented is in Sect. V.

## II. OPC Unified Architecture Key Features

### *A. Service Oriented Architecture*

At the very beginning of a new solution development the question about its fundamental paradigms and architecture must be addressed. Observing continuous evolution of the ICT domain, it seems that finding a solution that will guarantee an unlimited lifetime is a real challenge. However, decoupling the solution from any base technology increases the chance of its surviving the disappearance of the base technology from the market. Fortunately, as mentioned above, there are many options on how to get applications interconnected over the Internet. Developing services and deploying them using Service Oriented Architecture (SOA) is the best way to utilize ICT systems to meet this challenge. A service differs from an object or a procedure because it is defined by messages that it exchanges with other services. SOA defines the way in which services are deployed and managed. Adopting of the SOA approach increases reuse, lowers overall cost, and improves the ability to rapidly change and evolve systems, whether old or new.

To make systems interoperable, any even brilliant idea is not enough - a data transfer technology is needed, however – when defining data exchange in context of messages – we do not need to bother about different technologies used by the participants as long as they can absorb the messages.

Today, an ideal platform for the SOA concept implementation is Web Service technologies. Web Services are a set of standards based on XML (eXtensible Markup Language) and developed by W3C (World Wide Web Consortium) [2] marked with a WS-* symbol. Because the WS-* standards are developed without any initial assumption concerning the underlying system platform they are implemented on, they therefore must precisely define what must be on the "wire".

The WS-* standards are the basic foundation for OPC UA but, using them alone, would not be enough to reach the expected data throughput performance in industrial applications. To promote scalability, the OPC UA suite of protocols, therefore, expands the WS-* standards by defining a few proprietary ones that can be used alternatively. OPC

UA messages may be encoded as an XML text or in binary format for efficiency purposes.

### *B. Object Oriented Information Model*

OPC UA uses an object as a fundamental notion to represent data and activity of an underlying processes system. The objects are placeholders of variables, events and methods and are interconnected by references. This concept is similar to well-known object oriented programming (OOP) paradigm [9]. The OPC UA Information Model [4], [5], [10], [11] provides features such as data abstraction, encapsulation, polymorphism, and inheritance.

The OPC UA object model allows servers to provide type definitions for objects and their components. Type definitions may be abstract, and may be inherited by new types to reflect polymorphism. They may also be common or they may be system-specific. Object types may be defined by standardization organizations, vendors or end users. Each type must have a globally unique identifier that can be used to provide description of the information meaning, i.e. semantics from the defining body or organization. Using the type definitions to describe the information exposed by the server allows:

- Development against type definition;
- Unambiguous assignment of the semantics to the data expected by the client.

### *C.Abstraction and Mapping*

Interoperability of applications can be achieved if communicating parties are able to interchange streams of bits and assign to these streams the same meaning without any ambiguity. Unfortunately, the representation of information on the wire, and communication protocols are subject to continuous evolution, if not revolution nowadays. This could be dangerous for any long term initiatives. Because it is impossible to stop the progress of technology changes, some other precautions must be taken to keep the specification alive within a long term horizon. It is achieved by clear separation of definitions provided by the specification from their actual implementation. It makes OPC UA seamlessly portable from one technology to another. Mappings defined in the specification [13] set forth how to implement an OPC UA feature using a specific technology.

### *D.Security*

Security is the fundamental aspect of computer systems, in particular those dedicated to enterprise and process management. Especially in this kind of applications, security must be robust and effective. Security infrastructure should also be flexible enough to support a variety of security policies required by different organizations. OPC UA may be deployed in diverse environments; from clients and servers residing on the same hosts, throughout hosts located on the same operation network protected by the security boundary protections that separate the operation network from external connections, up to applications running in global environments using public network infrastructure. Depending on the environment and application requirements, the communication services must provide different protections to make the solution secure [14].

OPC UA Security is concerned with the authentication of clients and servers, the authorization of users, the integrity and confidentiality of their communications and the auditing of client server interactions. To meet this goal, security is integrated into all aspects of the design and implementation of OPC UA servers and clients.

OPC UA relies upon the site cyber security management system to protect confidentiality on the network and system infrastructure, and utilizes the public key infrastructure to manage keys used for symmetric and asymmetric encryption [15]. OPC UA uses symmetric and asymmetric encryption to protect confidentiality as a security objective, as well symmetric and asymmetric signatures to address integrity as a security objective.

### *E. Profiles*

OPC UA is designed to support integration of wide range of servers, from plant-floor control devices to enterprise management and operation systems. All of them are characterized by a variety of performances, execution platforms and functional capabilities. Therefore, OPC UA defines a comprehensive set of capabilities servers may implement a subset of. These subsets are referred to as *Profiles*, and servers may claim conformance to them.

### *F. Robustness*

Because it is to be used in the production environment including real-time process control applications, OPC UA is designed especially to provide robustness of the remote access to the underlying process data. OPC UA provides mechanisms for clients to quickly detect and recover from communication failures associated with transfers without having to wait for long timeouts provided by the underlying protocols [16].

## III. Information Model

### *A. Concept*

The primary objective of the OPC UA server is to expose information resources and services, which then can be used by clients to manage an underlying real-time process or the entire enterprise as a large whole with the main challenge of integrating systems and management resources into one homogenous environment. Information describes the state and behavior of the processes and the server must be able to transfer it in both directions. The main challenge of the OPC UA Information Model is to support this transfer by a unique and transparent means in spite of the process complexity and roles of clients in the enterprise management hierarchy.

Information is an abstract knowledge; therefore it cannot be directly processed by physical machines. To make information capable of being processed, it must be represented as the binary data. To define the relationship between information and binary data on the one-to-one basis, syntax and semantics are needed. Syntax defines rules of the vocabulary usage, and semantics maps valid bits pattern to the associated piece of information.

An Information Model for OPC Unified Architecture is such a collection of vocabulary, syntax and semantics. This collection plays a role similar to high level programming languages that describe data structures and an algorithm to be executed by the processor.

Information exposed by the OPC UA server may be complex. Clients may, therefore, want to obtain the information definition. Generally speaking, to select a particular target piece of information we have two options: random access or browsing. Random access requires that the target entity must have been assigned a globally unique address and the clients must know it in advance. We call them well-known addresses. The browsing approach means that the clients walk down available paths that build up the structure of information. For example hypertext document containing URL's locating recursively hypertext documents and other resources.

It seems that, in spite of the access method, we have to assign an address to all of the accessible items in the representation of the information structure. Therefore we call the collection of these items the Address Space [4], [5], [12]. This atomic addressable item is called a node. Each node is a collection of predefined set of attributes that have values accessible locally in context of the node. To represent information about the internal structure, nodes are interconnected by references.

Accessing information by clients is the first aspect of controlling the data stream between the clients and the underlying process environment of the server. Another one is creating and maintaining the Address Space in real-time.

To create the Address Space, we need to instantiate nodes and interconnect them by references. Instantiating nodes operation requires assigning appropriate values to attributes and adding references. To make information internally consistent as a large whole, we need rules governing the creation and modification processes. The Information Model implies these rules using the following two concepts:

- ***NodeClass*** – as a formal description of the node defining the allowed attributes and references;
- ***Type*** – as a formal description of the node defining the allowed attributes and references values.

For the client to understand the Information Model, it must be predefined or exposed.

Available NodeClasses are predefined, i.e. the specification provides a strictly defined non-extensible set of NodeClasses. Each one is assigned a dedicated function, e.g. *Variable* NodeClass defines nodes that provide a value, and *Method* NodeClass represents a function.

Like the NodeClass concept, the specification provides a set of predefined types, which is extensible. According to the above rule, all not predefined types must be exposed in the Address Space. To expose predefined and proprietary type definitions in the Address Space, there are dedicated NodeClasses, namely *ObjectType*, *VariableType* and *ReferenceType*. For example, nodes of the *VariableType* NodeClass provide clients with definitions of types derived from the *BaseVariableType* that is a base type for all variables. The main role of the types represented by the above NodeClasses is to provide a description of the Address Space structure and to allow clients to use this knowledge to navigate to desired information resources (represented by the *Variable* nodes) and services (represented by the *Method* nodes) in the Address Space exposed by the OPC UA server.

DataType NodeClass is also dedicated to describe types. In this case, the represented types have a special mission, because they describe underlying process data that client has access to using a connection to the OPC UA server. For example, a node of the DataType can provide information to clients that the data has a numeric value and the clients reading it can use this knowledge to interpret and process the obtained value.

Types are called metadata since they describe the data structure (context) not the actual data values.

Even though the OPC UA specification contains a rich set of predefined types, the type concept allows designers to freely define types according to the application needs. New types are derived from the existing ones. The derived types inherit all features of the base types but can include modifications to make the new types more appropriate for information that is to be represented.

The Address Space concept based on types can be a foundation for exposing any information that is required. Clients understand the Address Space concept and have a browse service to navigate through the Address Space. Since browsing is based on the incremental and relative passage among nodes it is apparent that each path must have a defined entry point, so the question as to "where to start" must be addressed. To meet this requirement, the Address Space must have a predefined template containing well defined nodes that can be used as anchors from which a client can start browsing the Address Space content. Thus to design an Address Space and define new types, they must be derived from the existing ones. At the very beginning the only existing types are the standard ones defined by the specification. The available standard types are briefly described in the Section III.B.

### *B. Standard Information Model*

The primary objective of the OPC UA Address Space is to provide a standard way for servers to represent objects to the clients. The Object NodeClass is used to define objects. Each object in the Address Space has an assigned *ObjectType*. The specification has provided a

*BaseObjectType* from which all other *ObjectTypes* shall either directly or indirectly inherit.

*Variable* NodeClass is dedicated to provide a value to the clients. To define a *Variable* two types must be provided:

- *VariableType*: describes the type of a variable. Each Variable node has the *HasTypeDefinition* reference to its type definition.
- *DataType*: describes the type of the value of the variable. It is assigned to the *DataType* attribute.

The type of data provided by the *Variable* Value attribute is defined by the associated *DataType*. *DataType* is pointed out by the *DataType* attribute of the *Variable* and *VariableType* nodes. In many cases, the value of the *DataType* attribute will be well-known to clients and servers. Well-known data types allow clients to use random addressing and interpret values without having to read the type description from the server.

To some standard data types – called built-in types - special rules apply. Built-in data types are a fixed set that should be known to all OPC UA products. Examples of built-in data types are *Int32* and *Double*. Most of the built-in data types are similar to those in programming languages.

Process data could be complex. *Structure* is an abstract data type defined as the base for all structured types. All complex data, if not defined in the specification explicitly as primitive, are created by defining of new types derived from the *Structure*.

Reference types are used to create interconnections between nodes. They are not instantiated, i.e. a NodeClass representing a reference is not defined. Instead of instantiating the references, they are added to a collection associated with each node. NodeClass of the node and its type decide what references are allowed to be added to this collection.

The base of all references is an abstract *References* type. There is no semantics associated with it. There are two disjoint sets of standard references:

- *HierarchicalReferences*
- *NonHierarchicalReferences*

This distinction reflects two fundamental relationship categories that can be generally distinguished: the association and the dependency. Associations are used to build information architecture – nodes hierarchy - that can be discovered by the clients using the browsing mechanism. An example of the association is the “parent/child” relationship. In this case it can be said that the target belongs to the source. A dependency of a source element (called the client) on a target element (called the supplier) indicates that the source element uses or depends on the target element. An example of dependency is the variable and the variable type relationship. In this case the target describes the source.

### C. Extending OPC UA Information Model

The standard OPC UA Information Model is expandable. For example, in 2008 the OPC Foundation announced support for Analyzer Devices Integration into the OPC Unified Architecture and created a working group composed of end users and vendors with its main goal to develop a common method for data exchange and an analyzer data model for process and laboratory analyzers. In 2009 the OPC Unified Architecture Companion Specification for Analyzer Devices was released [17]. To prove the concept a reference implementation has been developed containing ADI compliant server and simple client using the Software Development Kid released by the OPC Foundation [17].

It is an example of how OPC UA standard Information Model can be expanded by a selected domain application. Standardized expandability of the metadata used to provide a context of underling process data is key requirements of the presented Object Oriented Internet concept.

In this example, the model described in the specification is intended to provide a unified view of analyzers irrespective of the underlying device. This Information Model is also referred to as the ADI Information Model. As it was mentioned, analyzers can be further refined into various groups, but the specification defines an Information Model that can be applied to all the groups of analyzers.

The ADI Information Model is located above the DI Information Model [18] [19]. It means that the ADI model refers to definitions provided by the DI model, but the reverse is not true. To expand the ADI Information Model, the additional layers shall be provided.

## IV. Information Model Deployment

The OPC UA is a standard that allows clients to get access to the server underling processes. To meet this objective, each server instantiates and maintains an Address Space that is a collection of data to be exposed to clients. The OPC UA Address Space consists of nodes and references. The main role of the nodes is to expose the underlying processes state and behavior as a selectable, well-defined piece of information.

To create the Address Space the OPC UA servers must instantiate all nodes and interconnect them by means of references.

As it was stated previously, typical implementation architecture consists of OPC UA Clients, which are connected to an OPC UA server (Fig. 1). To get access to underlying *Processes* data a generic client does not need to have any awareness of the Information Model used to create the Address Space exposed by the sever in advance. However, in the production environment, the Information Model (types) knowledge may be useful to offer additional functions, like dedicated data processing, customized control panels or predefined structure of the database tables. Types knowledge also simplifies configuration of the clients, because all of the items composing the complex process information can be accessed simultaneously – they can have one single address – identifier.

To implement the Address Space two questions must be addressed [20]:

- How to couple the nodes bi-directionally with the underling process data sources?
- How to create and maintain it?

Using the instantiated nodes by means of a well-defined set of services [8] (*OPC UA interface*), clients get access to data representing a selected part of the underlying processes environment. Nodes are divided into classes. The *Variable* class is used to represent the values – has the Value attribute. To be used as the process state representation, the value of the Value attribute must be bound with a real data source, e.g. an analog signal or a database item. The *Method* class represents a function that can be called by the clients connected to the server. In this case the real-time process bindings are responsible for conveying the parameter values, invoking the represented series of operations and returning the execution result. In Fig. 1 both classes are the main building blocks of the architecture that allow the server to couple the exposed Address Space with the current state and behavior of the underlying *Processes*.

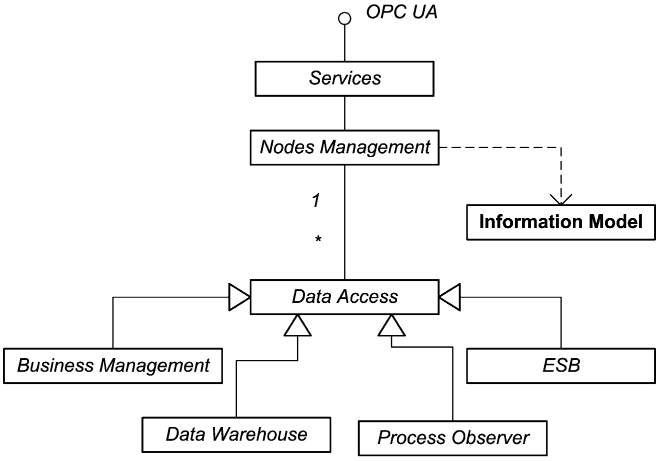


Fig. 2 Process Observer archetype diagram

The technique of binding the nodes with process data is vendor specific, but it must be transparent to the *Clients*. Nodes management functionality on the *Client* part is standardized by the OPC UA Service Model [8] (*OPC UA* interface - as a set of services depicted in Fig. 2). Access to the values representing the current process state is provided by the Read/Write functions. The client can also be informed about changes of the process state using "data change" notifications. Invoke and event notification functionalities allow clients to use the *Methods*.

In Fig. 2 the proposed internal diagram of the *OPC UA Server* package is shown. To implement the functionality presented above, three coupled function classes shall be distinguished:

- *Services*
- *Nodes Management*
- *Data Access*

The diagram in Figure 2 shows the associations between the above function classes. In this architecture the *Data Access* is responsible for transferring process data up and down. The *Nodes Management* function class couples the *Processes* data with appropriate nodes instances representing underling process metadata and provides a homogenous picture to *Services* that finally exposes it to all connected clients.

Real-time process data can be obtained from any underling process, i.e. file system, database, device or even large scale highly distributed automation system. For embedded applications it may directly use internal controller registers of the device. The *Data Access* function class is able to obtain data using the random access or underlying communication infrastructure and vendor-specific protocols.

To create the Address Space - i.e. to instantiate all nodes and interconnect them by means of references - the *Nodes Management* function class uses a predesigned static *Information Context* (dependent on the *Information Model* – not shown in Fig. 2) providing a detailed description of all the nodes, including their attributes and references. Static means that the model is predefined for the selected environment, but it does not mean that the exposed Address Space is static. In this approach, nodes can be instantiated and linked dynamically, however this operations must conform to the model definition. Dynamic behavior of the Address Space can be controlled by the connected clients using services or by the current state of the process.

Before nodes making up the Address Space can be instantiated by the server, this Address Space must be designed first. Model designing is a process aimed at designing Information Model as a set of nodes and their associations and, next, creating the *Information Context* as its representation in a format appropriate for the implementation of the *Nodes Management* function class. Depending on the OPC UA server implementation, the Information Model representation and support for the modeling process varies. The main challenge that must be faced up is how to prepare *Information Context* seamlessly without programming. The designing process can be supported by the Address Space Model Designer tool [19], [21], [22] that is intended to help architects, engineers and developers accomplish *Information Context* preparation. Using the tool it could be similar to preparation of a hypertext document.

The tool developed by a team leaded by the author is very useful to make the publication of the process data in the context of metadata straightforward and without programming, but it is only proof of the solution concept. To promote the Object Oriented Internet concept in a wider scope more research is required with the goal to define a formal, widely accepted representation of Information Model, semantic validation methods, generation of the Address Space and custom complex data serialization to leverage the deliverables to the designers, developers, end user, etc. and to integrate them into other applications. It is proposed to carry on this research work as a common effort using an open source project [23] as the research workspace, which offers basic work framework and very convenient project management utilities available on the well-known GitHub platform. In other words, applications interoperability is yet granted by the OPC UA standard, the next step is to work out unification of the designing/deploying methods and supporting tools to make people cooperation possible and finally the Object Oriented Internet a real option.

## V. Accessing Information Resources

### A. Architecture

According to the definition the Internet is expected to provide access to information resources and services hereinafter referred to as data sources. For the architecture proposed in Fig. 2 the *Data Access* functional class is responsible for fulfilling this job. Because the underling information resources and services are to be exposed in the context of the Address Space the functional class *Nodes Management* is responsible for binding the underlining data sources with appropriate variable and method nodes embedded in the Address Space. These variables and methods are accessible by the remote clients using the standard *OPC UA* interface provided by the *Services* functional class

In the proposed approach there are no limits regarding possible data sources that can be coupled by the *Data Access* with *Nodes Management*. Generally, three classes of data sources can be distinguished:

- Data representing the current state and behavior of the underling real-time industrial processes;
- Archival data representing the behavior of the underling processes in time;
- Current processed data obtained from business supporting applications and other connectivity standards.

A typical example of the real-time physical processes is the industrial automation process control system. The process control contains digital plant floor devices responsible for measurements, controlling and condition monitoring of the real-time process locally. Usually, the predominant function in this case is accomplished using PLC (Programmable Logic Controller) or DCS (Distributed Control System) class products. In a distributed process, one can distinguish autonomous islands of automation, whose cooperation has to be harmonized by a supervisory system that is responsible for controlling the process as a larger whole.

To get access to the plant floor devices and couple them to the *Nodes Management* functional class underlying proprietary communication links must be instantiated. Although from the design point of view this communication can be considered transparent, its availability and reliability is crucial for the final result. Assuming transparency, it simplifies the problem to a great extent, provided that the assumption is valid.

To instantiate a link we need a medium. To transfer data over the medium, we have to use selected protocols controlling access to the medium and responsible for robust data transfer. Additionally, the protocol and medium often limit the bandwidth and medium access. Any of these requirements can cause that the above assumption and, in consequence, this approach becomes unreal. Therefore, we need to look for a compromise between an unacceptable complexity and unreal assumption.

To make the plant floor device interoperable with the *Data Access* functional class, both have to use the same vendor-specific or standard-compliant protocol. Relying on vendor-specific solutions limits future solution expandability. Generally, it is, therefore, not recommended and vendors usually offer a standard protocol for plant floor devices. Unfortunately, there are hundreds of “open standards” defined in the automation marketplace.

For the highly distributed process control systems (like smart grid, smart heat distribution networks, etc.) assuming that the whole system uses one common communication medium is not feasible [24].

Lack of common medium coverage of the whole area that the controlled process is dispersed over requires engaging simultaneously many communication infrastructures, and dealing with a multidimensional communication network. The main advantage of using many infrastructures is the possibility of improving robustness of the system by providing communication redundancy [24] in overlapping areas provided that it is possible to utilize them alternatively.

To transfer the data, we need a medium, but to use the medium, we need to engage an infrastructure: a technology (Internet, GSM, satellite, ISDN, etc.) governed by technical standards and an organization governed by regulations, procedures, practice, etc. A platform optimal today may be useless for future because technology is progressing rapidly and economical standing of organizations may fluctuate.

To address all issues described above the *Data Access* functional class has to be expanded to employ appropriate communication functionality. It is proposed to implement Process Observer architecture presented in next section as an extension to manage the underling communication infrastructure and transfer process data in real-time in a systematic manner.

The next example of the underlying data source is a repository containing manufacturing process information, like data base or even a data warehouse. Data warehouses are designed to facilitate reporting and analysis. This kind of application focuses on data retrieving and analysis, to extract, transform and load data.

To interconnect with an archival processes data repository the *Data Warehouse* extension of the *Data Access* functional class has been added to the proposed architecture in Fig. 2.

Data without context has no meaning, hence metadata is critical to a data strategy. Designing a data binding mechanism both data and metadata must be considered. The *Data Warehouse* extension is responsible for providing an appropriate translation (according to the OPC UA Information Model) between metadata of the underling process data and the context of the server Address Space where the data is made available to the clients. Simplicity of this relationship is crucial to the business, because metadata exposed by the OPC UA server and metadata describing the underling repository content must be designed on the basis of the same semantics rules. In the design process, where the metadata originates and how to synchronize it should be addressed first.

Usually, apart from the historical data access mechanism OPC UA clients use real-time data access subscribing to current data changes. Therefore, the *Data Warehouse* must be smart enough to provide updates by following the repository modifications.

Business Intelligent (BI) applications are a keystone for macro optimization at the enterprise level because they provide an insight into data, which allows analysts and executives to easily uncover patterns and abnormalities in the business [26]. In the late 90s organizations also implemented enterprise resource planning (ERP) and customer relationship management (CRM) software that can be candidates for the next data source. There are many other business level applications (BLA) processing information and providing results that can be published by the OPC UA server in the Internet using the Object Oriented approach.

Usually a data warehouse (DW) is a central part of today's BLA and real-time process control deployment and hence the archival data may be available also indirectly via the *Business Management* or *Process Observer* data bindings.

The cornerstone of a successful BI application is its capability to provide business users fast and easy access to data for analysis. Online analytical processing (OLAP) tools are a foundation of BI application. In the discussed architecture, another option is to distribute BI application over the Internet and couple the OPC server exposing OLAP functionality to the remote applications.

The implementation of the above described functional classes requires a dedicated link used to manage data transfer. Data transfer for the most popular database management systems are governed using Structured Query Language [27]. It is a language rather than connectivity, but can be used together with widely used vendor services to standardize the data access and simplify the *Data Warehouse* implementation.

To make the enterprise more and more beneficial, the applications supporting automation and business processes have to be integrated. From integration, we should expect additional performance improvement as a result of synergy and real-time macro optimization effects. Enterprise Service Bus (ESB) [28] is a standard-based concept and hence it is well suited for integration projects. The ESB provides a highly distributed, event-driven Service Oriented Architecture (SOA) that combines Message Oriented Middleware (MOM), web services, XML data transformation and intelligent routing based on content.

Using ESB as a foundation for the applications integration allows for implementation of the OPC UA server data bindings by interconnecting of the *Data Access* with this bus. In the architecture presented in Figure 2 this role is fulfilled by the ESB extension of the *Data Access*.

### B. *Process Observer Architecture*

*The Process Observer* architecture described in [24], [25] is proposed to be used as a consistent sole representation of a distributed real-time process (Fig. 3). It is an extension of the *Data Access* class (Fig. 2).

In the presented architecture the following classes are distinguished:

- *Cache* is a collection of the latest values of the process data.
- *Controller* holds the plant-floor device data description.
- *Channel* is used to represent independent communication threads conducted simultaneously to each other.
- *Segment* represents a single communication path and is responsible for managing communication resources and data transfer from a group of devices that is to be accessed using the same transport connection.
- *DataProvider* is responsible for providing a stream of data to the *Segment*.
- The *Pipe* is a collection of *Ports*, where only one of them is active at any time.
- The *Port* represents a bidirectional device data streaming functionality.

The description represented by the *Controller* is used to schedule in time all read operations to update the data in the *Cache*.

Usually, lower layer communication requires multidimensional networks. The *Channel* class allows creation as many simultaneous communication paths as it is necessary. To assure mutually exclusive access to common resources, the *Channel* activates only one *Segment* at any time.

To provide a consistent process data from multidimensional network environment and using custom protocols the proposed solution enables to create many *DataProviders* instances by a *Channel* and use them by a data transfer algorithm realized by the *Segment*. Each *Segment* can use only one *DataProvider*, but one *DataProvider* can be used by many *Segments* associated with the same channel.

To provide polymorphism for the environment specific needs, the *DataProvider* is located outside the main software package and inherits an interface ensuring flexible management of the communication medium and transfer of the process data. This solution makes it possible to keep the core software unchanged and adapt a Software Development Kit to the specific needs. In this scenario, the late binding approach is supported. Late binding is useful if it is required to replace a part of software package without recompiling of the code base. In this case, a variety of protocols might be supported with a separate module for each protocol specification. A declarative configuration can be used to tell the application to use a specific module at runtime. Another scenario where late binding can be useful is to enable users of the system to provide their own customization through a plug-in. Again, the system can be instructed to use a specific customization by using a configuration setting.

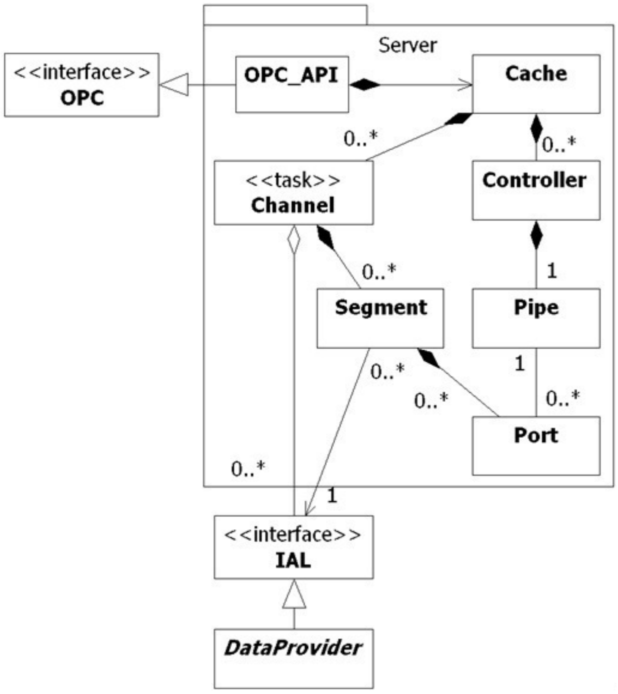


Fig. 3 Process Observer Architecture

In a real environment, apart from accessing underlying process data, monitoring and management of the recourses and communication infrastructure are often of the same importance. Monitor class (Fig. 4) represent this functionality. To commence factory tests or provide a state observer a simulation environment is required. Simulator class is responsible to provide the simulated data and can be used in place of the Protocol class for testing purpose. This concept makes it possible to publish all of the mentioned types of information in the same way using the defined interface and late binding approach.

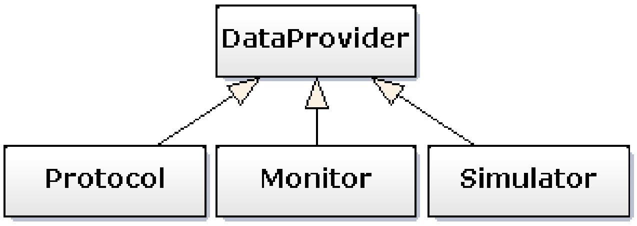


Fig. 4 DataProvider functions

In the proposed model (Fig. 3), the *Pipe* concept is used to assure redundancy. After detecting a failure of the active path, another *Port* belonging to the same pipe is activated immediately. *Segment* uses only active *Ports* and, therefore, the data is transferred over the network once only. The *Pipe* checks availability of non-active paths periodically. Using paths redundancy additional spare plant-floor devices can be used seamlessly as the next level of redundancy.

The main job of the communication software is to make `best effort' to keep the process data fresh and allow clients to access the data randomly. From the communication point of view, two independent communication environments can be distinguished (Fig. 1):

- *Processes* connecting plant-floor devices to an intermediate component (server);
- OPC UA Interface connecting the intermediate component (server) to *OPC UA Clients*.

Because both are used to transfer the process data, it is vital how these data transfer processes are related to one another. To finally design an appropriate sampling scheduling mechanism on the process side, we need to take into consideration:

- Needs of the *OPC UA Clients*.
- Current real-time process state;
- Current communication path load and its throughput;

All of them can change in time and, therefore, it is proposed to implement the following two unique closely coupled costs saving algorithms providing process data just in time and preserving communication bottlenecks:

- *Adaptive Sampling Algorithm* (ASA): responsible for adjusting the plant-floor devices sampling rate according to the current process state.
- *Optimal Transfer Algorithm* (OTA): responsible for minimizing the difference between requirements of client individual process data update rate and current sampling rate of a process control devices;

To minimize the data transfer costs, the sampling rate is adapted to the current process needs.

The Process Observer architecture is widely used as a communication engine in highly distributed systems. The supervisory control of a metropolitan heating system located in the city of Lodz – Poland [24],[25] is an example. The heat distribution network of Lodz (750k citizens) is supplied from heat and power plants with total thermal output of 2.5GW. It consists of:

- 3 heat and power plants,
- 2 backbone pumping stations,
- Hundreds of backbone heat chambers
- Thousands of local distribution points.

Their optimal utilization requires a control system to allow working on common supplying areas. As the system is distributed geographically (about 800km of pipes), safe communication between nodes (automation islands) is very important. An implementation [25] of Process Observer

Architecture proves the concept in highly distributed application.

The architecture presented in this section has been already integrated with the OPC UA services, but further research is required to integrate it with Information Model designing methodology consistently. The main challenge is how to support custom complex data. The complex data must:

- be factored using components gathered from the underling process,
- follow the DataType declarations in the Information Model,
- be transparently serialized over the wire.

## VI. Conclusion

Nowadays, in such a fiercely competitive environment, modern manufacturing and transportation automation systems have to be involved. Such systems usually consist of numerous different ICT systems located at business and process management levels. They are frequently dispersed geographically in multi-division organizations.

The Internet is a globally available communication infrastructure that makes it the first and practically the only candidate to be used as a platform to build a universal solution for the above objectives and even to integrate systems belonging to cooperating organization groups to benefit from the synergy effect and global optimization.

The freely expandable Object Oriented archetype and its practical implementation presented in the article prove that the above goal can be achieved and the final solution offers the following features:

- It provides application to application robust interoperability over the Internet;
- On the server side, it makes it possible to freely publish and update information and services in a contextual (semantics aware) environment;
- On the client side, it makes it possible to get a description, discover and finally get access to the requested information and services;
- Information resources and services exposed by the server that represent the state and behavior of the underlying processes allow clients to manage and control them over the Internet/Intranet;
- Client and server software can be offered by independent vendors as generic off-the-shelf products;
- The products can be tested for interoperability independently of each other.

To accomplish this it is proposed to distribute publisher (server) main tasks to three functional classes:

- A predefined fixed set of services based on the SOA concept conforming to the OPC Unified Architecture specification;
- Information context management using the object oriented programming paradigm;
- A pluggable proprietary data binding mechanism.

Development of generic communication software that can be interoperable requires specification compliance testing. It is proposed that OPC Unified Architecture, a new emerging industrial standard that fulfils requirements derived from this architecture should be used because it provides a definition of an appropriate: set of services and Information Model concept dedicated to formally describe the Address Space – context for the exposed information resources and services. It has wide industrial support and a well defined compliance test procedure governed by the OPC Foundation.

Available reference applications and commercial products pointed out in the article prove that the data binding concept can be successfully implemented as dedicated application-dependent pluggable components. The components must be able to couple the proprietary underling data access mechanism with the server mechanism managing the context where the data is embedded and made available to connected clients.

The approach to represent the underlying data processing environment as presented in the paper can be used for countless applications, from exposing the representation of measurement devices to building multi-enterprise management and remote process control systems. Smart networks, i.e. smart grid, smart district heat distribution networks, utility distribution, oil distribution, railways, etc. are an example of applications like that.

It is worth nothing that to promote the Object Oriented Internet concept in a wider scope more research is required with the goal to define a formal, widely accepted representation of Information Model, semantic validation methods, generation of the Address Space and custom complex data serialization to leverage the deliverables to the designers, developers, end user, etc. and to integrate them into other applications. In other words, applications interoperability is yet granted by the OPC UA standard, the next step is to work out unification of the designing/deploying methods and supporting tools to make people cooperation in this respect possible and finally the Object Oriented Interned a real option.

It is proposed to carry on this research work as a community effort using the open source project [23] as the research workspace on the well-known GitHub platform.

## References

[1] Network Protocols Handbook, Javvin Press, 2007;

[2] http://www.w3.org/ The World Wide Web Consortium (W3C), 2015.

[3] M. Postol, UA Specifications, in J. Lange, F. Iwanitz, T. J. Burke, OPC – from Data Access to Unified Architecture, Hüthig Fachverlag, 2010.

[4] http://www.commsvr.com/UAModelDesigner/ OPC Unified Architecture e-book, 2015.

[5] W. Mahnke, S. Helmut L., M. Damm. OPC Unified Architecture. Berlin: Springer, 2009.

[6] OPC UA Specification: Part 1 – Concepts, Version 1.0 or later. OPC Foundation, 2009.

[7] https://opcfoundation.org/, The OPC Foundation - The Interoperability Standard for Industrial Automation, 2015.

[8] OPC UA Specification: Part 4 – Services, Version 1.0 or later. OPC Foundation, 2009.

[9] E. Gamma, R. Helm, R. Johnson, J. Vlissides: Design Patterns – Elements of Reusable Object-Oriented Software, Addison-Wesley 1995

[10] OPC UA Specification: Part 5 – Information Model, Version 1.0 or later. OPC Foundation, 2009.

[11] M. Postol, Information model, in J. Lange, F. Iwanitz, T. J. Burke, OPC – from Data Access to Unified Architecture, Hüthig Fachverlag, 2010.

[12] OPC UA Specification: Part 3 – Address Space Model, Version 1.0 or later. OPC Foundation, 2009.

[13] OPC UA Specification: Part 6 – Mappings, Version 1.0 or later. OPC Foundation, 2009.

[14] OPC UA Specification: Part 2 – Security, Version 1.0 or later. OPC Foundation, 2009.

[15] C. Adams, S. Lloyd: Understanding PKI: Concepts, Standards, and Deployment Considerations, Second Edition, Addison-Wesley Professional, 2002;

[16] http://www.commsvr.com/Products/OPCUA/CommServerUA.aspx- CommServerUA: Redundant, Multi-protocol, Multi-channel OPC UA Server For Highly Distributed Systems, 2015.

[17] OPC Unified Architecture Companion Specification for Analyser Devices. OPC Foundation, 2009.

[18] OPC Unified Architecture Companion Specification for Devices. OPC Foundation, 2009.

[19] M. Postol, OPC UA Information Model Deployment, CAS, 2015, http://goo.gl/HqYjvy

[20] M. Postol, Design and Modelling of the Address Space, in J. Lange, F. Iwanitz, T. J. Burke, OPC – from Data Access to Unified Architecture, Hüthig Fachverlag, 2010.

[21] http://www.commsvr.com/Products/UAModelDesigner.aspx – OPC UA Address Space Model Designer software, 2010

[22] M. Postol, UA Address Space Model Designer, in J. Lange, F. Iwanitz, T. J. Burke, OPC – from Data Access to Unified Architecture, Hüthig Fachverlag, 2010.

[23] OPC UA Object Oriented Internet, Opc-ua-ooi open source project on GitHub, http://mpostol.github.io/OPC-UA-OOI/, 2015

[24] M. Postol, Real-Time Communication for Large Scale Distributed Control Systems; International Multiconference on Computer Science and Information Technology; Wisła (2007) PIPS, pp. 849–859 ISSN 1896-7094

[25] M. Postol, Large scale distributed process and business management integration; 14th International Congress of Cybernetics and Systems of World Organization of Systems and Cybernetics, Wroclaw (2008), pp. 632-642, ISBN 978-83-7493-400-8

[26] W. A. Giovinazzo: Internet-Enabled Business Intelligence, Prentice Hall; 2002

[27] T. Connolly, Database Systems (2nd ed.). Addison-Wesley, 1999.

[28] David A Chappell: Enterprise Service Bus, O'Reilly Media, Inc., 2004;

[29] J. Lange, F. Iwanitz, T. J. Burke, OPC – from Data Access to Unified Architecture, Hüthig Fachverlag, 2010.

DOI: 10.15439/2015F239
ACSIS, Vol. 5

# The development of InterNetwork channel Emulation platform for Surgical Robot Telemanipulation control system (INSeRT)

Maciej Rostański,
Paweł Buchwald,
Krystian Mączka
University of Dąbrowa
Górnicza, Poland

Paweł Kostka
Silesian University of Technology,
Biomedical Engineering Faculty,
Zabrze, Poland

Zbigniew Nawrat
Foundation for Development
of Cardiac Surgery
in Zabrze, Poland

***Abstract*—In this paper, we describe the RobinHeart surgery robot development related project in which entire robot operation is supposedly done remotely, using wide area network connection. In such environment, any vision and telemanipulation data packets are subject to delay, limitations, issues and failures, as any network connections do. It has become necessary to create an internetwork emulation for the purposes of robot-related trials. The INSeRT platform has been developed to fulfill that role. INSeRT platform design is presented and discussed and first results of validation with real-life example are shown. Network channel manipulation techniques can be implemented at data-link OSI layer level, or it may be induced at network layer level – this approach relies on network layer packet manipulation and poses a promising start, as well as raises questions about whether implementing different distributions for delay or reordering is going to have significant impact on channel emulation and traffic parameters.**

***Index Terms*—Network performance, simulation, emulator, WAN, internetwork, telemanipulation, remote operation**

## I. Introduction

THE research and development presented in this paper is driven by the development of a surgical robot for long distance operation. The family of Robin Heart [1] tele-manipulators was founded in the Foundation for Development of Cardiac Surgery of Prof. Zbigniew Religa (FRK) in Zabrze, in collaboration with specialists from several academic centres (Lodz, Gliwice and Warsaw Technical Universities). According to project preliminary assumptions, the robot would have a segment-like structure allowing different configurations set up for various types of soft-tissue surgery. In particular, it should have an independent arm of the endoscopic video track with a wide range of application. The project has evolved together with the increase of experience gathered by the construction team.

In the first phase of the Polish Robin Heart Project, three robot models: Robin Heart 0, 1 & Robin Heart 2 (Fig.1) were created [2], differing from one another by the concept of control system and mounting. Between 2007 and 2008, the Robin Heart Vision, a robot for tracking video endoscopic channel, was constructed and tested. In 2010, after only one year of work, a new model, the Robin Heart mc2 appeared in the laboratory and in the Animal Experimental Medicine Centre for in vitro and in vivo tests. The robot fulfils the role of three operators: the first & second surgeon as well as the assistant holding the vision channel. In the same year, after completing a 2-year project, novel mechatronic tools, the Robin Heart Uni System were created. These innovative tools allowed to put into practice the idea of using the same surgery tools both on the robotic arm, tele-controlled by the operator, and in given cases in the other way: similar to traditional laparoscopic tools - manually but driven by means of a special handle with micro-motors mechanisms.

At the same time, effective systems of Man-Machine interface including ergonomic Surgeon (Master) environment with comfortable operating position of the Surgeon, high quality vision system and intuitive contact with Master tool of tele-manipulator (including force feedback, which is still optional, in research phase) are being developed [3].

In this paper, we describe the part of the project in which entire robot operation is supposedly done remotely, using wide area network connection. In such environment, any vision and telemanipulation data packets are subject to delay, limitations, issues and failures, as any network connections do. It has become necessary to create an emulation of internetwork for the purposes of robot-related trials. The INSeRT platform has been developed to fulfill that role.

This paper is organized as follows: the questions and problems with remote long distance operations in RobinHeart are presented, then the short review on network channel simulation state-of the-art and related issues are mentioned. INSeRT platform design is presented and discussed and first results of validation with real-life example are shown. Conclusions are drawn and the most important fields of further study are pointed out and summary is described at the end.

## II. Telemanipulation Control System for Long Distance Operation by Means of Robin Heart Family of Surgery Robotic Manipulators

Telemanipulator invented for less, minimally invasive cardiac surgery is a computer-controlled device, located between surgeon's hands and the tip of a surgical instrument or endoscopic vision channel. Basic requirements for this device are stable operative field of view, direct surgeon control and high level of precision. Main advantages, motivating the introduction of tele-manipulators to minimal invasive surgery field are also:

- Scaling of movement between Master console and Slave arm
- Filtering and tremor removing

978-83-60810-66-8/$25.00©2015, IEEE

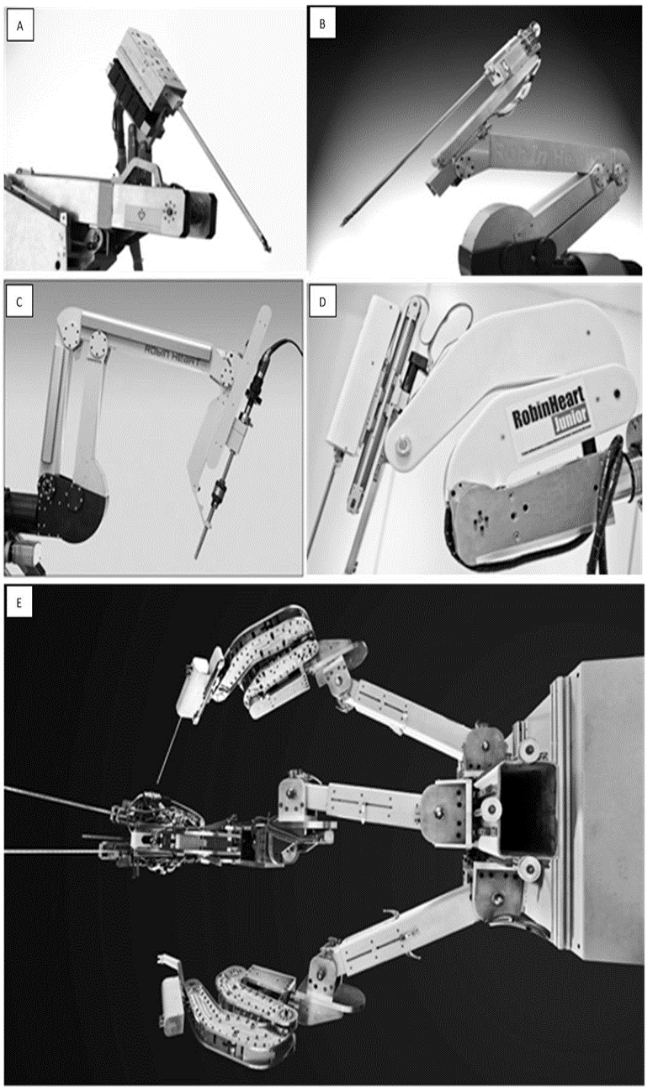


Fig 1. Progress in RobinHeart project development. - The Robin Heart family robots.: Robin Heart 0 (A), Robin Heart 1(B), Robin Heart Vision (C), Robin Heart Junior (D), Robin Heart mc$^2$(E)

- Comfort of surgeon work improvement with high quality HD (stereoscopic) imaging from operation field with zooming and filtering

General structure of Master-Slave (Operator-Robot arm) teleoperation system with video and manipulation data channel is presented on Fig. 2. Optional force-feedback track is presented, that is used to transfer tool-tissue interaction to Surgeon/Operator.

In this type of robotic arm navigation the slave manipulator mimics movements of master controller, driven by Surgeon/Operator (Fig. 3). A brushless DC motors working in connection with the robust local CAN bus with dedicated low level controllers were used as a driven units for every arm with four degree of freedom. Control system exchange data between master and slave part of control system, which process and transfer data by means of *network shared variables*. Both video transmission as well as manipulation transmission channels performances are closely related to outcome performance of an operator. In case of setting up any of these channels (and practically that would mean both of them) using internetwork, the network channel performance characteristic becomes critical to the system.

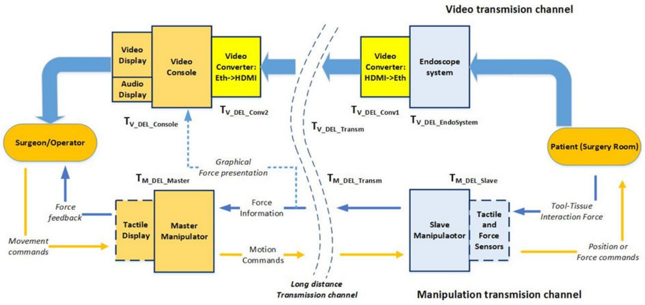


Fig 2. General structure of Operator-Robotic_Arm bidirectional data transfer: video and manipulation channel

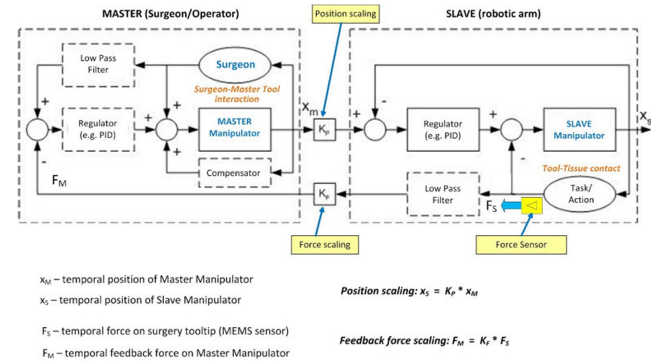

MASTER (Surgeon/Operator)
Position scaling
SLAVE (robotic arm)
Low Pass Filter
Surgeon
Surgeon-Master Tool interaction
+
Regulator (e.g. PID)
+
+
MASTER Manipulator
$X_m$
$K_P$
+
-
Regulator (e.g. PID)
+
SLAVE Manipulator
$X_s$
+
-
Compensator
$F_M$
$K_F$
Low Pass Filter
-
Tool-Tissue contact
Task/ Action
$F_S$
Force scaling
Force Sensor
$x_M$ – temporal position of Master Manipulator
$x_S$ – temporal position of Slave Manipulator
$F_S$ – temporal force on surgery tooltip (MEMS sensor)
$F_M$ – temporal feedback force on Master Manipulator
Position scaling: $x_S = K_P * x_M$
Feedback force scaling: $F_M = K_F * F_S$

Fig 3. Bilateral, Master-Slave control system structure of Robin Heart Robot

## III. Selected Issues of the Network Channel Emulation

The need for analysis and simulation of network traffic is constantly increasing. A large number of studies in this field derives from the need to improve the quality of services (QoS), complete the  service level agreement (SLA) or ensure proper network security level. As a first example, authors in [4] discuss the characteristics of actual Internet traffic in the context of flows and investigates the cause for the high fluctuation in the number of flows. They defined and analyze measurement metrics (such as link utilization, packet arrival rate, and the number of flows) and also test it by the use of university network with over 6000 end hosts and servers.

It is worth mentioning that types and patterns of actual network traffic flows changed dramatically.  As the authors show in [5] the proportion of traditional traffic (determined by well-known port based flows) is decreasing for p2p (peer-to-peer) traffic or other media as streaming or network gaming therefore a better method is to classify the traffic according to application layer programs. The authors indicates that it has become difficult to detect newly developed Internet applications which use random port numbers rather than static and registered ports.

Many aspects of traffic characterization and the dynamics and patterns of network usage was discussed by the authors in [6]. Authors proposed characterization of the traffic by linking the flow measurement architecture with the estimation algorithm. Developed framework was used to estimate the distribution and sample space of the underlying traffic by the use of nonparametric Parzen window (which is an information theory technique, possible to use for classification with various types of data sets [7]). Köandgel in [8] shows how important is characteristic of networks by the use of one-way delay measured, or taken, between any two points in the network. At his work author highlights that precision of one-way delay calculation for flow data depends on the capturing devices, especially on theirs timestamp resolution. Also, [9] presents distributions of flow length, packet size, throughput, for the popular and bandwidth consuming applications as stochastic phenomenon. Flow duration, flow bytes, packet size for popular network applications in the probability density function was presented for the flows collected at several locations on a corporate Intranet.

The classification of network traffic may be performed using machine learning approaches [10]. Network traffic classification can be used for network management (automated problems detection) or automated intrusion detection systems. The authors compare the research in this area. Very interesting subject is the collation of features, used by various researchers, to describe the Internet traffic dataset.

## IV. System design and technical implementation

Taking into account presented works, the most important distinguished features included:

- the necessity of control over the basic parameters of the network connection, in compliance with [11] and just like previously discussed in [12] and [13]: delay (latency), data transmission rate (throughput), lost packets ratio (packet loss) and delay variation (jitter) - this requirements form and review phases were conducted with methods described by one of authors in [14];
- as the requirements elicitation process has proven, specific system functionality has been necessary, which are pre-defined sets of network parameters and their variability trends over time (more on this subject below);
- the system should be able to gather information about the network parameters induced into channel in specific moment in time – the availability of such data will help to identify the correlation between the parameters of telemanipulation channel and assumed measure factors of remote controller operation.

Due to specific project stakeholders demands, the requirement of solution simplicity and capability of operation on available common hardware and software infrastructure has also been considered a ‘must-have’ (entire system had to be mobile and easily transferred, as well as we want to maintain the positive effect on system integration with other components).

The research and business enquiries allowed to distinguish two groups of solutions. Network channel manipulation techniques can be implemented at data-link OSI layer level, or it may be induced at network layer level, with most important examples of:

- Linux operating system kernel providing couple of built-in and patched capabilities, such as *tc* and *netem* package, *Iptables statistics* and/or *random* module, *Iproute2* packet capabilities exploiting *mark* in *iptables* software;
- FreeBSD operating system’s packet *dummynet*, providing similar capabilities;

Both concepts – bridge (Layer 2) or router (Layer 3) based, provide desired functionality in different manners, which will be the subject of another authors’ study. Taken into account the requirement of easy implementation in virtualized environment and the API functionality, the layer 3 approach has been chosen. There are a couple of entire similar systems as well, such as NIST’s project NISTNet (no longer maintained), or Tata’s Performance Engineering Research Centre *WANem* (The Wide Area Network emulator) which was actively taken into consideration, but the decision was made to create own solution, with API dedicated for Robin surgical robot environment and capable of data collection for channel testing purposes. Due to Linux popularity and for stability, CentOS OS 6.6 has been recommended for implementation.

The logical network infrastructure is presented on Figure 4.

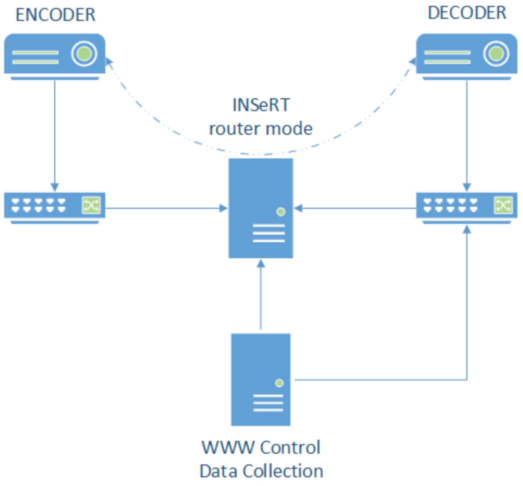


Fig 4. . INsERT system topology

Due to the need for the reduce of the interaction of the transmission line emulator with additional services such as HTTP graphical configuration module or database for connection parameters collection, they have been implemented on a separate server. Communication between a web application that offers GUI functions and module responsible for traffic manipulation takes place on a separate, dedicated channel. Commands are transferred with SSH2 protocol, which allows the use of a wide range of solutions, including certificates, so storing passwords in configuration files is avoided completely. Remote connection using SSH2 library enables standards console stream input and thus the introduction of any previously prepared user commands responses to an event related to changes in network parameters.

## V. Validation of Network Channel Emulation with INsERT

Two network interfaces are used to manipulate outgoing traffic, thus enabling asymmetrical network connection emulation (by varying parameters for both directions). Example of basic manipulation is presented in Table 1.

Table 1.
Basic testbed set example

| Parameter | Values set | Values measured |
|---|---|---|
| **Latency** | 100.0ms | 104.120ms |
| **Packet loss** | 10% | 9% |
| **Jitter** | 50ms | 46.541ms |

Currently, after test trials, three basic sets, or link categories, were defined, illustrating typical internetwork connections that might be used for telemanipulation and/or vision channels:

- Good quality connection – corresponding to leased WAN line;
- Limited quality connection – corresponding to channel realized using Internet;
- Problematic connection – corresponding to ‘challenged links’, for example cellular or even satellite connection.

Those three categories were implemented as a presets of parameters, using simple statistical models of network traffic. Statistical characteristics are based on traffic recorded during experiments with every channel type, but those simple models do not take into the account the non-standard delay distribution, or correlation for packets delay, jitter or loss (which in turn do not emulate burst losses, for example) yet.

The example of limited quality connection emulation with real-life case study as a comparison is shown on Fig. 5 (TCP time/sequence graph) and Fig.6 (throughput). The simplicity of a model is visible on throughput graph comparison in particular – there is much more variation in real-life scenario, however the transmission characteristic is similar.

## VI. Conclusions and Summary

In presented article, the design and implementation process of internetwork channel emulation for surgical robot telemanipulation experiment was presented, as well as the discussion on various methods, issues and future possible work. The system is functional and provides means for emulating real-life scenarios when using internetwork links for remote operation of equipment using video and telemanipulation channels. System provides functionality for traffic manipulation, allowing to simulate different kinds of network links and situations. Open research question, to be pursued in further work, is whether implementing different distributions for delay, reordering, and/or packet loss is going to have significant impact on channel emulation and traffic parameters.

The INsERT system concept assumes that beside traffic characteristics manipulation, the system is designed for collecting the data on network parameters set when working remotely with surgical robot. The maintenance of ergonomic conditions of surgeon’s work is essential to surgical robot system and proper assessment of the network channel conditions impact on any activity is crucial to any experiments on that matter. The determination of the minimum acceptable channel parameters is necessary to anticipate problems with the telemanipulation track, based on changing network traffic characteristics. Thanks to implemented system, it becomes

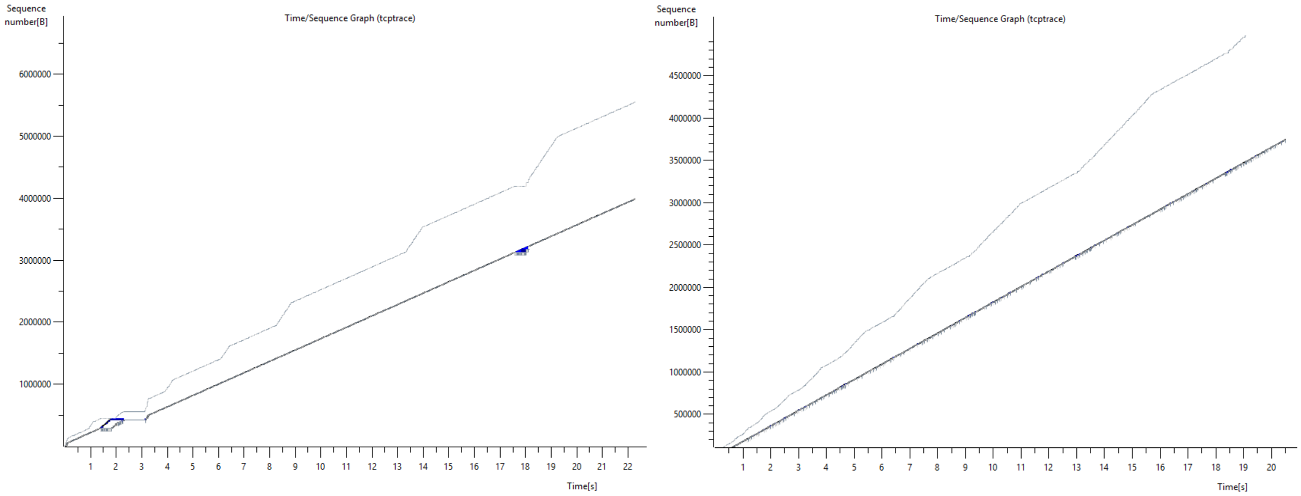


Fig 5. General structure of Operator-Robotic_Arm bidirectional data transfer: video and manipulation channel

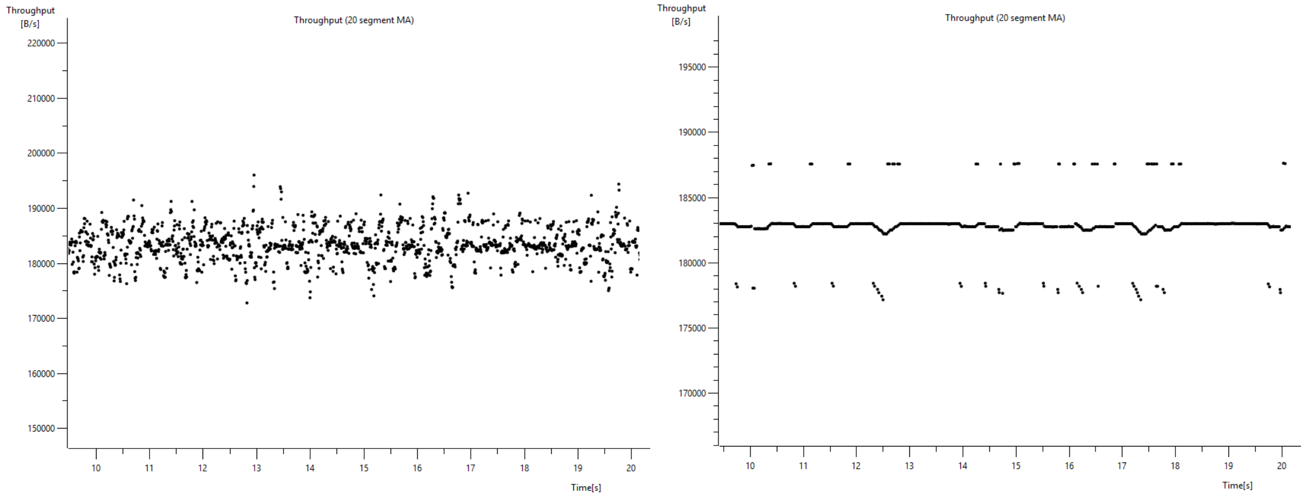


Fig 6. General structure of Operator-Robotic_Arm bidirectional data transfer: video and manipulation channel

possible to designate parameters values in an empirical manner with exercises with real surgeons in controlled training environment. Obtained data will allow for network link validation for surgical robot vision and telemanipulation purposes in real-time. Trial studies with control group of surgeon students were already conducted and will be covered in extensive manner in the future. Current research connected with the studies of influence of network parameters values on telemanipulation performance are tested on new RH Tele ™ arm (Fig. 7).

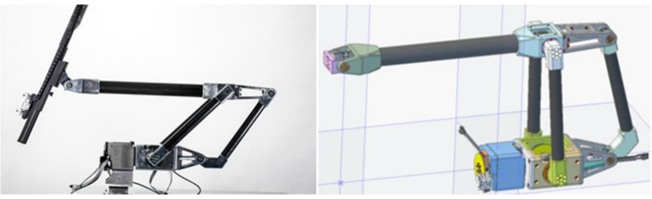


Fig 7. . RH Tele ™ arm, made from carbon base, light materials (final arm and project)

## Acknowledgements

This work is a part of project „**TeleRobinSurgery – Creating and study on new technical systems for remote controlled surgical procedures by means of Robin Heart family telemanipulators**" (ID **181019**) financially supported by Polish National Center of Reseach and Development, based on contract nb **1/A3/2/2012.**

## References

[1] Nawrat, Z. and Kostka, P. (2006), Polish cardio-robot 'Robin Heart'. System description and technical evaluation. Int. J. Med. Robotics Comput. Assist. Surg., 2: 36–44. doi: 10.1002/rcs.67

[2] Nawrat, Z. and Koźlak, M. (2007). Robin Heart system modelling and training in virtual reality. Journal of Automation Mobile Robotics and Intelligent Systems, 1, 62-66.

[3] Kostka, P. and Nawrat, Z. (2012). Wybrane interfejsy chirurg-maszyna w strukturze systemu wizyjnego i sterowania telemanipulatorów chirurgicznych rodziny Robin Heart. Pomiary, Automatyka, Robotyka, 16, 420-423.

[4] Kim, Myung-Sup, Young J. Won, and James W. Hong. (2006). Characteristic analysis of internet traffic from the perspective of flows. Computer Communications 29.10, 1639-1652.

[5] Saroiu S., Gummadi K.P., Dunn R.J., Gribble S.D., Levy H. M. (2002). An Analysis of Internet Content Delivery Systems, Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI), Boston, MA, Dec. 2002

[6] Kundu S.R., Pal S., Basu K. and Das S.K. (2009). An architectural framework for accurate characterization of network traffic. IEEE Transactions on Parallel and Distributed Systems 2009;20(1):111–23.

[7] Biesiada, J., Duch, W., Kachel, A., Maczka, K., and Palucha, S. (2005). Feature ranking methods based on information entropy with Parzen windows. In: International Conference on Research in Electrotechnology and Applied Informatics (Vol. 1, p. 1).

[8] Köandgel J. (2011). One-way delay measurement based on flow data: quantification and compensation of errors by exporter profiling. In: 2011 International conference on information networking (ICOIN), 2011. pp. 25–30.

[9] Liu D, Huebner F. (2002). Application profiling of IP traffic. In: 27th annual IEEE conference on local computer networks, 2002. Proceedings. LCN 2002: 220–229.

[10] Nguyen T., Armitage G. (2008) A survey of techniques for Internet traffic classification using machine learning. Communications Surveys Tutorials, IEEE 2008;10 4): 56–76

[11] Paxson et al.: Framework for Internet Protocol Performance Metrics, RFC 2330, RFC 7312

[12] Rostanski M., Pikiewicz P. (2010). TCP Congestion Control Algorithms Performance in 3G networks with moving client. In: Performance Modelling and Evaluation of Heterogeneous Networks, Proceedings of 6th Working International Conference HET-NETs 2010, ISBN: 978-83-926054-4-7, pp. 379-390

[13] Grzywak A., Pikiewicz P., Rostański M.: Sieci bezprzewodowe, Wyższa Szkoła Biznesu w Dąbrowie Górniczej, Dąbrowa Górnicza 2010, ISBN: 978-83-88936-74-6

[14] Duda J., Rostański M., Borczyk W., Grochla K. (2015). Applying Kano model into goal/requirements elicitation for crossplatform mobile content technology. In: Conference Proceedings of Strategic Management and its Support by Information Systems 2015 (in print)

DOI: 10.15439/2015F52
ACSIS, Vol. 5

# Simulation of Mobile Wireless Ad Hoc Networks for Emergency Situation Awareness

Andrzej Sikora[1] and Ewa Niewiadomska-Szynkiewicz[1,2] and Mateusz Krzyszton[2]
Research and Academic Computer Network (NASK)[1]
Wawozowa 18, 02-796 Warsaw, Poland
Institute of Control and Computation Engineering[2]
Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
Email: Andrzej.Sikora@nask.pl, ens@ia.pw.edu.pl

***Abstract*—Mobile self-organizing ad hoc networks (MANETs) can significantly enhance the capability to coordinate the emergency actions as well as monitor contaminated areas, explore unmanned space, inspect and control working environments. The management of networks that can dynamically and freely organize into temporary topologies raises interesting problems, that are particularly challenging for networks formed by mobile devices. Due to the inherent complexity of these systems, the development of applications relying on mobile network nodes and wireless communication protocols would be greatly simplified by the use of specific tools for supporting testing and performance evaluation. Modeling and simulation are widely used in the design and development of wireless ad hoc networks. In this paper we model mobile ad hoc network (MANET) using discrete event systems methodology (DEVS) and describe the functionality and performance of the Java-based simulation tool for the performance evaluation of self-organizing and cooperative networks for emergency situation awareness. The simulator can provide a useful support for the verification of the design of a network system, employed communication protocols, control and coordination algorithms, allowing the user to display a step-by-step evolution of the network in a suitable graphical interface. The practical case studies are provided to illustrate the operation and performance of the presented software.**

## I. Introduction

THE AD hoc networking is a relatively new area of research that has become extremely popular over the last decade and is rapidly increasing its advance into different areas of technology. A mobile, wireless, and ad hoc network (MANET) is formed of wireless mobile devices (network nodes) that can dynamically and freely self-organize into temporary network topologies. The topology of MANET may change rapidly and unpredictably. Moreover, in many application scenarios there is no prearrangement assumption about specific role a given node should perform. Each device makes its decision independently, based on the situation in the domain and its knowledge about the network. Nodes communicate wirelessly and share the same radio channel. The devices located within their transmission range can communicate directly without the need for an established infrastructure and centralized administration. For communicating with devices located beyond the transmission range, the node needs to use intermediate nodes to relay messages hop by hop. Thus, in general, routes between mobile nodes may include multiple hops.

Currently research effort is directed toward the specifics and constraints in ad hoc networks, such as: limited transmission range, limited link bandwidth and quality of transmission, constrained resources, mobility nature of the network and transmission security [1], [2], [6].

To design a self-organizing MANET that can support the monitoring in emergency situations and/or support coordination of emergency actions the following problems have to be solved: (i) how to determine a minimum number of devices to monitor an area, explore an unmanned space or control working environment; (ii) how to determine the optimal positions for all devices and how to manage internode communication to imply connectivity among the working set of devices and a base station, and how to coordinate and control all devices; (iii) how to schedule devices to reach the destination positions, etc.

Design, development and evaluation of MANET is a nontrivial task, especially as it is envisioned to be deployed in a large scale. It is obvious that the complexity and scale of modern ad hoc networks limit the applicability of purely analytic analysis. Therefore, investigation of MANETs is achievable by resorting either to software simulators or to testbed networks. In most applications the lack of flexibility of testbeds and high costs of their development make simulation unavoidable.

In this paper, we model a MANET system using discrete event systems methodology. We describe a flexible interactive simulation environment for the development of self-organizing, cooperative wireless networks formed of static and mobile devices that can be used to monitor contaminated area, inspect a harsh environment, create a communication infrastructure to collect measurements and transmit them to the base station, and support coordination of emergency actions.

## II. Wireless Networks Simulators

Due to the complex nature of MANET, its simulation is a challenging task. It needs models of hardware, wireless propagation, mobility, energy usage, decision making, etc. Hence, simulators are inherently complex and they generally require huge computational resources to execute. In general,

978-83-60810-66-8/$25.00©2015, IEEE

they rely on various techniques for improving their accuracy, usability, scalability, speed, etc.

A variety of software environments simulating wireless networks are available today. Simulators rely on various methods and technologies for improving their accuracy, usability, efficiency and scalability. An overview of state of the art ad hoc network modeling and simulation tools available commercially and from open source is presented in [16], [40] and [14]. Researchers and engineers can choose among publicly available products or alternatively, can develop their own simulator. The commonly used network simulators like Riverbed Modeler (OPNET) [36], ns2 [20], ns3 [21], OMNeT++ [30] or GloMoSim [11], and its commercial version QualNet [31] can simulate ad hoc networks. Moreover, several software tools for mobile robots simulation, like v-rep [39] can be used. These tools provide the facility to simulate the protocols in different layers (MAC protocols, routing protocols, etc.), and some of them simulate movement of nodes (wireless devices). There are a number of possible sets of criteria that could be used for network simulators comparisons, e.g. time of execution, memory requirements, scalability, available functionality, programming interface, etc. Different tools are optimized for different purposes.

Most of available ad hoc networks simulators require costly shared-memory supercomputers to run even medium size network simulation. We are involved to large scale mobile network systems simulation and their practical applications, and our goal was to develop scalable simulator operating in real time. Hence, to provide high performance and scalability we utilized the paradigm of federating disparate simulators [7] and asynchronous distributed simulation technology [22], [40]. This is the main difference between our software and the other tools. Moreover, most existing ad hoc networks simulators focus on the MAC protocols implementation with the lack of the radio management and mobility modeling. Usually only simplified wireless transmission models and obstacles free simple mobility models are provided (ns-2, Castalia project in OMNeT++, v-rep). The other reason for developing a new simulator was the complicated architecture of available tools and limitations in results visualization and user-system interaction. In case of OPNET, OMNeT++ or ns-2 and ns-3 systems a user must read a large number of manuals to learn how to use the tool. The source coding is usually specialized for a given simulator and it is not easy to implement a given example and add modules developed by the user. Many network simulators do not support both the user interactions during the experiments and animation.

The current version of our simulator called MobASim provides implementations of radio propagation models, mobility models handling obstacles and tools for an environment (simulation scene) modeling. All these models are described in the next section. The open design of the architecture of MobASim, and its extensibility to include other open source modules or modules developed by the user, which are specific to a given application, was chosen in the hope that the system will be a useful platform for research and education in ad hoc networks. The federated approach to simulation of networks and provided functionality make different our tool from mentioned popular software systems for simulation.

## III. Ad Hoc Wireless Network Modeling

### A. Network Modelling

The aim is to model an ad hoc network formed by $N$ static and mobile, self-organizing and cooperating wireless devices $D$ (network nodes) equipped with sensors. All nodes can move with the speed $v \in [v_{min}, v_{max}]$ in the workspace $W$ avoiding the existing obstacles and communicate through radio.

In our research we use the discrete event systems methodology (DEVS) to model a mobile network, i.e., the network operation being modeled is understood to advance through events. The considered DEVS system is composed of several components responsible for different functionalities. We distinguish three types of such components: *node* – a mobile or static device that executes the assigned task, *mobility manager* that is responsible for tracking the nodes on the map and collision avoidance and *communication manager* that models the wireless communication between all nodes. Hence, each MANET simulator consists of three groups of logical processes LPs, responsibly Ns, MMs and CMs. The wireless communication and mobility models that are implemented in our simulation software are described below.

### B. Mobile Network Node Modelling

A high quality mobility modeling is a critical aspect that has great influence on the performance characteristics of each network node and the whole ad hoc system simulation. In literature we can find several less and more realistic models, such as a random mobility model with Brownian-like motion, a random waypoint model (RWP) with randomly generated target point and velocity, a random direction model with randomly generated direction of movement. Map-based mobility models are used for applications in which nodes are constrained to move within defined paths. The surveys of mobility models and main directions to mobility modeling of moving wireless devices one can find in [2], [8], [19], [37].

Widely used approach to mobility modeling, is to apply a concept of an artificial potential field that can be viewed as a landscape where the mobile device moves from a high-value state to a low-value state, [5], [37]. The artificial potential function $V$ is a differentiable real valued function, which value can be viewed as energy. The gradient of $V$ results in a force $F$, which points in the direction that locally maximally increases $V$. The potential function can be constructed as a sum of repulsive $V_{-}$ and attractive $V_{+}$ potentials. The obstacle repels the mobile device while the goal attracts it. The sum of $V_{-}$ and $V_{+}$ draws a device to the target while deflecting it from obstacles. The concept of a potential function is used in particle-based mobility schemes [37], where network node considered as a self-driven moving particle is characterized by a sum of forces, describing its desire to move to the target and avoiding collisions with other nodes and obstacles.

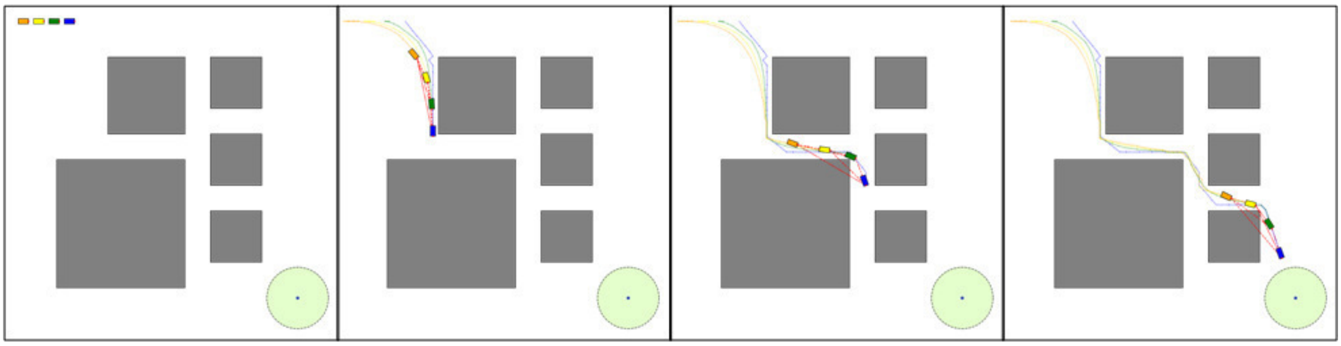


Fig. 1: Convoy formed by the team of mobile devices.

In our research we investigate a group mobility model for calculating of collision-free motion trajectories for $N$ wireless devices that form a cooperative network. Our computing scheme adopts two techniques, the concept of an artificial potential field and the concept of a particle-based mobility. This model can be used to cooperative and fully connected networks design. The detailed description of our model is provided in [24], therefore in this paper we present the summary.

We model each mobile device $D_i$ by a polygon in the workspace. We define its reference point $\mathbf{c}^i = [x^i, y^i]$, which is the location of its antena and a target location – the destination point $\mathbf{c}_g^i = [x_g^i, y_g^i]$. It is obvious that to maintain a constant communication with the base station at least one another device has to be within the transmission range of each $D_i$. Therefore, the motion trajectory of each $D_i$ depends on the potential $U_g^i$ between $D_i$ and its target and the potentials $U_k^i$ between $D_i$ and other devices $D_k$, $k = 1, \ldots, N$, $k \neq i$. Hence, a simple artificial potential function $U^i$ for $i$th device can been calculated

$$U^i(c^i) = U_g^i + \sum_{k=1, k\neq i}^{N} U_k^i(c^i), \tag{1}$$

where

$$U_g^i(c^i) = \epsilon_g^i \Big(\frac{\overline{d}_g^i}{d_g^i} - 1\Big)^2, \qquad d_g^i = ||c^i - c_g^i||, \tag{2}$$

$$U_k^i(c^i) = \epsilon_k^i \left(\frac{\overline{d}_k^i}{d_k^i} - 1\right)^2, \qquad d_k^i = ||c^i - c_k^i||. \tag{3}$$

$\epsilon_g^i \geq 0$ and $\epsilon_k^i \geq 0$ in (2) and (3) denote weighting factors determining the importance of the target point $g$ and the device $D_k$ respectively. $d_g^i$ and $d_k^i$ are real Euclidean distances between $\mathbf{c}^i$ and $\mathbf{c}_g^i$ and $\mathbf{c}_k^i$ after a network transformation, $\overline{d}_g^i$ and $\overline{d}_k^i$ the reference distances calculated due to the current signal strength measurements.

Taking into account the potential function defined in (1) we can formulate the optimization problem (4) and calculate the expected position of the reference point $\mathbf{c}^i$ of $D_i$ after moving the device:

$$\min_{\mathbf{c}^i} U^i(c^i). \tag{4}$$

The Tangent Bug algorithm described in [5] is applied to avoid obstacles in the workspace while moving the devices.

In summary, the following algorithm is used to motion trace calculation of $i$th device at time instants $t_0, t_1, \ldots, t_i, \ldots$:

1) Calculate the reference distance $\overline{d}_g^i$ to the target point $g$ and the reference distances $\overline{d}_k^i$ to all neighbour nodes due to the current maximal radio range and environment characteristics.
2) Calculate the displacement for the device $D_i$ for results of step 1.
3) Move $D_i$ to the new position in $W$.
4) Rotate $D_i$ (if necessary).
5) Calculate and broadcast to the network the new position of the device $D_i$.
6) Return to step 1.

We assume that each network device can calculate its localization in the workspace based on the measurements from GPS or other localization systems ([15], [17], [29], [38]. The algorithm terminates when the minimal the potential function (1) reaches the minimum value, and all the devices don't change their positions. The example results of application of our model to calculate the motion patterns for a group of devices forming a convoy are presented in Fig. 1

### C. Internode Communication and Reference Distance Calculation

A signal strength measurement using the RSSI (*Received Signal Strength Indicator*) are often used to estimate the reference distances $d_g^i$ and $d_k^i$ in real networks. For purposes of the simulation we applied the long-distance path loss model to radio transmission modelling, [35]. It indicates that received signal power decreases with a distance, both in outdoor and indoor environments. We assumed that the radio coverage region of the transceiver of $D_i$th device is a disc centered at $\mathbf{c}^i$. Hence, the signal degradation $PL(d)$ ("path loss") with distance $d$ can be defined as follows $PL(d)[dB] = P_t(d)[dBm] - P_r[dBm]$, where $P_t$ denotes power used by a sender to transmit the signal and $P_r$ power of the signal received by a receiver. In the long-distance path loss model $PL(d)$ is modeled as a random variable with log-normal distribution ([2], [9], [32]).

$$PL(d)[dB] = PL(d_0)[dB] + 10nlog\left(\frac{d}{d_0}\right) + X_\sigma, \qquad (5)$$

where $d_0$ denotes a reference distance ($d_0$=1 m for IEEE 802.15.4) and $X_\sigma$ a zero-mean Gaussian distributed random variable with standard deviation $\sigma$ (all in dB).

To enable the communication the signal strength received by neighboring nodes should exceed a receiver sensitivity $P_s$. The Q-function may be used to determine the probability that the received signal level will exceed $P_s$. It is defined as follows

$$Q(z) = \frac{1}{\sqrt{2\pi}}\int_z^x \exp\left(-\frac{x^2}{2}\right)dx, \qquad Q(z) = 1 - Q(-z). \quad (6)$$

The probability that the received signal level $P_r$ in distance $d$ will exceed a value $P_s$ can be calculated from the cumulative density function as

$$P[P_r(d) > P_s] = Q\left(\frac{P_s - \overline{P_r(d)}}{\sigma}\right). \qquad (7)$$

A tabulation of the Q-function for various values of $z$ is given by Rapaport in [35]. The received signal exceeds the receiver sensitivity with probability of 99% for

$$\frac{P_s - \overline{P_r(d)}}{\sigma} = -2.3, \qquad \overline{P_r(d)} = P_t - PL(d). \qquad (8)$$

From (8) we can calculate $PL(d)$

$$PL(d) = -2.3\sigma - P_s + P_t. \qquad (9)$$

Using (9) and definition of $PL$, (5) we can estimate the reference distance $\hat{d_k^i}$ between $i$-th and $k$-th mobile nodes

$$\hat{d_k^i} = d_0 10^{\frac{-2.3\sigma - P_s^k + P_t^i - PL(d_0)}{10n}}, \qquad (10)$$

where $P_t^i$ denotes a transmission power of a node $i$ and $P_s^k$ the sensitivity of a node $k$.

The value of $X_\sigma$ and $n$ in (10) depend on the workspace conditions, and can be calculated using linear regression such that the difference between the measured and estimated path losses $PL$ is minimized over a wide range of measurement locations. The values of $n$ estimated for various environments are provided in [35].

### D. Environment Modeling

To create a workspace of a network to be simulated a user can define simple objects in the domain as polygons. For more detailed description of a terrain to be considered the MobASim simulator provides the interface to the GeoTools toolkit. The GeoTools [10] is the open source Java coded library containing standard methods for the manipulation of geospatial data, for example to implement GIS (Geographic Information System).

Wireless devices are widely use to establish sensing systems for contaminated areas monitoring. To perform simulations with the MANET system for environmental monitoring it is necessary to model propagation of a pollution. Very dangerous for human beings are clouds created by heavy gases – gases with density greater than that of air. They can move close to the ground for significant time at high level of gas concentration [33].

The models of heavy gas dispersion are divided into categories based on different criteria. Three main groups: empirical, research and engineering models are distinguished [18]. Empirical models are developed based on environmental measurements and laboratory experiments. Research models – formulated by sets of partial differential equations dependent on time and three space coordinates – provide complete and detailed description of the physical process of the heavy gas dispersion. The trade off are engineering models that are widely used in practical applications.

The current version of the MobASim system provides the box model for instantaneous releases to simulate heavy gas dispersion. It belongs to the group of engineering models. It is a simple model that assumes that a gas cloud forms a uniform cylinder. The detailed description is in [23]. Hence, in this section we present the summary of the box model. In general, it is formed by a set of three linear ordinary differential equations:

$$\frac{dc_c}{dt} = v_c, \qquad (11)$$

$$\frac{dr}{dt} = v_f, \qquad (12)$$

$$\frac{dm_a}{dt} = \rho_{air}(\pi r^2)v_t + \rho_{air}(2\pi rh)v_e, \qquad (13)$$

where $c_c = [x_c, y_c]$ denotes the position of the centre of a cloud, $v_c$, $v_f$, $v_e$, $v_t$ denote following velocities: transport, gravitational and entrainment for edge and top of a cloud. $r$ is the radius of a cloud, $h$ its height and $m_a$ is the entrained mass. $\rho_{air}$ denotes the air density.

Equation (11) describes the spreading of the centre of a cloud, (12) the puff horizontal spreading influencing the cloud radius. Mass conservation is described by formula (13).

The gravitational velocity $v_f$ can be calculated for a given standard gravity $g$, the height of a cloud $h$, a cloud $\rho_c$ and an air $\rho_{air}$ densities according to following formula:

$$v_f = C_F\sqrt{\frac{g(\rho_c - \rho_{air})h}{\rho_{air}}}, \qquad (14)$$

where $C_F$ denotes the Froude number of the front. In case of the dense gas models typically assumed $C_F = 1.1$ [23].

The relationship between the edge entrainment velocity $v_e$ and the gravitational velocity is as follows:

$$v_e = \alpha v_f, \qquad (15)$$

where $\alpha \in [0.6, 0.9]$ [18].

Finally, the entrainment velocity for a cloud can be calculated due to the following equation

$$v_t = u_* \left( \frac{\kappa}{1 + \beta \frac{g(\rho_c + \rho_{air})h}{\rho_{air} u_*^2}} \right), \tag{16}$$

where $\kappa = 0,4$ denotes the von Karman constant [18], $u_*$ is the friction velocity [13], $\beta$ denotes the parameter (suggested value $\beta = 0.125$ [3]).

To solve above equations the current cloud density $\rho_c$ and its height $h$ have to be determined. They depend on the concentration $c$ of gas in a cloud. The value of $c$ dynamically changes due to cloud mixing with an ambient air. It can be calculated due to the following formula

$$c = \frac{\frac{m_0}{M}}{\frac{m_0}{M} + \frac{M_a}{M_{air}}}, \tag{17}$$

where $m_0$ denotes a mass of contaminant gas, $M$ and $M_{air}$ the molar weights of the gas and air respectively.

The relation between the gas concentration and its density notable influences a cloud dynamics. This relation is affected by many factors. In our work we assume that mixing of gas with ambient air is the only source of density change. We neglect chemical reactions and occurrence of any aerosol formations in the cloud. Based on these assumptions we can calculate the density of gas

$$\rho_c = \rho_{air} \left( \frac{1 + c\frac{M - M_{air}}{M_{air}}}{1 + \frac{c\Delta H_0}{\left((1-c)M_{air}c_p^{air} + cMc_p\right)T_{air}}} \right), \tag{18}$$

where $\Delta H_0$ denotes the enthalpy difference between the release material at the source and ambient conditions, $c_p$ and $c_p^{air}$ are specific heat capacities of gas and air respectively. $T_{air}$ is the temperature of an ambient air.

Then, the height of a cloud $h$ can be computed

$$h = \frac{V}{\pi r^2}, \qquad V = \frac{m}{\rho_c}, \tag{19}$$

where $V$ is the volume of a cloud.

The Euler method is used to solve the model (11)–(13). It should be pointed that the presented model can be employed to simulate a gas dispersion in a flat area without obstacles. In general, the model is valid until the occurrence of one of two conditions [18]:

- the difference between the density of cloud and air is less than a small assumed value,
- the growth of a cloud radius in single step is small enough.

## IV. Mobile Ad hoc Networks Simulation Platform

MobASim (*Mobile Ad hoc network Simulation*) [25] is a general purpose software framework for wireless, mobile ad hoc networks simulation and a library of Java classes that can be used to build a MANET simulator. Each simulator is designed as federation of disparate simulators of subnetworks that compose the considered MANET or a federation of simulators of independent, geographically dispersed MANETs or WSNs (wireless sensor networks) that cooperate from time to time. The components of such simulators can be easily reused in many computations. MobASim can run on a single or parallel machine or a computer cluster.

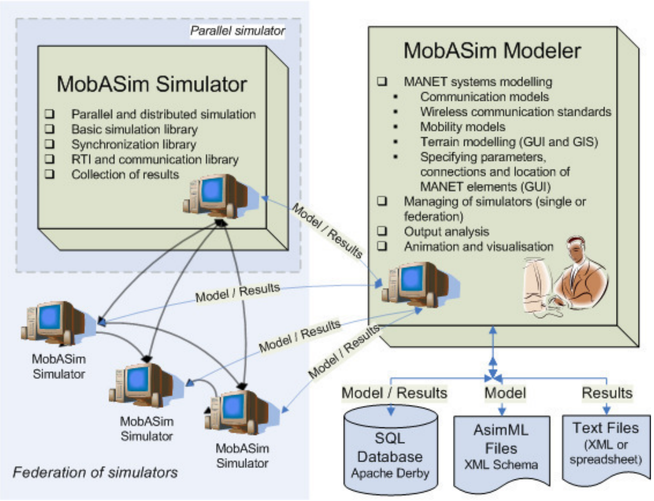

Parallel simulator
MobASim Simulator
❑ Parallel and distributed simulation
❑ Basic simulation library
❑ Synchronization library
❑ RTI and communication library
❑ Collection of results
MobASim Modeler
❑ MANET systems modelling
• Communication models
• Wireless communication standards
• Mobility models
• Terrain modelling (GUI and GIS)
• Specifying parameters, connections and location of MANET elements (GUI)
❑ Managing of simulators (single or federation)
❑ Output analysis
❑ Animation and visualisation
Model / Results
Model / Results
Model / Results
Model / Results
MobASim Simulator
MobASim Simulator
MobASim Simulator
Federation of simulators
Model / Results
Model
Results
SQL Database Apache Derby
AsimML Files XML Schema
Text Files (XML or spreadsheet)

Fig. 2: The MobASim application.

The MobASim system is completely based on Java. At the heart of the MobASim technology is the ASimJava library developed by the authors, and described in [26], [28]. MobASim is composed of four main components presented below.

### A. MobASim Modeler

The model of the network considered can be generated using MobASim GUI or can be loaded from the disc file in the XML format. The configuration of the system is saved into the disc file for reuse in other experiments. The following attributes and parameters have to be provided by the user: number of wireless devices that compose a network, wireless transmission model, MAC protocol, radio communication range, type of the mobility model, minimal and maximal velocity, destination, simulation time horizon.

Three libraries are provided to implement a simulator: wireless transmission modeling library (WTML), Mobility Models Library (MML) and Environment Modeling Library (EML). WMTL contains a collection of classes implementing models of wireless transmission and wireless communication standards. The currently available version of MobASim provides the implementation of the long-distance path loss model described in section III-C and three classes of MAC protocols based on the method that they handle the hidden and the exposed terminal problems. Class 1 – protocol assuming random access to the wireless channel (the hidden and exposed node problem is unsolved), class 2 – the protocol solves the hidden node problem but leaves the exposed node problem unsolved, class 3 – the protocol solves both the hidden node and the exposed node problems, but requires the deployment of an additional signaling channel.

MML contains classes implementing three types of mobility models: random mobility model, waypoint model and group mobility model described in section III-B. All models utilize discrete event systems methodology. The state of each mobile node is described by three variables: location within the deployment region, orientation (an angle between X axis and the direction of a node movement) and velocity.

EML contains a collection of classes that allow to model a simulation scene. They provide interfaces to tools used to create a workspace. This library is under development. We plan to implement models for spread of contaminations and various types of gases dispersion. Nowadays, EML provides the heavy gas dispersion model described in section III-D.

*a) MobASim Simulator:* responsible for the simulator implementation and performance. It provides three libraries: basic library – a collection of classes implementing basic elements of a simulator (logical processes, events, etc.), synchronization protocols library – a collection of classes implementing synchronization algorithms and runtime infrastructure (RTI) for parallel and distributed simulation. The functions from these libraries are basic components of a given simulator.

*b) MobASim Database:* stores all geographical information – a map of a workspace, and all network nodes' positions.

*c) MobASim User Interface:* – an integrated GUI that can help a user to define a simulation scenario, create a network topology and provide all attributes of all devices that form a network to be considered. The GUI is organized in a set of nested setting and presentation windows. The setting windows are used to facilitate the configuration phase. The presentation windows are used to display the calculation results – network animation, statistics, etc.

## V. Case Studies

In this paper we demonstrate some of the main capabilities of the MobASim platform by studying the performance of various ad hoc networks application scenarios concerned with an environmental monitoring and coordination of an emergency action. The results of experiments performed on single machine (AMD Sempron 1.67 GHz, 512 MB RAM) were compared with those obtained on parallel machine (Intel Core2 Duo 2.2 GHz, 2038 MB RAM), and computer cluster consisting of three machines (AMD Sempron, Intel Core2 Duo and AMD Athlon-M 1.2 GHz, 512 MB RAM). In parallel and distributed implementations the networks considered were divided into subnetworks simulated in parallel by three processors.

### A. Rescue Action Coordination

The MobASim software can be used to support the design of mobile ad hoc network for the rescue actions – monitoring of the situation after various disasters: earthquake, fire, flood or explosion, etc. In this paper we present the results of hypothetical usage of MANET to support the rescue action in case of real life explosions, which occurred at Buncefield Oil Storage Depot, Hemel Hempstead, Hertfordshire in December 2005. The damages were significant, a large area around the site was evacuated, forty people were injured. The fire burned for several days, destroying most of the site and emitting large clouds of smoke into the atmosphere.

It is obvious that usually the most of the communication infrastructure, i.e., wired phone lines, base stations for cable networks, etc. are devastated after explosions. Hence, mobile, self-configuring and cooperative wireless network can be used to support the rescue action. Let us consider that we plan to send several rescue teams to work on the disaster scene. A coordination action can be achieved only if rescuers are able to communicate, both within their team, the members of the other teams and an emergency action coordination center. Thus, one of the priorities in the disaster management is to reinstall the communication infrastructure as quickly as possible and monitor the situation. It can be done by deploying temporary communication equipment, e.q. vehicles or robots equipped with radio transceivers and sensors, and form an ad hoc network. Computer simulation can be used to support the decision about the number of devices used to create the given network, their destination positions, and finally number of rescue teams [27].

A series of experiments for various network topologies used for reestablishing the communication infrastructure and monitoring were performed. Each network system was built by three types of nodes:

A – mobile devices equipped with transceivers and a set of sensors,
B – static devices equipped with transceivers and a set of sensors; a few of them were mobile (backup devices),
C – base station (emergency action coordination center).

In this paper the results of tests performed for two network configurations are presented:

- Net 1: 23 nodes A, 12 nodes B, 1 node C (36 LPs).
- Net 2: 81 nodes A, 46 nodes B, 1 node C (81 LPs).

In both cases the maximal velocity of all mobile nodes was equal to 10 m/s. Three variants of implementation – sequential, parallel and distributed on three hardware platforms depicted in Table I were compared. The goal of all experiments was to simulate 900 seconds of physical ad hoc network operation. The execution times of each experiment are given in Table II.

TABLE I: Computer systems: sequential, parallel and distributed implementations.

| Variant | AMD Athlon-M<br>1,2 GHz<br>512 MB RAM | AMD Sempron<br>1,67 GHz<br>512 MB RAM | Intel Core2 Duo<br>2,2 GHz<br>2038 MB RAM |
|---|---|---|---|
| Sequential | | X | |
| Parallel | | | X |
| Distributed | X | X | X |

The last two columns collect the speedup for parallel and distributed calculations, defined as $s(p) = T(1)/T(p)$, where

TABLE II: Performance evaluation – comparison of execution times.

| Ex. | Simulation time [s] | | | s(2) | s(4) |
|---|---|---|---|---|---|
| | S | P | D | P | D |
| Net 1 | 64.2 | 49.2 | 49.9 | 1.31 | 1.29 |
| Net 2 | 182.2 | 152.8 | 118.9 | 1.19 | 1.53 |

$T(1)$ denotes execution time in case when all calculations were performed on a single processor and $T(p)$ execution time in case of $p$ processors or machines.

During the experiments the animation of time varying network topologies – all nodes moving from the source to their destination, avoiding the obstacles – was presented in the MobASim display window. The snapshots of initial, temporary and final network topologies are depicted in Figures 3 and 4.

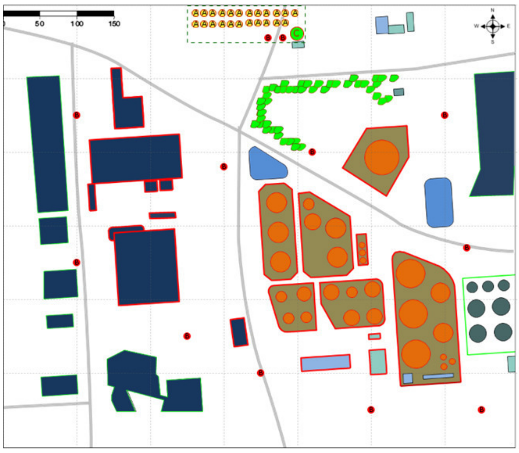


Fig. 3: The snapshot of initial network topology.

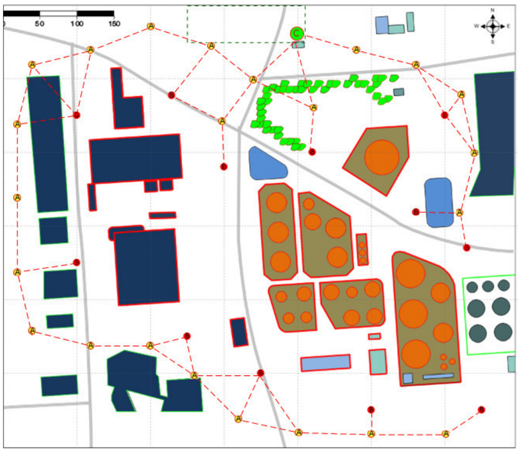


Fig. 4: The snapshot of final network topology.

From the simulation results we see that by using multihop wireless communication and mobile nodes, the communication between the base station and rescuers will be possible without the need for reestablishing the fixed communication infrastructure. We can observe that federated, parallel and distributed simulation developed based upon MobASim software can speedup simulation of MANETs operation w.r.t. sequential implementation. As expected, the calculation speedup depends on the size of network model and assumed degree of parallelism – the speedup factor increases with the problem dimension and complexity. It can be observed that the sequential part of execution concerned with MobASim internal servers and database initializations depends on the size of problem to be considered, and seriously influences the whole computation time. As a final observation, we can point that we obtained the speedup of calculations even in case of small dimension and strongly interconnected network models. We expect much better results for higher dimension networks operating in inherently parallel environments, e.g. cooperating MANETs or sensor networks.

### B. *Gas Cloud Detection and Exploring*

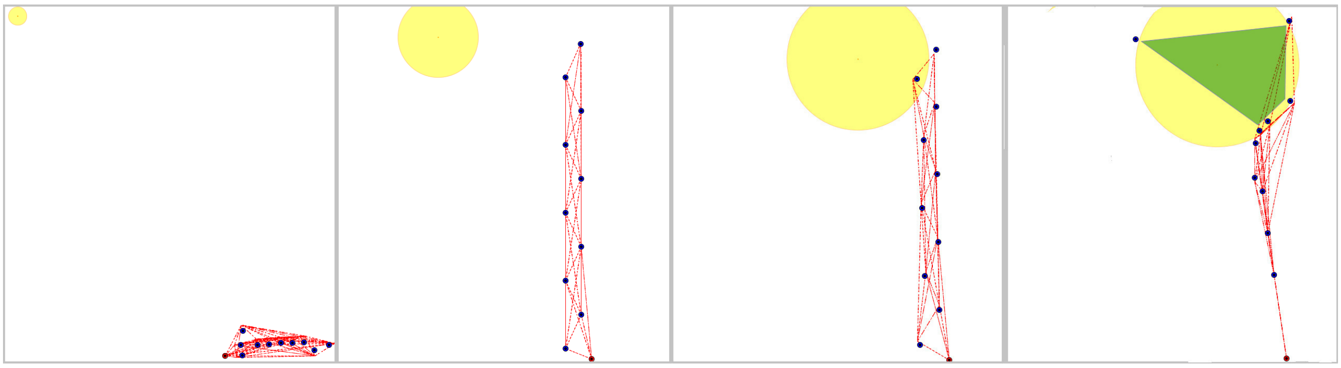
A wireless network formed by mobile devices can reduce lack of situation awareness in area prone to emergencies such as environmental pollution. The MobASim platform was used to simulate various scenarios of application of MANETs for environmental monitoring. The goal of the first series of experiments was to design and develop robot-assisted wireless sensor networks for outdoor and indoor monitoring. The simulation was used to determine the optimal number of sensing devices, calculate collision and obstacle-free motion trajectories for mobile robots carrying the sensor devices and calculate the optimal positions for all sensors (network nodes). Various deployment strategies were considered, i.e., regular and pre-defined grid, self-configuring and hybrid deployments. The results are described and discussed in [27]. The objective of the second series of experiments was to design and develop a MANET for detection, exploration and tracking of a heavy gas cloud. Such cloud may be created as an effect of an instantaneous release of gas heavier than air from a tank [4].

Let us assume that the emergency team has to find, surround and suppress a chlorine gas cloud. The network formed by mobile devices equipped with punctual gas sensors can successfully support the team in detecting and surrounding the cloud. Hence, the set of self-organized mobile devices were deployed in a suspicious area, and formed a coherent network to discover the cloud. After detecting the cloud the devices were divided into two groups. The devices from the first group were responsible for exploring and discovering the shape of the cloud. The task of the remaining devices was to maintain communication between the first group of sensors and the base station.

The results of simulations performed with the example ad hoc network are shown in Fig. 5. The values of all parameters occurring in the equations (18)-(13) are given in Table III [12], [34], [23].

## VI. Summary and Conclusions

The evolution of wireless, mobile ad hoc networks and improved designs will strongly depend on the ability to predict their performance using simulation methods. In this paper we described the application of our software platform MobASim to design mobile ad hoc networks used to detect threats and support decision makers in emergency situations. The presented case studies confirm that the MobASim system can support researches and engineers during the design and implementation of MANETs applications and verification of new MANET's technologies. The tool is especially useful in



(a) (b) (c) (d)

Fig. 5: MANET for gas cloud detection and tracking: (a) initial network topology, (b) cloud detection, (c) cloud was detected, (d) exploring the cloud and discovering its shape (in green the detected cloud shape).

TABLE III: Values of parameters of the chlorine gas cloud simulation model.

| Symbol | Value | Units |
|---|---|---|
| $c_c$ | [200, 200] | $[m,m]$ |
| $v_c$ | [3, 1] | $[\frac{m}{s}, \frac{m}{s}]$ |
| $r$ | 10 | $m$ |
| $m_0$ | 2000 | $kg$ |
| $m_a$ | 0 | $kg$ |
| $g$ | 9.81 | $\frac{m}{s^2}$ |
| $\rho_{air}$ | 1.20 | $\frac{kg}{m^3}$ |
| $M$ | 71 | $\frac{g}{mol}$ |
| $M_{air}$ | 29 | $\frac{g}{mol}$ |
| $c_p$ | 0.48 | $\frac{kJ}{kg*K}$ |
| $c_p^{air}$ | 1.01 | $\frac{kJ}{kg*K}$ |
| $u_*$ | 0.15 | $\frac{m}{s}$ |
| $\Delta H_0$ | 661 | $\frac{kJ}{kg}$ |
| $T_{air}$ | 293.15 | $K$ |
| $\alpha$ | 0.9 | - |

large scale applications in which the speed of simulation is of essence, such as real time ad hoc networks simulation.

## REFERENCES

[1] Aggelou G., *Mobile Ad Hoc Networks. From Wireless LANs to 4G Networks*, McGraw-Hill, USA 2005.

[2] Basagni S., Conti M., Giordano S. and Stojmenovic, I., *Mobile Ad Hoc Networking*, Wiley-Interscience, IEEE Press, 2004.

[3] Britter, R.E. and Simpson, J.E., *Experiments on the dynamics of a gravity current head*, Journal of Fluid Mechanics, Vol. 88, pp. 223-240, 1978. doi:10.1017/S0022112078002074

[4] Cameron, I. and Raman, R., *Process Systems Risk Management*, Elsevier Academic Press, pp. 246–247, 2005.

[5] Choset, H., Lynch, K.M., Hutchinson, S., Kantor, G., Burgard, W., Kavraki, L.E., and Thrun, S., *Principles of Robot Motion*, The MIT Press, Cambridge, 2005.

[6] Daniluk, K., and Niewiadomska-Szynkiewicz, E., *Energy-efficient security in Implantable Medical Devices*, in Proc. FedCSIS, pp.773-778, 2012.

[7] Ferenci, S.L., Perumalla, K.S. and Fujimoto, *An Approach for Federating Parallel Simulators*, Proc. of PADS 2000, Bologna, pp. 63–70. doi:10.1109/PADS.2000.847145

[8] Fongen, A., Gjellerud, M. and Winjum, E., *A Military Mobility Model for MANET Research*, Proc. of IASTED International Conference on Parallel and Distributed Computing and Networks, 2009.

[9] Forouzan, B.A., *Data Communications and Networking*, McGraw-Hill, 2004.

[10] GeoTools The Open Source Java GIS Toolkit, http://geotools.codehaus.org/ .

[11] GloMoSim homepage, http://pcl.cs.ucla.edu/projects/glomosim/ .

[12] Hanna, S.R. and Drivas P.J., *Guidelines for use of vapour cloud dispersion models*, Center for Chemical Process Safety, Institute of Chemical Engineers, New York, 1989.

[13] HGSYSTEM, *The Heavy Gas Dispersion Model HEGADAS*, HGSYSTEM Technical Reference Manual, http://www.hgsystem.com/tech_ref/Chap07.pdf .

[14] Kasch, W.T., Ward, J.R. and Andrusenko, J., *Wireless Network Modeling and Simulation Tools for Designers and Developers*, Comm. Mag., Vol. 47, Issue 3, pp. 120–127, IEEE Press, March 2009. doi:10.1109/MCOM.2009.4804397

[15] Kasprzak, W., Szynkiewicz, W., Karolczak, M., *Global colour image features for discrete self-localization of an indoor vehicle*, Lecture Notes in Computer Science 3691, pp. 620627, Springer Verlag, Berlin Heidelberg, 2005.

[16] Kurkowski, S., Camp, T. and Colagrosso, M., *MANET Simulation Studies: The Incredibles*, SIGMOBILE Mob. Comput. Commun. Rev., Vol. 9, Issue 4, pp. 50–61, ACM, 2005. doi:10.1145/1096166.1096174

[17] Mao, G., Fidan, B. *Localization algorithms and strategies for wireless sensor networks*, Inforamtion Science Reference, USA, 2009. doi:10.1007/BF02136831

[18] Markiewicz, M.T.,*Mathematical modeling of heavy gas atmospheric dispersion over complex and obstructed terrain*, Archives of Environmental Protection, Vol. 36, no. 1, pp. 81–94, 2010.

[19] Musolesi, M. and Mascolo, C., *Mobility Models for Systems Evaluation. A Survey*, State of the Art on Middleware for Network Eccentric and Mobile Applications (MINEMA), Springer, 2009.

[20] Network Simulator ns-2 homepage, http://www.isi.edu/nsnam/ns/.

[21] Network Simulator ns-3 homepage, http://www.nsnam.org/.

[22] Nicol, D.M. and Fujimoto, R. (1994) *Parallel Simulation Today*, Annals of Operations Research, Vol. 53, pp. 249–285. doi:10.1007/BF02136831

[23] Nielsen, M., *Dense Gas Dispersion in the Atmosphere*, Riso National Laboratory, Roskilde, Denmark, 1998.

[24] Niewiadomska-Szynkiewicz, E. and Sikora A., and Kołodziej, J., *Modeling Mobility in Cooperative Ad Hoc Networks*, Mobile Networks and Applications, Vol. 18, No. 5, pp. 610-621, 2013. doi:10.1007/s11036-013-0450-2

[25] Niewiadomska-Szynkiewicz, E., Sikora, A., *A Software Tool for Federated Simulation of Wireless Sensor Networks and Mobile Ad Hoc Networks*, Applied Parallel Scientific Computing K. Jonasson (ed.), LNCS7133, part I, pp. 303-313, Springer-Verlag, 2012.

[26] Niewiadomska-Szynkiewicz, E. and Sikora A., *ASim/Java: A Java-based Library for Distributed Simulation*, Journal of Telecommunications and Information Technology, No. 3, pp. 12-17, 2004.

[27] Niewiadomska-Szynkiewicz, E. and Sikora, A., *Simulation-Based Evaluation of Robot-Assisted Wireless Sensors Positioning*, Progress in Automation, Robotics and Measuring Techniques, Vol. 351, pp. 181-190, Springer International Publishing, 2015. doi:10.1007/978-3-319-15847-1_18

[28] Niewiadomska-Szynkiewicz, E. and Sikora, A., *A Federated Approach to Parallel and Distributed Simulation of Complex Systems*, International Journal of Applied Mathematics and Computer Science, Vol.17, Issue 1, pp. 99106, 2007. doi:10.2478/v10006-007-0009-0

[29] Niewiadomska-Szynkiewicz, E. *Localization in Wireless Sensor Networks: Classification and Evaluation of Techniques*, International Journal of Applied Mathematics & Computer Science, University of Zielona Gora Press, vol. 22, No 2. pp. 281-297, 2012.

[30] OMNeT++ homepage http://www.omnetpp.org/.

[31] QualNet homepage http://web.scalable-networks.com/content/qualnet.

[32] Santi, P., *Topology Control in Wireless Ad Hoc and Sensor Networks*, John Wiley & Sons, Ltd, 2006.

[33] Scargiali, F., Di Rienzo, E., Ciofalo, M., Grisafi, F., and Brucato, A., *Heavy Gas Dispersion Modelling Over a Topographically Complex Mesoscale: A CFD Based Approach*, Process Safety and Environmental Protection, Volume 83, Issue 3, Pages 242-256, ISSN 0957-5820, May 2005. doi:10.1205/psep.04073

[34] Schmittinger, P., Florkiewicz, T., Curlin, L.C., Lke, B., Scannell, R., Navin, T., Zelfel, E. and Bartsch, R., *Chlorine*, Ullmann's Encyclopedia of Industrial Chemistry, Wiley-VCH Verlag, 2006.

[35] Rappaport, T.S., *Wireless Communications. Principles and Practice*, Prentice Hall, USA, 2009.

[36] Riverbed Modeler (OPNET) homepage http://www.riverbed.com/products/performance-management-control/network-performance-management/network-simulation.html.

[37] Roy, R.R., *Handbook of Mobile Ad Hoc Networks for Mobility Models*, Springer, USA, 2010.

[38] Van Haute, T., Rossey, J., Becue, P., De Poorter, E., Moerman, I., and Demeester, P., *A hybrid indoor localization solution using a generic architectural framework for sparse distributed wireless sensor networks*, Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Vol. 2, pp. 1009–1015, 2014. doi:10.15439/2014F20

[39] V-rep (Virtual Robot Experimental Platform) homepage, http://www.coppeliarobotics.com/.

[40] Zeigler, B.P., Praehofer, H. and Kim, T.G., *An Overview of MANETs Simulation*, Jour. of Electronic Notes in Theoretical Computer Science (ENTCS), Vol. 150, Issue 1, pp. 81-101, 2006. doi:10.1016/j.entcs.2005.12.025

# 2nd Workshop on Emerging Aspects in Information Security

ADMITTEDLY,information security works as a backbone for protecting both user data and electronic transactions. Protecting the communication and data infrastructure of an increasingly inter-connected world has become vital nowadays. Security has emerged as an important scientific discipline whose many multifaceted complexities deserve the attention and synergy of the computer science, engineering, and information systems communities. Information security has some well-founded technical research directions which encompass access level (user authentication and authorization), protocol security, software security, and data cryptography. Moreover, some other emerging topics related to organizational security aspects have appeared beyond the long-standing research directions.

The Emerging Aspects in Information Security (EAIS'15) workshop focuses on the diversity of the information security developments and deployments in order to highlight the most recent challenges and report the most recent researches. The workshop is an umbrella for all information security technical aspects. In addition, it goes beyond the technicalities and covers some emerging topics like social and organizational security research directions. EAIS'15 is intended to attract researchers and practitioners from academia and industry, and provides an international discussion forum in order to share their experiences and their ideas concerning emerging aspects in information security met in different application domains. This opens doors for highlighting unknown research directions and tackling modern research challenges. The objectives of the EAIS'15 workshop can be summarized as follows:

- To review and conclude researches in information security and other security domains, focused on the protection of different kinds of assets and processes, and to identify approaches that may be useful in the application domains of information security
- To find synergy between different approaches, allowing to elaborate integrated security solutions, e.g. integrate different risk-based management systems
- To exchange security-related knowledge and experience between experts to improve existing methods and tools and adopt them to new application areas
- To present latest security challenges, especially with respect to EC Horizon 2020

## Topics

Topics of interest include but are not limited to:

-
- Biometric technologies
- Human factor in security
- Cryptography and cryptanalysis
- Critical infrastructure protection
- Hardware-oriented information security
- Social theories in information security
- Organization- related information security
- Pedagogical approaches for information security
- Individual identification and privacy protection
- Information security and business continuity management
- Decision support systems for information security
- Digital right management and data protection
- Cyber and physical security infrastructures
- Risk assessment and risk management in different application domains
- Tools supporting security management and development
- Emerging technologies and applications
- Digital forensics and crime science
- Misuse and intrusion detection
- Security knowledge management
- Data hide and watermarking
- Cloud and big data security
- Computer network security
- Security and safety
- Assurance methods
- Security statistics

## Event Chairs

**Awad, Ali Ismail,** Luleå University of Technology, Sweden

**Bialas, Andrzej,** Institute of Innovative Technologies EMAG, Poland

## Program Committee

**AbdAllah, Mohamed Mostafa,** Yanbu Industrial College, Saudi Arabia

**Bun, Rostyslav,** Lviv Polytechnic National University

**Clarke, Nathan,** Plymouth University, United Kingdom

**Cyra, Łukasz,** European Commission - Joint Research Centre Institute for the Protection & Security of the Citizen

**Dworzecki, Jacek,** Police Academy in Szczytno

**Fernandez, Eduardo B.,** Florida Atlantic University, United States

**Furnell, Steven,** Plymouth University, United Kingdom

**Furtak, Janusz,** Military University of Technology, Poland

**Geiger, Gebhard,** Technical University of Munich, Faculty of Economics

**Grzenda, Maciej,** Orange Labs Poland and Warsaw University of Technology, Poland

**Hämmerli, Bernhard M.,** Hochschule für Technik+Architektur (HTA), Switzerland

**Hasssaballah, M.,** South Valley University, Egypt

**Kalbarczyk, Zbigniew,** University of Illinois at Urbana-Champaign

**Kapczynski, Adrian,** Silesian University of Technology, Poland

**Klamka, Jerzy,** Polish Academy of Sciences

**Kosmowski, Kazimierz,** Gdansk University of Technology

**Mamojka, Mojmír,** Police Academy in Bratislava

**Pańkowska, Małgorzata,** University of Economics in Katowice, Poland

**Rot, Artur,** Wroclaw University of Economics, Poland

**Soria-Rodriguez,** Pedro, Atos Research & Innovation

**Stokłosa, Janusz,** Poznań University of Technology, Poland

**Suski, Zbigniew,** Military University of Technology

**Szmit, Maciej,** Orange Labs Poland, Poland

**Thapa, Devinder,** Luleå University of Technology

**Yen, Neil,** The University of Aizu, Japan

**Zamojski, Wojciech,** Wrocław University of Technology

**Zieliński, Zbigniew,** Military University of Technology, Poland

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1099–1106

DOI: 10.15439/2015F77
ACSIS, Vol. 5

# Experimentation tool for critical infrastructures risk management

Andrzej Bialas
Institute of Innovative
Technologies EMAG,
ul. Leopolda 31,
40-189 Katowice, Poland
Email: andrzej.bialas@ibemag.pl

*Abstract*—**The paper concerns a risk assessment and management methodology in critical infrastructures. The research objective is to adapt a ready-made risk manager, supporting information security- and business continuity management systems, to a new domain of application – critical infrastructure protection. First, a review of security issues in critical infrastructures was performed, with special focus on risk management. On this basis the assumptions were discussed how to adapt the OSCAD risk manager designed for the information security/business continuity applications. According to these assumptions, the OSCAD risk manager was adapted to its new domain of application, i.e. critical infrastructures. The aim of this work is to assess the usefulness of such a solution and to elaborate requirements for the advanced critical infrastructure risk manager to be developed from scratch.**

## I. Introduction

CRITICAL infrastructures (CIs) consist of large scale infrastructures whose degradation, disruption or destruction would have a serious impact on health, safety, security or well-being of citizens or effective functioning of governments and/or economies. Typical examples of such infrastructures are energy-, oil-, gas-, finance-, transport-, telecommunications-, and health sectors. CIs provide products and services for the society. In order to function, CIs need many different assets. What is more, they are based on complex processes interrelated with other processes across different sectors. CIs are extremely important for effective functioning of today's societies, especially those of well-developed countries. Critical infrastructures ensure proper relationships between citizens and governments. Each society is very sensitive to any disturbance of a CI. Security and safety issues are very important, but due to the CIs complexity, multi-dimensional interdependencies, large scale and heterogeneity, the problems emerging in these areas are often hard to solve.

A CI can be disturbed by different kinds of threats and hazards. The most important are: natural disasters and catastrophes, technical disasters and failures, espionage, international crime, physical- and cyber terrorism. A new, holistic approach to CI protection is applied by programmes and activities which are understood as critical infrastructure protection (CIP). It is a common effort of the infrastructure owners and operators, manufacturers, users, R&D institutions, governments, international bodies and regulatory authorities. The aim of these efforts is to keep the performance of CIs in case of failures, attacks or accidents and minimize the recovery time and damages.

Well developed countries, including the EU countries, pay more and more attention to the protection of their critical infrastructures. The European Council (EC) Directive [1] specifies the CIP related needs on the EU and member state levels. It precisely defines the rules of the CI identification based on casualties-, economic- and public criteria, risk analysis and management programmes. The EC Directive defines the term ECI (European critical infrastructure). ECI means a critical infrastructure located in member states, whose disruption or destruction would have a significant impact on at least two member states. There are two ECI sectors distinguished:

- energy (electricity, oil, gas),
- transport (road transport, rail transport, air transport, inland waterways transport, ocean and short-sea shipping and ports).

The European Programme for Critical Infrastructure Protection (EPCIP), aimed at both European and national infrastructures, was launched in 2006. The revised and more practical implementation of EPCIP is presented in the EU document [2].

Risk assessment is the basis for critical infrastructures protection programmes. Dozens of EU or worldwide CIP R&D projects which focus on risk methodologies and tools have been completed or are running (FP6, FP7, Horizon 2020, CIPS), which is a proof that the CIP issue is still a challenge.

The researches presented in the paper can be considered preliminary activities of the Ciras[1] project [3]. Ciras was

This work was not supported by any organization

[1] This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the European Commission cannot be held responsible for any use which may be made of the information contained therein (Grant Agreement clause).

978-83-60810-66-8/$25.00©2015, IEEE

launched by the international consortium (ATOS, CESS, EMAG) including the author's organization.

The motivation for researches presented in the paper is to elaborate the experimental OSCAD-Ciras tool to get an input for the Ciras project. Particularly, the following aims are planned to be accomplished:

- to implement the requirements [4] on the OSCAD software platform [5] elaborating the CI risk manager to be used as an experimental platform,
- to assess, using near real data, if the basic requirements of the risk manager can be implemented on the OSCAD software, and
- to summarize the whole experiment, acquiring indispensable knowledge about the usability of this risk manager and to identify directions of the future works.

Apart from the requirements identified during the stakeholders' workshop and the reviewed state of the art of the risk management methodology, the Ciras project will get input from the results of this OSCAD-Ciras feasibility study. This input will be used for the Ciras Toolset development.

The paper includes an introduction to risk management in critical infrastructures (section II), summarizes the preferred features of the risk management tool discussed in the work [4] (section III), presents the functionality of the OSCAD software platform (section IV), gives the specifics of OSCAD's adaptation to be a CI risk manager (section V), and finally draws some conclusions for future works.

## II. Risk management in critical infrastructures protection

Critical infrastructure is a heterogeneous, distributed, adaptive, and, first and foremost, very complex socio-technical system. Such a system encompasses hardware, software, liveware, environmental, management, and organizational elements. The main objective of a CI is to provide products and/or services for the society. This aim can be accomplished when this complex socio-technical system is well harmonized and the disturbances within the system are under control – the system processes its work smoothly and the assets needed to perform this job are well protected. The CI countermeasures, selected on the risk basis, should be properly managed and composed into CIP programmes.

Collaborating critical infrastructures (systems), e.g. electricity, rail transport, gas, oil, telecommunications, constitute a more complex structure, called a system-of-systems (SoS).

Different mutual dependencies (i.e. interdependencies) between particular CIs exist within SoS too. An interdependency [6] is a bidirectional relationship between two infrastructures (systems) through which the state of each infrastructure influences or is correlated to the state of the other [7].

The CIs failures are usually causally linked – the impacts of incidents may pass across the CIs. Certain CI-specific effects are observed. A cascading effect is [8] a sequence of component failures: the first failure shifts its load to one or more nearby components – these components fail and, in turn, shift their loads to other components. This sequence is repeated. An escalating failure is when there is a disruption in one infrastructure which causes an independent disruption in another infrastructure [6]. The effects of hazardous events may escalate outside the area where they occur and exacerbate the consequences of a given event (generally in the form of increasing the severity or the time for recovery of the second failure). Different failures may have a common cause (failures implied by a single shared cause and coupling to other systems mechanisms) and may occur almost concurrently. An important issue is the CI resilience, which is understood as an ability of a system to react to and recover from unanticipated disturbances and events.

The critical infrastructure protection concept comprises preparedness and response to serious incidents which occur within critical infrastructures. To ensure the preparedness and incident response ability, it is necessary to imply the risk source, character and value. In addition, the right countermeasure should be applied and embedded into the risk management framework, sometimes supported by tools.

The comprehensive approach to risk management in critical infrastructures still remains a challenge, due to CIs complexity, interdependencies, specific effects (common cause failures, cascading, escalating effects), different abstract levels applied to manage CIs, and other factors.

The risk management methodology and tools are a subject of current R&D on the national and international levels, including the EU level. Very comprehensive reviews of R&D results can be found in the following knowledge sources:

- the report [9] of the Institute for the Protection and Security of the Citizen, an EC Joint Research Centre (JRC); the report assesses and summarizes 21 existing risk management methodologies/tools on the EU and global level; it identifies their gaps and prepares the ground for R&D in this field, like Ciras project [3];
- the EURACOM report [10]; it presents a study of 11 risk assessment methodologies related to the energy sector;
- the book [7]; in its Appendix C there is a comparison of the features of about 22 commonly used risk analysis methods;
- the ISO 31010 standard [11] characterizes about 30 risk assessment methods for different applications;
- the ENISA website [12] includes an inventory of risk management/assessment methods, mostly ICT-focused.

A very exhaustive review of the state of the art is reported in [13]. To select the most favourable methods/tools features for implementation during the Ciras project, the document summarizes the assessment of: 14 methods (from 46 preselected), 22 tools (from 150 preselected) and considers 19 projects and 8 frameworks.

Usually, methods/tools are focused on the confined domain and they do not address properly the holistic view and resilience. The problem is how to consider CIs interdependencies in the risk management process. This requires to distinguish the internal and external causes of hazardous events as well as the internal and external consequences implied by these events.

The survey on the representative methodologies and tools for the CI risk management was made in [4]. Based on these researches, the most favourable features of the CI risk manager are specified in the next section.

## III. Preferred Features of Risk Management Tools for Critical Infrastructures

The paper [4] discusses the basic requirements of the risk manager to be applied in critical infrastructure protection. This section gives a short overview of these issues.

### A. *Conceptual model of the risk manager*

The implementation of the bow-tie risk concept in the tool is advantageous for CI risk management [4].

The bow-tie conceptual model [8] embraces both multiple and complex causes of the given hazardous event and its diversified and multidirectional consequences (Fig. 1). The triggered hazards or threats, which exploit certain vulnerabilities, can degrade proactive barriers (countermeasures) existing in the system. As a result, an event may occur which is hazardous for assets. The consequences of such an event are usually diversified and multidirectional. To mitigate them, reactive barriers are applied. These barriers can be weakened or even removed by vulnerabilities. Generally, barriers are identified with different kinds of countermeasures. The countermeasures are applied with respect to the risk value and are monitored and maintained – according to the risk management principles. The bow-tie model is focused on risk assessment and can be used to reassess the risk after new/updated barriers are applied.

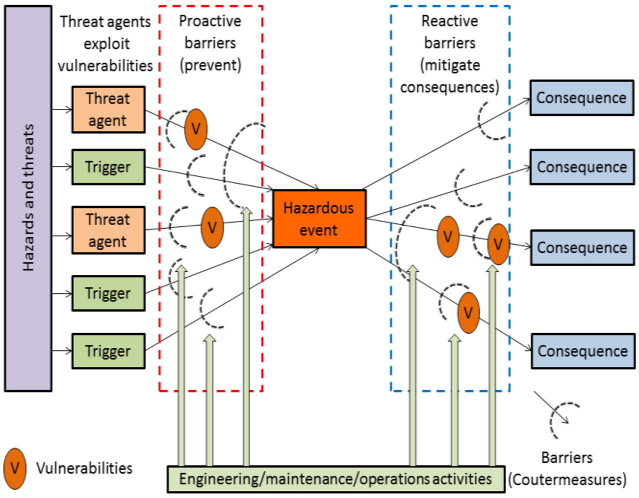

Hazards and threats
Threat agents exploit vulnerabilities
Proactive barriers (prevent)
Reactive barriers (mitigate consequences)
Threat agent
Trigger
Threat agent
Trigger
Trigger
V
Hazardous event
Consequence
Consequence
Consequence
Consequence
V Vulnerabilities
Engineering/maintenance/operations activities
Barriers (Coutermeasures)

Fig. 1 Bow-tie model

The bow-tie model encompasses the cause analysis and the consequences analysis. These risk analyses can be implemented in less or more complex ways [11], e.g. using FTA (Fault tree analysis) [14] and ETA (Event tree analysis) [15].

There is no analysis of interdependencies in this model, therefore it is necessary to supplement the model in this respect.

### B. *Risk register and risk related data*

The tool should support a CI owner in elaborating and maintaining a risk register as the managed inventory of hazardous events. The listed items (data records) should include at a minimum: related hazards/threats, a possible corresponding hazardous event, probability of the event and its consequences. The risk management process is performed during the CI life cycle, so the risk register can be continuously updated. It is used in CIP programmes. There are some data associated with each item of the risk register, like assets, societal critical functions (SCF) ensuring the basic needs of a society (e.g.: life and health, energy supply, law and order, national security), hazards, threats, vulnerabilities, countermeasures, etc.

### C. *Risk measures and the assessment process*

Risk measures depend on the applied methodology. A common method is to assess the likelihood (probability, frequency) of a hazardous event, e.g.: fairly normal, occasional, possible, remote, improbable, and to assess the consequence severity in different dimensions using the enumerative scales, e.g.: negligible, minor, major, catastrophic damages. The risk is the function of both, usually expressed by a risk matrix, as presented in [4].

### D. *Interdependencies and critical infrastructure specific phenomena*

The risk assessment/management methods/tools (Section II) are focused on the given environment with protected assets and processes, and they do not consider interdependencies between other environments. The interdependencies ought to be considered in the risk management process because they are essential for the CI protection. The risk assessment methodology should be able to take into account the CI specific phenomena mentioned in Section II.

## IV. Functionality of the OSCAD Software Platform

The identified requirements are experimentally implemented on the OSCAD[2] platform [5]. The OSCAD software was originally elaborated to support business continuity management according to BS 25999 (ISO 22301)

[2] developed at the Institute of Innovative Technologies EMAG within a project co-financed by the National Centre for Research and Development (NCBiR).

and information security management according to ISO/IEC 27001. It is used to control factors which disturb business processes or breach information assets in an institution (business, public) leading to negative consequences, to limit losses when an incident occurs, and to help in the recovery process.

The solution is open and flexible and thus, after certain modifications, possible to implement in other application domains, e.g.: flood protection [16], railway safety management systems [17] and coal mining [18].

OSCAD offers the following functions:

- general purpose functions: system management, document and tasks management, reporting, dictionaries, business continuity planning, auditing, etc.;
- functions allowing to assess the system effectiveness: acquiring data, assessing effectiveness, permanent improvement actions;
- external communications functions with ERP, SCADA, GSM.

Additionally, OSCAD offers risk management and incident management functions, which are discussed here in the CI context.

OSCAD is equipped with tools to analyze causes of hazardous events:

- AORA – Asset Oriented Risk Analyzer,
- PORA – Process Oriented Risk Analyzer,

and tools analyzing their multidimensional consequences:

- ABIA – Asset Oriented Business Impact Analyzer,
- PBIA – Process Oriented Business Impact Analyzer.

Countermeasures are selected based on the assessed risk value and their total investment/maintenance costs. Then the risk is reassessed with respect to the acceptance level.

The incident management functions allow for events acquisition. They also enable to assess their severity according to the elaborated criteria. Serious events, which are incidents, are managed according to standards. The incident statistics and corrective actions are prepared too.

## V. Implementation of risk manager requirements on the OSCAD software

The section discusses the author's proposals how to implement the above-listed requirements into the existing OSCAD software platform.

### A. *Bow-tie model implementation in the OSCAD software platform*

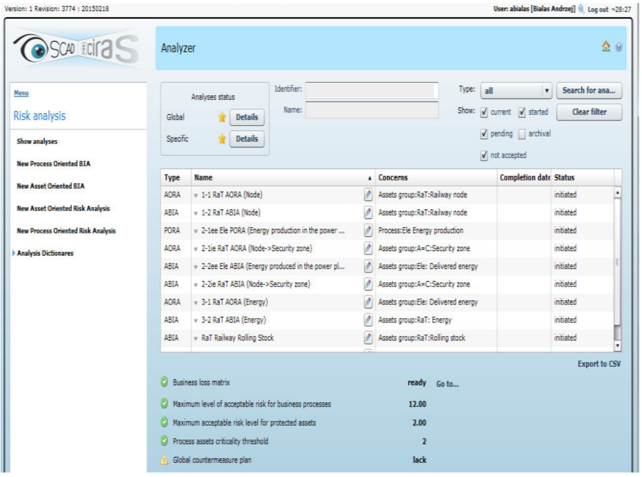
The bow-tie model is not directly implemented in the OSCAD software but its existing risk analyzing tools can be used to compose it.

The cause analysis part of the bow-tie model is implemented on the basis of AORA or PORA. AORA analyzes each threat-vulnerability pair which can breach the given asset, while PORA does the same with respect to the given process.

The consequences analysis part of the bow-tie model is implemented on the basis of ABIA or PBIA. For a given asset (process), which is under a hazardous event, multi-dimensional consequences can be assessed with the use of the loss matrix.

Both parts of the bow-tie model are not coupled directly by the hazardous event, but by the threatened asset (or process) related to this event.

Examples of analyses pairs composing the bow-tie model are shown in Fig. 2. The "1-1 RaT AORA (Node)" and "1-2 RaT ABIA (Node)" create one of pairs related to the railway node belonging to the Railway transport (RaT) European critical infrastructure (ECI) [1].



Version: 1 Revision: 3774 : 20150218    User: abialas [Bialas Andrzej]  Log out ~28:27

Global ★ Specific ★

Identifier:
Name:
Type: all
Show: ☑ current ☑ started ☑ pending □ archival ☑ not accepted

| Type | Name | Concerns | Completion date | Status |
|---|---|---|---|---|
| AORA | 1-1 RaT AORA (Node) | Assets group:RaT:Railway node | | initiated |
| ABIA | 1-2 RaT ABIA (Node) | Assets group:RaT:Railway node | | initiated |
| PORA | 2-1ee Ele PORA (Energy production in the power ... | Process:Ele Energy production | | initiated |
| AORA | 2-1ie RaT AORA (Node->Security zone) | Assets group:A=C:Security zone | | initiated |
| ABIA | 2-2ee Ele ABIA (Energy produced in the power pl... | Assets group:Ele: Delivered energy | | initiated |
| ABIA | 2-2ie RaT ABIA (Node->Security zone) | Assets group:A=C:Security zone | | initiated |
| AORA | 3-1 RaT AORA (Energy) | Assets group:Ele: Delivered energy | | initiated |
| ABIA | 3-2 RaT ABIA (Energy) | Assets group:RaT: Energy | | initiated |
| ABIA | RaT Railway Rolling Stock | Assets group:RaT:Rolling stock | | initiated |

| | |
|---|---|
| Business loss matrix | ready |
| Maximum level of acceptable risk for business processes | 12.00 |
| Maximum acceptable risk level for protected assets | 2.00 |
| Process assets criticality threshold | 2 |
| Global countermeasure plan | lack |

Fig. 2 OSCAD risk analyses composing the bow-tie model

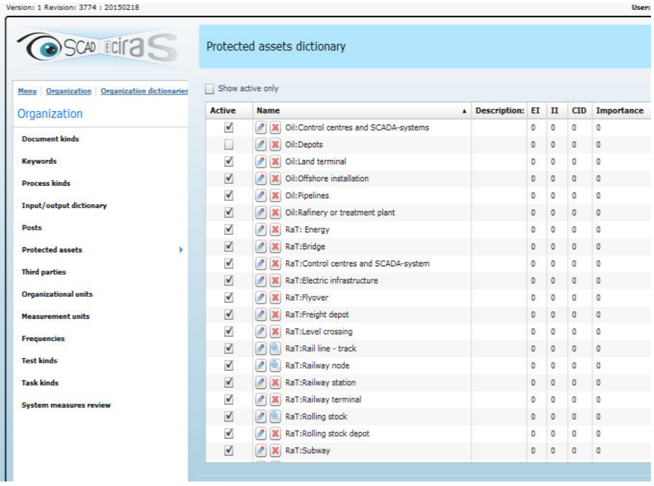
The bow-tie model is rather asset-oriented, similarly to the risk analysis in CIs. For this reason AORA/ABIA may be more convenient for CIs. The given AORA analysis groups the threats related to the given asset. Threats and hazards have the same representation – they are simply the "OSCAD threats". The PORA/PBIA pair represents the process approach. It allows to see a CI from a point of a view of processes, not only assets. Process-oriented risk analyses in CIs need further research.

### B. *Risk register and risk related data – the OSCAD representation*

The basic risk-related data are assets belonging to the critical infrastructures which need protection.

The general CIs taxonomy and assets are implemented in OSCAD. Two groups of CIs are distinguished: ECI (European CI), embraced by the EU Directive [1] and others (non-ECI). Currently only the ECI ones are implemented.

In OSCAD the protected assets dictionary is a simple flat list. For this reason, the assets belonging to the given CI are preceded by a label standing for a CI name: Ele (Electricity), Oil (Oil), Gas (Gas), RoT (Road Transport), RaT (Rail Transport), AiT (Air Transport), IWT (Inland Waterways Transport), Sea (Ocean and short-sea shipping and ports). Figure 3 presents different assets, belonging to the ECI, distinguished according to their labels. Each protected asset can be the central point of any AORA or ABIA. They play a role of primary assets. Please note three attributes CID (CI degradation), IE (Internal escalations), EE (external escalations) which express three types of consequences when the given asset is breached (this will be explained later, when ABIA/PBIA will be discussed).



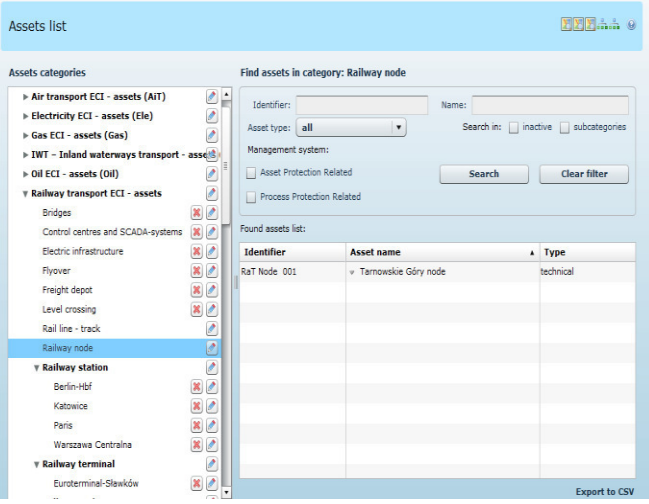
☐ Show active only

| Active | Name | Description: | EI | II | CID | Importance |
|---|---|---|---|---|---|---|
| ☑ | Oil:Control centres and SCADA-systems | | 0 | 0 | 0 | 0 |
| ☐ | Oil:Depots | | 0 | 0 | 0 | 0 |
| ☑ | Oil:Land terminal | | 0 | 0 | 0 | 0 |
| ☑ | Oil:Offshore installation | | 0 | 0 | 0 | 0 |
| ☑ | Oil:Pipelines | | 0 | 0 | 0 | 0 |
| ☑ | Oil:Rafinery or treatment plant | | 0 | 0 | 0 | 0 |
| ☑ | RaT: Energy | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Bridge | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Control centres and SCADA-system | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Electric infrastructure | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Flyover | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Freight depot | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Level crossing | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Rail line - track | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Railway node | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Railway station | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Railway terminal | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Rolling stock | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Rolling stock depot | | 0 | 0 | 0 | 0 |
| ☑ | RaT:Subway | | 0 | 0 | 0 | 0 |

Fig. 3 Protected assets of ECIs in the OSCAD software

It is possible to create a group of the related secondary assets (technical, personal, immaterial, playing role of countermeasures, etc.) around the given primary asset. This group of assets can be defined in the assets inventory module (Fig. 4). The Railway node asset is represented by a node located in the city of Tarnowskie Góry (Upper Silesia, Poland).

For each of the protected assets, the AORA analysis can be performed. PORA can be done for the processes (not discussed here) in a similar way.



Find assets in category: Railway node

Identifier:
Name:
Asset type: all
Search in: ☐ inactive ☐ subcategories
Management system:
☐ Asset Protection Related
☐ Process Protection Related

Found assets list:

| Identifier | Asset name | Type |
|---|---|---|
| RaT Node 001 | Tarnowskie Góry node | technical |

Fig. 4 Different assets belonging to the given protected assets category

To perform a risk analysis for different barriers, security zones, etc., which play the role of countermeasures, an auxiliary category is defined: A=C (Countermeasures considered as assets), for example A=C:Security zone can be added to the Railway node, and an additional risk analysis for it can be performed when internal escalation effects are analyzed (will be explained further).

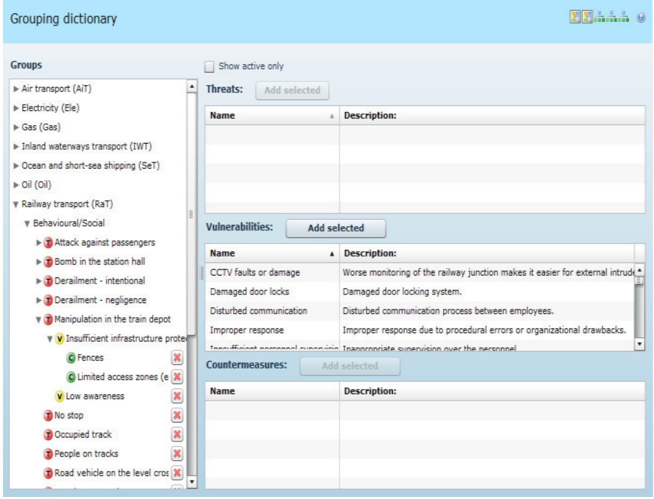
The general formula of the threats/hazards scenario is:

[*Threat agent*] exploiting [*vulnerability*] causes [*adverse action*] to [*asset*] or [*process*].

Assuming that a threat agent is identified as the hazard trigger, the common description of threats and hazards is possible in OSCAD. The threat specification includes key terms essential for the risk analysis. Threats specified during the AORA/PORA analyses are considered risk register items. OSCAD is equipped with the incident management functionality (registering, assessment, solving, lessons learnt, statistics). The incidents which have already occurred are assigned to the threat items too. For this reason, the predicted risk scenarios and occurred incidents (materialized risk scenarios) are consistent. OSCAD is able to build statistics of incidents (not discussed here).

Summing up, the risk register is defined in OSCAD as a set of risk scenarios worked out during AORA or PORA, and compatible with the incident inventory.

OSCAD has predefined lists of threats, vulnerabilities and countermeasures. They are flat, but a special grouping mechanism is applied as the hierarchical grouping dictionary. On the upper hierarchy level threats can be ordered according to critical infrastructures, next according to these threats character (Behavioural/Social, Natural/Force majeure, Organizational, Technological). For the given threat (T), relevant vulnerabilities (V) are given, and to the given pair threat-vulnerability, recommended countermeasures (C) can be assigned. These predefined relations speed up the countermeasures selection.


Grouping dictionary
Groups
Air transport (AiT)
Electricity (Ele)
Gas (Gas)
Inland waterways transport (IWT)
Ocean and short-sea shipping (SeT)
Oil (Oil)
Railway transport (RaT)
Behavioural/Social
T Attack against passengers
T Bomb in the station hall
T Derailment - intentional
T Derailment - negligence
T Manipulation in the train depot
V Insufficient infrastructure prote
C Fences
C Limited access zones (e
V Low awareness
T No stop
T Occupied track
T People on tracks
T Road vehicle on the level cros
Add selected
Add selected
Add selected

☐ Show active only

Threats:

| Name | Description: |
|---|---|

Vulnerabilities:

| Name | Description: |
|---|---|
| CCTV faults or damage | Worse monitoring of the railway junction makes it easier for external intrude |
| Damaged door locks | Damaged door locking system. |
| Disturbed communication | Disturbed communication process between employees. |
| Improper response | Improper response due to procedural errors or organizational drawbacks. |
| Insufficient personnel supervisio | Inappropriate supervision over the personnel |

Countermeasures:

| Name | Description: |
|---|---|

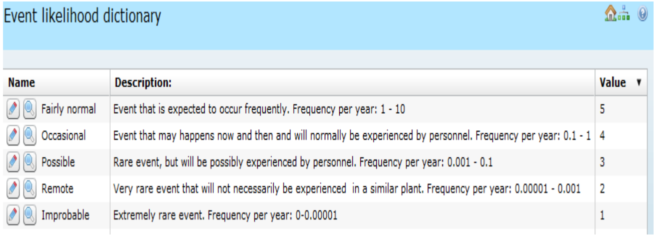
Fig. 5 Grouping dictionary with rail transport relevant data

Figure 5 presents the hierarchical structure of the grouping dictionary and some examples concerning railway transport.

### C. *Risk measures and assessment process in the OSCAD software*

For the AORA and PORA analyses two issues should be defined: likelihood of the event (Fig. 6) and its consequences (Fig. 7).

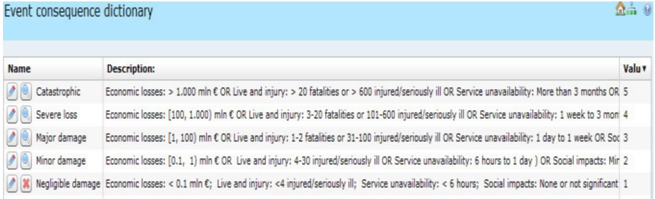
Event likelihood measures with their interpretation are direct implementation of the measures presented in Table 2 included in [4].



| Name | Description: | Value ▾ |
|---|---|---|
| Fairly normal | Event that is expected to occur frequently. Frequency per year: 1 - 10 | 5 |
| Occasional | Event that may happens now and then and will normally be experienced by personnel. Frequency per year: 0.1 - 1 | 4 |
| Possible | Rare event, but will be possibly experienced by personnel. Frequency per year: 0.001 - 0.1 | 3 |
| Remote | Very rare event that will not necessarily be experienced in a similar plant. Frequency per year: 0.00001 - 0.001 | 2 |
| Improbable | Extremely rare event. Frequency per year: 0-0.00001 | 1 |

Fig. 6 Event likelihood measures

The consequences measures are implemented in the same way . They are based on Table 1 from [4].



| Name | Description: | Valu ▾ |
|---|---|---|
| Catastrophic | Economic losses: > 1.000 mln € OR Live and injury: > 20 fatalities or > 600 injured/seriously ill OR Service unavailability: More than 3 months OR | 5 |
| Severe loss | Economic losses: [100, 1.000) mln € OR Live and injury: 3-20 fatalities or 101-600 injured/seriously ill OR Service unavailability: 1 week to 3 mon | 4 |
| Major damage | Economic losses: [1, 100) mln € OR Live and injury: 1-2 fatalities or 31-100 injured/seriously ill OR Service unavailability: 1 day to 1 week OR Soc | 3 |
| Minor damage | Economic losses: [0.1, 1) mln € OR Live and injury: 4-30 injured/seriously ill OR Service unavailability: 6 hours to 1 day ) OR Social impacts: Mir | 2 |
| Negligible damage | Economic losses: < 0.1 mln €; Live and injury: <4 injured/seriously ill; Service unavailability: < 6 hours; Social impacts: None or not significant | 1 |

Fig. 7 Event consequences measures

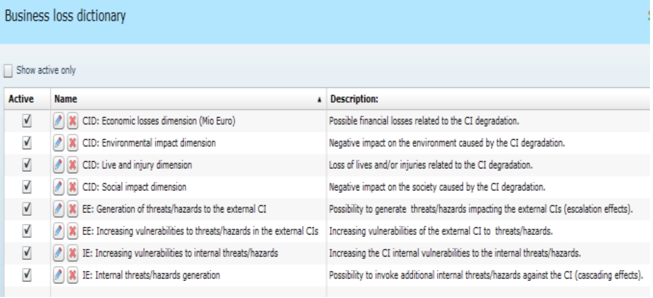
The risk value (AORA/PORA) is calculated with the use of the simple formula:

$$\text{Risk} = \frac{\text{Event likelihood} * \text{Event consequences}}{\text{Countermeasure class} * \text{Countermeasure impl.lev.}}$$

The "Countermeasure class", if used, i.e. when it is > 1, can express countermeasure assurance (basic, advanced). The "Countermeasure implementation level" expresses the stage of the implementation (not implemented, partially, fully implemented). These two additional parameters are used for more advanced applications.

The measures of multidimensional consequences of the hazardous event (Fig. 8) are key issues for the ABIA/PBIA analyses. Three groups of consequences are distinguished. The basic one is the CID (CI degradation) category which expresses different kinds of damages within the given CI. To consider the CI specific effects analyses, two additional categories are introduced:

- IE (Internal escalations), expressing new internally generated threats or new or increased vulnerabilities which influence the considered CI, caused by the hazardous event,
- EE (External escalations), expressing generated threats which impact the external CIs or new or increased vulnerabilities in the external CIs, caused by the hazardous event.



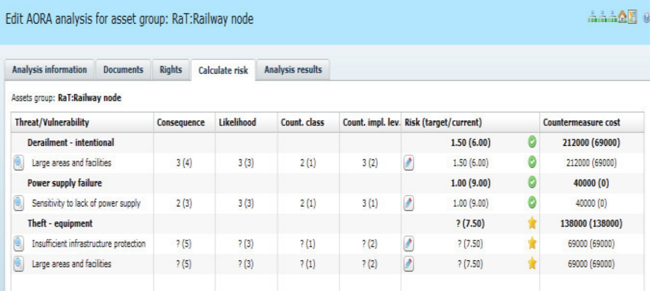
□ Show active only

| Active | Name ▴ | Description: |
|---|---|---|
| ☑ | CID: Economic losses dimension (Mio Euro) | Possible financial losses related to the CI degradation. |
| ☑ | CID: Environmental impact dimension | Negative impact on the environment caused by the CI degradation. |
| ☑ | CID: Live and injury dimension | Loss of lives and/or injuries related to the CI degradation. |
| ☑ | CID: Social impact dimension | Negative impact on the society caused by the CI degradation. |
| ☑ | EE: Generation of threats/hazards to the external CI | Possibility to generate threats/hazards impacting the external CIs (escalation effects). |
| ☑ | EE: Increasing vulnerabilities to threats/hazards in the external CIs | Increasing vulnerabilities of the external CI to threats/hazards. |
| ☑ | IE: Increasing vulnerabilities to internal threats/hazards | Increasing the CI internal vulnerabilities to the internal threats/hazards. |
| ☑ | IE: Internal threats/hazards generation | Possibility to invoke additional internal threats/hazards against the CI (cascading effects). |

Fig. 8 Event impacts measures with CID, IE and EE categories

The implementation of the bow-tie model is presented by the pair AORA-ABIA with respect to the given asset (here: railway node of the RaT infrastructure). The process approach (PORA-PBIA), though possible, is not discussed here.

AORA is shown in Fig. 9.



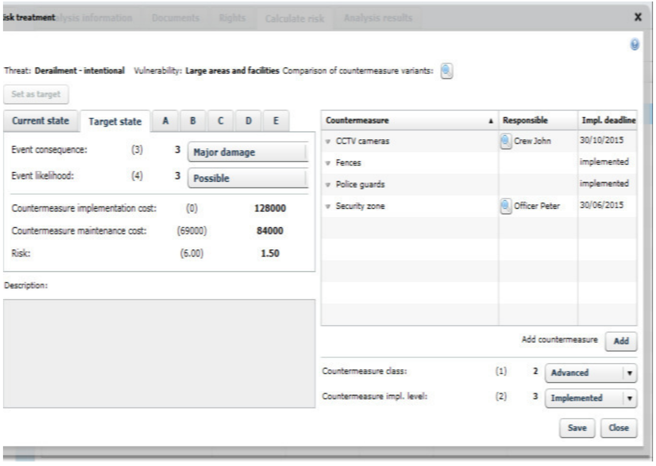
Assets group: **RaT:Railway node**

| Threat/Vulnerability | Consequence | Likelihood | Count. class | Count. impl. lev. | Risk (target/current) | | Countermeasure cost |
|---|---|---|---|---|---|---|---|
| **Derailment - intentional** | | | | | **1.50 (6.00)** | ✓ | **212000 (69000)** |
| Large areas and facilities | 3 (4) | 3 (3) | 2 (1) | 3 (2) | 1.50 (6.00) | ✓ | 212000 (69000) |
| **Power supply failure** | | | | | **1.00 (9.00)** | ✓ | **40000 (0)** |
| Sensitivity to lack of power supply | 2 (3) | 3 (3) | 2 (1) | 3 (1) | 1.00 (9.00) | ✓ | 40000 (0) |
| **Theft - equipment** | | | | | **? (7.50)** | ★ | **138000 (138000)** |
| Insufficient infrastructure protection | ? (5) | ? (3) | ? (1) | ? (2) | ? (7.50) | ★ | 69000 (69000) |
| Large areas and facilities | ? (5) | ? (3) | ? (1) | ? (2) | ? (7.50) | ★ | 69000 (69000) |

Fig. 9 Example of the AORA analysis for a railway node

Please note three threats (Derailment – intentional, Power supply failure, Theft – equipment) and vulnerabilities associated with them. For each pair threat-vulnerability, which influences the asset, the risk value can be determined according to the above presented formula. Inherent risk ("risk before") is in parentheses, while the current risk ("after measures applications") – without parentheses. The same rule applies to the cost of countermeasures. Each pair threat-vulnerability is considered a risk register item.

If the risk value is unaccepted, extra (other) countermeasures can be selected (Fig. 10).



Threat: **Derailment - intentional** Vulnerability: **Large areas and facilities** Comparison of countermeasure variants:

| | | | |
|---|---|---|---|
| Event consequence: | (3) | 3 | Major damage |
| Event likelihood: | (4) | 3 | Possible |
| Countermeasure implementation cost: | (0) | 128000 | |
| Countermeasure maintenance cost: | (69000) | 84000 | |
| Risk: | (6.00) | 1.50 | |

Description:

| Countermeasure ▴ | Responsible | Impl. deadline |
|---|---|---|
| ▾ CCTV cameras | Crew John | 30/10/2015 |
| ▾ Fences | | implemented |
| ▾ Police guards | | implemented |
| ▾ Security zone | Officer Peter | 30/06/2015 |

Add countermeasure

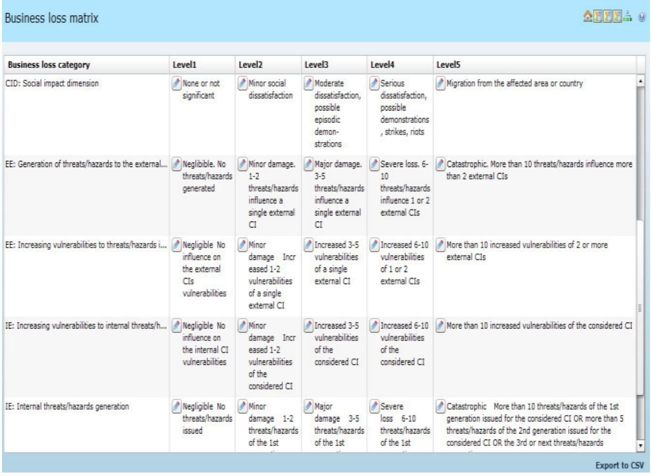
| | | | |
|---|---|---|---|
| Countermeasure class: | (1) | 2 | Advanced ▾ |
| Countermeasure impl. level: | (2) | 3 | Implemented ▾ |

Fig. 10 CI risk management – countermeasures selection

The risk manager can consider up to five variants of decisions with respect to the possible risk reduction and the cost of this reduction.

The next step is the consequences analysis of a given hazardous event embraced by ABIA. It is possible to apply ABIA in the case of each hazardous event, but in most cases it will be more convenient to perform ABIA according to assets.

The basic ABIA tool is the loss matrix (Fig. 11). For each subcategory of CID, IR, EE, losses are assessed with the use of 5 levels. A number of subcategories and levels are configurable. As a result, the CI degradation is assessed.



Business loss matrix

| Business loss category | Level1 | Level2 | Level3 | Level4 | Level5 |
|---|---|---|---|---|---|
| CID: Social impact dimension | None or not significant | Minor social dissatisfaction | Moderate dissatisfaction, possible episodic demon- strations | Serious dissatisfaction, possible demonstrations , strikes, riots | Migration from the affected area or country |
| EE: Generation of threats/hazards to the external... | Neglibible. No threats/hazards generated | Minor damage. 1-2 threats/hazards influence a single external CI | Major damage. 3-5 threats/hazards influence a single external CI | Severe loss. 6-10 threats/hazards influence 1 or 2 external CIs | Catastrophic. More than 10 threats/hazards influence more than 2 external CIs |
| EE: Increasing vulnerabilities to threats/hazards i... | Negligible No influence on the external CIs vulnerabilities | Minor damage Incr eased 1-2 vulnerabilities of a single external CI | Increased 3-5 vulnerabilities of a single external CI | Increased 6-10 vulnerabilities of 1 or 2 external CIs | More than 10 increased vulnerabilities of 2 or more external CIs |
| IE: Increasing vulnerabilities to internal threats/h... | Negligible No influence on the internal CI vulnerabilities | Minor damage Incr eased 1-2 vulnerabilities of the considered CI | Increased 3-5 vulnerabilities of the considered CI | Increased 6-10 vulnerabilities of the considered CI | More than 10 increased vulnerabilities of the considered CI |
| IE: Internal threats/hazards generation | Negligible No threats/hazards issued | Minor damage 1-2 threats/hazards of the 1st | Major damage 3-5 threats/hazards of the 1st | Severe loss 6-10 threats/hazards of the 1st | Catastrophic More than 10 threats/hazards of the 1st generation issued for the considered CI OR more than 5 threats/hazards of the 2nd generation issued for the considered CI OR the 3rd or next threats/hazards |

Fig. 11 Use of the loss matrix during BIA

Additionally, we can identify new threats (or vulnerabilities) caused by a hazardous event. These breaches usually concern assets which are also countermeasures (C=A category). Here AORA-ABIA must be performed with respect to the given asset to identify internal escalation or cascading consequences. Similarly, threats/vulnerabilities which influenced external CIs are identified. This requires extra AORA-ABIA for external CIs with respect to the influenced asset.

Fig. 12 presents an example of a risk assessment scenario driven by CIs phenomena:

1. In CI#n an external event occurs and triggers the hazardous event HE(#n,#m) impacting the primary asset #m which belongs to this infrastructure.
2. AORA(#m) identifies the risk related to this hazardous event, while ABIA(#m) – its multidimensional consequences. The internal degradation caused by the HE(#n,#m) is assessed (CID). ABIA identifies that this event breaches the security zone (#m->#k) which is a secondary asset of #m (IE) and influences the external infrastructure #p (EE).
3. Due to the internal escalation (IE) an extra analysis of the secondary asset (#m->#k), playing the role of a countermeasure, is required: AORA(#m->#k)-ABIA(#m->#k). The related ABIA identifies CI degradation caused by the security zone breaching but does not identify any further IE or EE impacts.
4. Due to the external escalation (EE) an extra analysis of asset #s of the CI#p is performed: AORA(#s)-ABIA(#s). The related ABIA identifies the CI degradation caused by an externally generated threat but does not identify any internal impacts (IE). Moreover, the backward external impacts to the infrastructure #n are identified.
5. Due to the external threat generated by the CI #p for the CI#n on its primary asset #z, an extra pair of analyses is issued: AORA(#z)-ABIA(#z). The CI internal degradation is assessed, and no internal/ external escalations are detected.

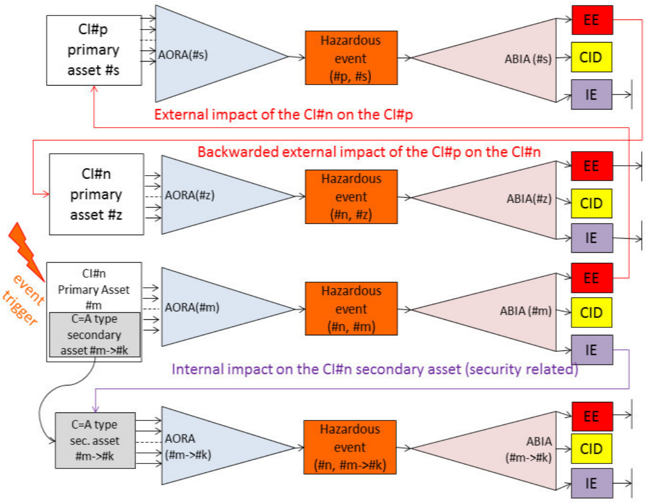


Fig. 12 Risk management in interdependent critical infrastructures

This scenario shows a general plan of analyses driven by the situations occurring in the interdependent CIs.

### D. Interdependencies and critical infrastructure specific phenomena

There is no specific tool in OSCAD to analyze interdependencies, especially the strength of coupling inside CIs. This task should be solved outside the system, e.g. by preparing a map of interdependent CIs. Using this map it is possible to further analyze risk within a set of interdependent infrastructures, which was shown in Section V, subsection C.

There is a mechanism introduced that allows to explicitly distinguish CI internal and external causes of hazardous events, and to distinguish CI internal non-escalating consequences, consequences generating hazards/threats in the same infrastructure, and consequences generating external hazards/threats for other collaborating infrastructures.

## VI. Conclusions

The short feasibility study provided in the paper confirms a possibility to adapt the ready-made OSCAD platform for CI risk management according to the previously [4] identified requirements.

The CI related data were prepared and implemented in the system. Some OSCAD functions and system messages were changed to better express the CI domain. No software changes were needed. Most of the required functionalities of the CI risk manager were successfully implemented. This way the OSCAD-Ciras tool was prepared for further researches.

The key advantage of the presented method, which allows to consider effects implied by interdependencies in risk management, is to distinguish the direct CI degradation (CID) and the internal (IE) and external (EE) escalation/cascading effects.

As far as more complicated CIs relationships are concerned, more iterations of analyses are needed. Here it is required to introduce identifiers of particular analyses, additional managing and reporting. The Ciras-OSCAD tool is currently used to elaborate use cases in the CIRAS project and to design the Ciras Toolset. The OSCAD tool can be integrated into the toolset but should be supported in the range of analyses and interdependencies management. More experiments based on the elaborated use cases are planned.

## Acknowledgment

The author thanks the colleagues from the CIRAS project consortium for reviewing this paper and discussing the presented concept.

## References

[1] Council Directive 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection (2008).

[2] Commission Staff Working Document on a new approach to the European Programme for Critical Infrastructure Protection Making European Critical Infrastructures more secure. European Commission. Brussels, Aug 28 2013, SWD(2013) 318 final

[3] Ciras project: http://Cirasproject.eu/content/project-topic (access date: June 2015).

[4] Bialas A.: Critical infrastructures risk manager – the basic requirements elaboration, In: Zamojski W., Mazurkiewicz J., Sugier J., Walkowiak T., Kacprzyk J (Eds.): *Theory and Engineering of Complex Systems and Dependability Proceedings of the Tenth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX, June 29 – July 3 2015, Brunów, Poland,* Advances in Intelligent Systems and Computing Vol. 365, 2015, Springer-Verlag: Cham, Heidelberg, New York, Dordrecht, London, pp. 11-24, DOI: 10.1007978-3-319-19216-1_2.

[5] OSCAD project: http://www.oscad.eu/index.php/en/ (access date: June 2015).

[6] Rinaldi, S.M., Peerenboom, J.P., Kelly, T.K.: Identifying, Understanding and Analyzing Critical Infrastructure Interdependencies. *IEEE Control Systems Magazine*. December, 11–25 (2001).

[7] Hokstad, P., Utne, I.B., Vatn, J. (Eds): *Risk and Interdependencies in Critical Infrastructures: A Guideline for Analysis* (Springer Series in Reliability Engineering). Springer-Verlag London (2012), DOI: 10.1007/978-1-4471-4661-2_2.

[8] Rausand, M., *Risk Assessment: Theory, Methods, and Applications*. Series: Statistics in Practice (Book 86), Wiley (2011).

[9] Giannopoulos, G., Filippini, R., Schimmer, M.: *Risk assessment methodologies for Critical Infrastructure Protection. Part I: A state of the art*. European Union (2012).

[10] Deliverable D2.1: Common areas of Risk Assessment Methodologies. Euracom (2007).

[11] ISO/IEC 31010:2009 – Risk Management – Risk Assessment Techniques.

[12] ENISA: http://rm-inv.enisa.europa.eu/methods (access date: June 2015).

[13] D1.1 State of the Art of Methods and Tools, Ciras report (Dissem. level: RE/CO), 2015.

[14] EN 61025 Fault tree analysis (FTA) (IEC 61025:2006), CENELEC (2007).

[15] EN 62502 Event tree analysis (ETA) (IEC 62502:2010), CENELEC (2010).

[16] Białas A.: Risk assessment aspects in mastering the value function of security measures. In: Zamojski W., Mazurkiewicz J., Sugier J., Walkowiak T., Kacprzyk J (Eds.): *New results in dependability and computer systems*. Advances in Intelligent and Soft Computing, Vol. 224, 2013, Springer-Verlag: Cham, Heidelberg, New York, Dordrecht, London, pp. 25-39. http://link.springer.com/chapter/10.1007%2F978-3-319-00945-2_3#page-1 DOI: 10.1007/978-3-319-00945-2_3.

[17] Bialas A.: Computer support for the railway safety management system – first validation results. In: Zamojski W., Mazurkiewicz J., Sugier J., Walkowiak T., Kacprzyk J. (Eds.): *Proceedings of Ninth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX. June 30 – July 4, 2014, Brunow, Poland*, Advances in Intelligent Systems and Computing Vol. 286, Springer Cham, Heidelberg, New York, Dordrecht, London, 2014, ISBN 978-3-319-07012-4, DOI 10.1007/978-3-319-07013-1, pp. 81-92.

[18] Białas A.: Business continuity management, information security and assets management in mining, *Mechanizacja i Automatyzacja Górnictwa*, Nr 8(510), 2013, Instytut Technik Innowacyjnych EMAG, Katowice, English version: pp. 125-138.

DOI: 10.15439/2015F88
ACSIS, Vol. 5

# A Random Traffic Padding to Limit Packet Size Covert Channels

Anna Epishkina, Konstantin Kogos
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Cybernetics anf Information Security Department,
31 Kashirskoe shosse, 115409, Moscow, Russia
Email: {avepishkina, kgkogos}@mephi.ru

***Abstract*—This paper observes different methods for network covert channels constructing and describes the scheme of the packet length covert channel. The countermeasure based on random traffic padding generating is proposed. The capacity of the investigated covert channel is estimated and the relation between parameter of covert channel and counteraction tool is examined. Practical recommendation for using the obtained results are given.**

## I. Introduction

THE IDEA of covert channel was introduced by Lampson in 1973. The covert channel is a communication channel that was not intended for information transfer at all, such as the service program's effect on the system load [1]. It is obvious that covert channel may lead to information leakage and it cannot be eliminated by techniques to detect network anomalies [2], malware activities [3], etc. TCSEC postulates that the covert channel is a communication channel which allows the transfer of data and violation of security policy [4].

Presently, the most popular covert channels are built on packet switching data networks because of some features available in the TCP/IP protocol suite [5]. Moreover, traditional security measures based on traffic encryption also permit the design of different types of covert channels.

Covert channels are divided by the data transfer technique into two classes such as, timing and storage channels [4]. Storage channel allows the direct or indirect storage recording by one process and the direct or indirect reading of it by another. Timing channel allows one process to signal information to another process by modulating of system resources (e. g. CPU time) usage so that the change in response time observed by the second process would provide information.

The first technique to design a storage channel in the IP network is to modulate packet header fields, e. g. TTL [6], IP ID [7], ToS [8]. The second technique is based on the modification of the packet length. Different timing channels in the IP networks use alteration of the inter-packet delays [9], [10], e. g. by JitterBug [11] and packet transfer rate [12]. In addition, the packet reordering could be used to build a timing channel [13]. Timing channel is a channel with noise since a packet timing is a random variable whose distribution depends on the network load [14].

A capacity of undetectable packet size covert channels can be higher than a capacity of timing channels, therefore these channels can lead to a serious security breach. The authors of the paper research this type of covert channels and propose the technique to estimate and limit their capacity which can be useful in different types of information systems, e.g. in data storage system in a cloud [15].

**Our Contribution.** Since a technique to choose the quantitative characteristics of countermeasures in order to keep balance between capacity of covert and communication channels has not yet been proposed, we offer an approach to gain it. This paper describes a technique to estimate and limit the capacity of a packet size covert channel based on random traffic padding generation. The investigation carried out is significant because such type of covert channels could be constructed even if data encryption is used. There are complicated undetectable covert channels which have no noise in contrast to timing channels.

This paper is organized as follows. It gives an analysis of different types of packet size covert channels. The investigated covert channel scheme and the counteracting technique are shown. Then the capacity of the covert channel is estimated and the technique to generate dummy packets is given. The main results and further research guidelines are summed up in the conclusion.

## II. Related Work

For the first time Padlipsky [16] and Girling [17] suggested to modulate the length of data link layer frames in order to accomplish the hidden data transfer. The main idea of the technique is as follows: sender and receiver share the rule used to compose a byte of the covert message depending on the frame length. To describe any byte of the transmitted message, one should use 256 different frame lengths.

Yao constructed a covert channel in which a sender and a receiver shared the periodically updated matrix with elements representing unique unsorted packet lengths in 2008 [18]. The sender using the bits of hidden transmitted message determines the matrix row and randomly chooses a packet length from the row. The receiver finds the gained packet length in the matrix and reconstructs bits of the message according to the row number. Because packet length distribution given by covert channel is not equal to packet length distribution of normal traffic, this type of covert channel is detectable.

978-83-60810-66-8/$25.00©2015, IEEE

Ji suggested a protocol-independent covert channel in 2009 [19]. Before the transmission starts, sender and receiver form the dynamically updated reference of packet lengths by fixing packets lengths in normal traffic. In order to transfer a hidden message the sender transmits a special packet. The length of the special packet is chosen from the reference using the algorithm known to the sender and receiver. The length of the next packet is a sum of lengths of the previous packet and the number corresponding to the message bits. The reference is updated by adding the length of the transmitted packet. The receiver recovers the message bits by evaluating the difference of the packets length gained. The disadvantage of the covert channel is as follows: the lengths of hidden messages are added to the reference, therefore the packet length distribution with the covert channel is not equal to packet length distribution of the normal traffic and this type of covert channels is detectable in case of the large data volume.

Ji worked out another protocol-independent covert channel in 2009 [20]. Before the transmission starts, sender and receiver form a non-updating reference of packet lengths by fixing packets lengths in the normal traffic. The main advantage of the technique is a small space and time complexity of the decoding, since the sender stores the whole reference and receiver saves only the lower and upper bound of each basket. To transmit the hidden data the sender randomly chooses the packet length from the basket, the receiver determines the number of the basket and restores the message bits. The regularity in the distribution of the transmitted message bits could cause a highly probable detection of the channel.

Hussain improved the technique and designed high capacity covert channel based on the alteration of packets lengths and information content in 2011 [20]. Sender and receiver share periodically updated matrix with elements representing unsorted packet lengths in normal traffic. The sender using bits of hidden message determines the matrix row and randomly chooses the packet length from the row. If the chosen length belongs to the stego-column, then the data is transferred in the information content of the packet, otherwise the data is transferred in the number of matrix row. The receiver finds the length of gained message in the matrix, detects the transfer method using the matrix row and recovers bit of the message. The disadvantage of the channel is that information content of the packet has to be used as the hidden container and is more complicated in comparison with the other techniques.

Edekar improved the method in 2013 [22] and realized it using TCP. Packets lengths in the shared matrix are unique. Each matrix element $a$ is associated with the binary vector $(v, y)$, where $v = 1 \Leftrightarrow a$ belongs to the stego-column and $y = 1 \Leftrightarrow a$ belongs to the stego-row. Then if $(v, y) = (1, 1)$ the packet is ignored; if $(v, y) = (1, 0)$ data is transferred in the packet information content; if $(v, y) = (0, 0)$ data is transmitted in the number of the matrix row; if $(v, y) = (0, 1)$ data is transferred in the number of the matrix row and the packet information content.

The way to eliminate covert channels based on length of transferred packets modulation is to equalize packets lengths and send packets with maximum length. However, the technique essentially diminishes the capacity of a communication channel. To limit a covert channel capacity, the random increase of packet lengths and generation of dummy packets can be used. Kiraly suggested the realization of the methods based on IPsec in 2008 in order to make traffic nontraceable [23].

However, a technique to choose the quantitative characteristic of the methods in order to keep balance between capacity of covert and communication channels has not yet been proposed. The authors of this paper offer an approach to gain it.

## III. The Covert Channel Scheme

Let the lengths of transferred packets possess the values from $l_{fix}$ to $l_{fix} + L$. The disjoint sets $L_0$ and $L_1$ are given and

$$\begin{cases} L_0 \cup L_1 = N_{L+l_{fix}} \backslash N_{l_{fix}-1}, \\ \mid L_0 \mid = \mid L_1 \mid \end{cases} \tag{1}$$

where $N_a$ stands for the set of positive integers from 1 to $a$.

Further, we consider a method to build a binary covert channel. In order to transfer «0» the sender communicates a packet with length $l \in L_0$, to transfer «1» the sender communicates a packet with length $l \in L_1$. It is obvious that the capacity of such a channel without counteraction is equal to 1 bit per packet. A large-scale site loses about 26 Gb of data annually if there is a covert channel with such a capacity [24].

If the symbol distribution in a transmitted message simulates a uniformly distributed random sequence (e.g. cryptographic keys sending), a random equiprobable choice of packets lengths from $L_0$ and $L_1$ leads to equally probable random distributed lengths of transmitted packets. Moreover, $L_0$ and $L_1$ could be periodically changed multisets so that a random choice of packets lengths from $L_0$ and $L_1$ induces the distribution of packets lengths to be close to the empirically obtained distribution of normal traffic.

To build such a covert channel, the sender must have one of the following possibilities:

- to modify lengths of transmitted messages;
- to form packets with undefined lengths;
- to buffer packets to be sent and transfer them to a channel at a specified moment.

The investigation proposes a technique to limit the capacity of covert channel based on random traffic padding. After $i$ data packets have been sent, random length dummy packets are created, the number of packets $i$ between dummy packets is the value of random variable that is uniformly distributed at the $N_k$, $k \in N$ where $k$ is the parameter of a counteraction tool. Let $\mu$ be the capacity of a communication channel, then a counteraction tool decreases the capacity of a communication channel and it equals

$$\frac{k+1}{k+3}\mu. \tag{2}$$

The countermeasure using traffic padding generation after equal number of packets passed is not resistant against the attack of traffic padding tracing. If the violator detects traffic padding ones, data transmitting via covert channel has no errors and can be processed without synchronization. The investigated countermeasure is resistant to this type of attack.

After a dummy packet is received, the mismatch between the hidden sender and hidden receiver takes place. To negotiate this fact SOF packets [25] are utilized after transferring $T-1$ packets within a covert channel. A receiver fixes $T-1$ packets gained after SOF packet and waits for the next SOF packet. Thus, $T$ is the parameter of a covert channel which estimates the synchronization frequency.

## IV. The Capacity of the Covert Channel

In 1987 Millen was the first to suggest the use of information theory to estimate a capacity of covert channels with noise [26]. The investigation was continued by Ventakraman [27]. The authors determine network covert channels and analyze techniques to audit and limit a capacity of covert channels utilizing indirect routing.

The capacity $C$ of the investigated covert channel is

$$C = \max_X I(X,Y) \tag{3}$$

where $I(X,Y)$ is the mutual information of random variables describing the input and output properties of the covert channel properly, the dimensionality of covert channel capacity is one bit per packet.

Let us consider the case when synchronization is done more rarely than a dummy packet sending, i.e. $T > k$. After a dummy packet is received, mismatch between sender and receiver occurs, therefore identification of the following received bits would be wrong until the next synchronization happens. Consequently, in order to build a covert channel the inequality $T < k+1$ is required.

Let the synchronization be not less frequent than dummy packet sending, i.e. $T < k+1$. Corresponding choice of parameters is explained in Fig. 1.

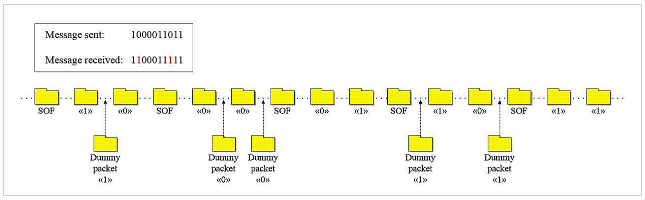


Fig. 1. The scheme of data transfer in the covert channel ($T = 3$, $k = 5$)

Since each $T$-th packet sent via a covert channel is not a data packet but a synchronization packet, the mutual information can be calculated using the following formula

$$I(X,Y) = \frac{T-1}{T} I^*(X,Y) \tag{4}$$

where $I^*(X,Y)$ is a mutual information of random variables describing the input and output properties of a covert channel without synchronization accordingly.

The mutual information $I^*(X,Y)$ is equal to the form of

$$I^*(X,Y) = H(Y) - H(Y|X) \tag{5}$$

where the entropy of random variable $Y$ is

$$H(Y) = -\sum_{y\in\{0,1\}} p(y)\log_2 p(y), \tag{6}$$

the conditional entropy of random variable $Y$ comparatively to random variable $X$ is expressed as

$$H(Y|X) = \\ = -\sum_{x\in\{0,1\}} \Big(p(x)\Big(\sum_{y\in\{0,1\}} (p(y|x)\log_2 p(y|x))\Big)\Big). \tag{7}$$

Since sizes of sets $L_0$ and $L_1$ are equal and lengths of passing through the covert channel dummy packets are chosen randomly and equiprobable, then

$$H(Y) = 1. \tag{8}$$

Whereas the values of conditional probabilities $p(x|y)$, $x, y \in \{0,1\}$ depend on the number of packets sent via a covert channel between the moment of synchronization and the moment of dummy packet receiving, the mutual information $I^*(X,Y)$ can be found using the following formula

$$I^*(X,Y) = \frac{\binom{k-T+2}{2} + \sum\limits_{i=0}^{T-2}\Big((k-i)(1-H_i(Y|X))\Big)}{\binom{k+1}{2}} \tag{9}$$

where $H_i(Y|X)$ is the conditional entropy of random variable $Y$ compared to random variable $X$ and it is evaluated when $i$ packets received between a moment of synchronization and the dummy packet's arrival.

The mutual information $I(X,Y)$ could be estimated as

$$I(X,Y) = \frac{(T-1)\Big(\binom{k-T+2}{2} + kS_1(T) - S_2(T)\Big)}{T\binom{k+1}{2}} \tag{10}$$

where

$$S_1(T) = \sum_{i=0}^{T-2}(1-H_i(Y|X)), \tag{11}$$

$$S_2(T) = \sum_{i=0}^{T-2} i(1-H_i(Y|X)). \tag{12}$$

Since

$$H_i(Y|X) = -\Bigg(\frac{T-1-i}{2(T-1)}\log_2\frac{T-1-i}{2(T-1)} + \\ + \frac{T-1+i}{2(T-1)}\log_2\frac{T-1+i}{2(T-1)}\Bigg) \tag{13}$$

then

$$S_1(T) = -(T-1)\log_2(T-1) + \\ + \frac{1}{2(T-1)}\Big(\sum_{i=0}^{T-2} f^+(i) + \sum_{i=0}^{T-2} f^-(i)\Big), \tag{14}$$

$$S_2(T) = \frac{1}{2(T-1)}\Big(\sum_{i=0}^{T-2} f^{2,+}(i) - \sum_{i=0}^{T-2} f^{2,-}(i) + \\ +(T-1)\sum_{i=0}^{T-2} f^{+}(i) - (T-1)\sum_{i=0}^{T-2} f^{-}(i)\Big), \tag{15}$$

where

$$f^{+}(i) = (T+i-1)\log_2(T+i-1), \tag{16}$$

$$f^{-}(i) = (T-i-1)\log_2(T-i-1), \tag{17}$$

$$f^{2,+}(i) = (T+i-1)^2\log_2(T+i-1), \tag{18}$$

$$f^{2,-}(i) = (T-i-1)^2\log_2(T-i-1). \tag{19}$$

In order to analyze functions $f^{+}(i)$, $f^{-}(i)$, $f^{2,+}(i)$, $f^{2,-}(i)$ we will examine analogue variable $\tilde{i}$, $\tilde{i} \in [0, T-2]$ instead of discrete variable $i$, in which case $f^{+}(\tilde{i})$, $f^{2,+}(\tilde{i})$ are strictly increasing and $f^{-}(\tilde{i})$, $f^{2,-}(\tilde{i})$ are strictly decreasing defined and continuous functions in the interval $[0, T-2]$. Then the values of the following forms

$$\sum_{i=0}^{T-2} f^{+}(i), \sum_{i=0}^{T-2} f^{-}(i), \sum_{i=0}^{T-2} f^{2,+}(i), \sum_{i=0}^{T-2} f^{2,-}(i) \tag{20}$$

could be approximated by means of functions $f^{+}(\tilde{i})$, $f^{-}(\tilde{i})$, $f^{2,+}(\tilde{i})$, $f^{2,-}(\tilde{i})$ integrating in the interval $[0, T-2]$ accordingly.

Now we explain how to gain the approximate value of the sum $\sum\limits_{i=0}^{T-2} f^{2,+}(i)$ and the other sum can be estimated in a similar,

$$\sum_{i=0}^{T-2} f^{2,+}(i) \approx \int\limits_{0}^{T-2} f^{2,+}(\tilde{i})d\tilde{i} + f^{2,+}(T-2) - \\ -\sum_{j=0}^{T-3}\frac{f^{2,+}(j+1) - f^{2,+}(j)}{2} = \int\limits_{0}^{T-2} f^{2,+}(\tilde{i})d\tilde{i} + \\ +\frac{f^{2,+}(0) - f^{2,+}(T-2)}{2}. \tag{21}$$

Values of the integrals could be found using the variable substitution and integration by parts

$$\int\limits_{0}^{T-2}(T+\tilde{i}-1)^2\log_2(T+\tilde{i}-1)d\tilde{i} = \\ =\mid T+\tilde{i}-1 = p \mid = \\ = \int\limits_{T-1}^{2T-3} p^2\log_2 p dp = \frac{1}{3}\int\limits_{T-1}^{2T-3}\log_2 p d(p^3) = \\ = \frac{1}{3}(p^3\log_2 p)\Big|_{T-1}^{2T-3} - \frac{1}{3}\int\limits_{T-1}^{2T-3} p^3 d(\log_2 p) = \\ = \Big(\frac{1}{3}p^3\log_2 p - \frac{p^3}{9\ln 2}\Big)\Big|_{T-1}^{2T-3}. \tag{22}$$

It follows that

$$I(X,Y) \approx \frac{(T-1)\Big(\binom{k-T+2}{2} + kA(T) + B(T)\Big)}{T\binom{k+1}{2}} \tag{23}$$

where

$$A(T) = \frac{2T-3}{2}\log_2\frac{2T-3}{T-1} - \frac{T-2}{2\ln 2}, \tag{24}$$

$$B(T) = -\binom{T-1}{2}\log_2(T-1) - \\ -\frac{(T-1)^2\log_2(T-1)}{6} + \frac{(T-2)^2}{12\ln 2} - \\ -\frac{(2T-3)(2T^2-6T+3)\log_2(2T-3)}{12(T-1)}. \tag{25}$$

Graphs of function $I(X,Y)$ from $k$ where $T \in \{2,3,4\}$ are illustrated in Fig. 2.

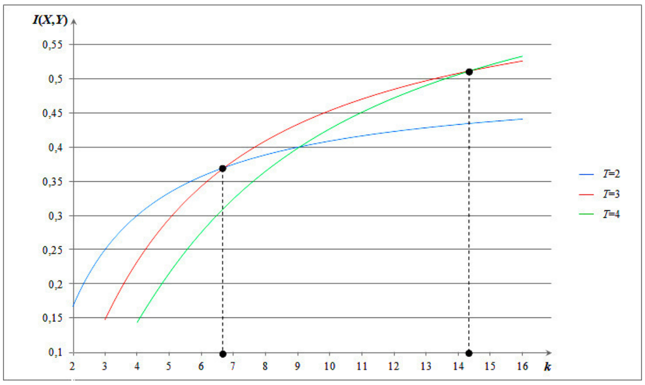


Fig. 2. $I(X,Y)$ function as a function from $k$ graph, $T \in \{2,3,4\}$

To build a covert channel the parameter $T$ is chosen while $I(X,Y)$ has a maximum value. Fig. 2 shows that when $k \in \{2,...,6\}$, parameter $T$ should be equal to 2, when $k \in \{7,...,14\}$, parameter $T$ should be equal to 3.

## V. The Practical Using of the Obtained Results

Let functioning of a covert channel with capacity less than $v_{\max}$ has no influence upon security (the dimensionality of $v_{\max}$ is one bit per sec) and the capacity of communication channel is $\mu$ bits per sec. Hence if the average length of transmitted packets is $\bar{l}$ bits, then the residual communication channel capacity is

$$\frac{(k+1)\mu}{(k+3)\bar{l}} \tag{26}$$

packets per sec. In case of the full using of the communication channel by the sender and the receiver the capacity of the covert channel is

$$v = \frac{(k+1)\mu C}{(k+3)\bar{l}} \tag{27}$$

bits per sec, where $C$ is the capacity of covert channel with the dimensionality bits per packet. Therefore when the allowable capacity of covert channel in limited, the following inequality is true

$$v \leq v_{\max}. \tag{28}$$

For example, according to [4] the functioning of a covert channel with capacity less than $v_{\max} = 100$ bits per sec can be acceptable in some cases. Let the capacity of the communication channel be $\mu = 100$ Mbits per sec (standard 100Base-T). The maximum length of IP packet is 65535 bytes, from which 20 bytes is a header length ($l_{fix} = 20$ bytes and $L = 65515$ bytes). Then the average length of sending packets is $\bar{l} = 262220$ bits in case of random equipropable choice of packets lengths from equinumerous sets $L_0$ and $L_1$.

Table I presents the relation between parameters of the covert channel and the counteraction tool (the symbol «*» means that the value of $v$ is round up to 2 digits).

TABLE I
RELATION BETWEEN PARAMETER OF COVERT CHANNEL, PARAMETER OF COUNTERACTION TOOL AND CAPACITY OF COVERT CHANNEL

| *k* | *T* | *C* | *v* |
|---|---|---|---|
| 2 | 2 | 0,17* | 41 |
| 3 | 2 | 0,25 | 67 |
| 4 | 2 | 0,30 | 86 |
| 5 | 2 | 0,33* | 100 |
| 6 | 2 | 0,36* | 112 |
| 7 | 3 | 0,38* | 122 |

Table I shows that in order to limit the capacity of covert channel up to 100 bits per sec the parameter of counteraction tool should be $k = 5$.

## VI. Conclusion

In this work the capacity of a packet size covert channel was examined using the information theory statements. The counteraction tool based on random traffic padding generating was designed. We proposed selecting the parameter of the counteraction tool when an allowable covert channel capacity is given.

The topic of the further work is to estimate the residual capacity in multi-symbol covert channels with traffic padding and to investigate the technique to limit covert channel capacity by random increase of packets sizes.

## References

[1] Lampson, B.W. 1973. A Note on the Confinement Problem. Communications of the ACM, 16(10):613–615, http://dx.doi.org/10.1145/362375.362389

[2] Szmit, M., Szmit, A., Kuzia, M. 2013. Usage of RBF Networks in prediction of network traffic. Annals of Computer Science and Information Systems. Position Papers of the 2013 Federated Conference on Computer Science and Information Systems, 1:63-66.

[3] Jasiul, B., Sliwa, J., Gleba, K., Szpyrka, M. 2014. Identification of malware activities with rules. Annals of Computer Science and Information Systems. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, 2:101-110, http://dx.doi.org/10.15439/978-83-60810-58-3

[4] Department of defense trusted computer system evaluation criteria. Department of defense standard, 1985.

[5] Zander, S., Armitage, G., Branch, P. 2007. A survey of covert channels and countermeasures in computer network protocols. IEEE Communications surveys and tutorials, 9(3):44–57, http://dx.doi.org/10.1109/COMST.2007.4317620

[6] Zander, S., Armitage, G., Branch, P. 2006. Covert channels in the IP time to live field. Proceedings of the 2006 Australian telecommunication networks and applications conference, 298–302.

[7] Ahsan, K., Kundur, D. 2002. Practical data hiding in TCP/IP. Proceedings of the 2002 ACM Multimedia and security workshop.

[8] Handel, T., Sandford, M. 1996. Hiding data in the OSI network model. Proceedings of the first International workshop on information hiding, 23–38, http://dx.doi.org/10.1007/3-540-61996-8_29

[9] Berk, V., Giani, A., Cybenko, G. 2005. Detection of covert channel encoding in network packet delays: Technical report TR2005-536. New Hampshire: Thayer school of engineering of Dartmouth College.

[10] Sellke, S.H., Wang, C.-C., Bagchi, S., Shroff, N.B. 2009. Covert TCP/IP timing channels: theory to implementation. Proceedings of the twenty-eighth Conference on computer communications, 2204–2212.

[11] Shah, G., Molina, A., Blaze, M. 2009. Keyboards and covert channels. Proceedings of the 15th USENIX Security symposium, 59–75.

[12] Yao, L., Zi, X., Pan, L., Li, J. 2009. A study of on/off timing channel based on packet delay distribution. Computers and security, 28(8):785–794, http://dx.doi.org/10.1016/j.cose.2009.05.006

[13] Kundur, D., Ahsan, K. 2003. Practical Internet steganography: data hiding in IP. Proceedings of the 2003 Texas workshop on security of information systems.

[14] Bovy, C.J., Mertodimedjo, H.T., Hooghiemstra, G., Uijterwaal, H., Mieghem, P. van. 2002. Analysis of end-to-end delay measurements in Internet, Proceedings of the 2002 ACM Conference Passive and Active Measurements.

[15] Shatilov, K., Boiko, V., Krendelev, S., Anisutina, D., Sumaneev, A. 2014. Solution for Secure Private Data Storage in a Cloud. Annals of Computer Science and Information Systems. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, 2:885-889, http://dx.doi.org/10.15439/978-83-60810-58-3

[16] Padlipsky, M.A., Snow, D.W., Karger, P.A. 1978. Limitations of end-to-end encryption in secure computer networks: Technical report ESD-TR-78-158. Massachusetts: The MITRE Corporation.

[17] Girling, C.G. 1987. Covert channels in LAN's. IEEE Transactions on software engineering, 13(2):292–296.

[18] Yao, Q., Zhang, P. 2008. Coverting channel based on packet length. Computer engineering, 34(3):183–185.

[19] Ji, L., Jiang, W., Dai, B., Niu, X. 2009. A novel covert channel based on length of messages. Proceedings of the 2009 Symposium on information engineering and electronic commerce, 551–554, http://dx.doi.org/10.1109/IEEC.2009.122

[20] Ji, L., Liang, H., Song, Y., Niu, X. 2009. A normal-traffic network covert channel. Proceedings of the 2009 International conference on computational intelligence and security, 499–503, http://dx.doi.org/10.1109/CIS.2009.156

[21] Hussain, Mehdi, Hussain, M. 2011. A high bandwidth covert channel in network protocol. Proceedings of the 2011 International conference on information and communication technologies, 1–6, http://dx.doi.org/10.1109/ICICT.2011.5983562

[22] Edekar, S., Goudar, R. 2013. Capacity boost with data security in network protocol covert channel. Computer engineering and intelligent systems, 4(5):55–59.

[23] Kiraly, C., Teofili, S., Bianchi, G., Cigno, R. Lo, Nardelli, M., Delzeri, E. 2008. Traffic flow confidentiality in IPsec: protocol and implementation. The International federation for information processing, 262:311–324, http://dx.doi.org/10.1007/978-0-387-79026-8_22

[24] Fisk, G., Fisk, M., Papadopoulos, C., Neil, J. 2002. Eliminating steganography in Internet traffic with active wardens. Proceedings of the fifth International workshop on information hiding, 18–35, http://dx.doi.org/10.1007/3-540-36415-3_2

[25] Cabuk, S., Brodley, C.E., Shields, C. 2004. IP covert timing channels: design and detection. Proceedings of the eleventh ACM conference on computer and communications security, 178–187, http://dx.doi.org/10.1145/1030083.1030108

[26] Millen, J.K. 1987. Covert channel capacity. Proceedings of the IEEE Symposium on Security and Privacy, 60–66, http://dx.doi.org/10.1109/SP.1987.10013

[27] Venkatraman, B.R., Newman-Wolfe, R.E. 1995. Capacity estimation and auditability of network covert channels. Proceedings of the IEEE Symposium on Security and Privacy, 186–198, http://dx.doi.org/10.1109/SECPRI.1995.398932

DOI: 10.15439/2015F127
ACSIS, Vol. 5

# Software Risk Assessment for Measuring Instruments in Legal Metrology

Marko Esche, Florian Thiel
Physikalisch-Technische Bundesanstalt
Abbestr. 2-12
10587 Berlin, Germany
Email: {marko.esche, florian.thiel}@ptb.de

***Abstract*—In Europe, measuring instruments subject to legal control are responsible for an annual turnover of 500 billion Euros and need to pass a conformity assessment with respect to European directives or national legislation before they can be used. Today, measuring instruments are frequently integrated into open networks and even branch into the areas of cloud computing and Internet of Things. Since software is one of the key components of such devices, Germany's national metrology institute, the Physikalisch-Technische Bundesantalt, is developing a method to assess the risks and evaluate current threats associated with software. The method uses the structure of and combines elements from the international ISO/IEC standards 27005 and 15408. It could be helpful for conformity assessment bodies and industry alike and supports the comparability of risk assessment results. Despite its focus on legal metrology, the method is applicable to other areas where software risk assessment is required, too.**

## I. Introduction

CERTAIN types of measuring instruments, like gas meters, taximeters, fuel pumps and grain moisture meters are subject to legal control in the European Union. Before making them available on the market, such measuring instruments have to undergo a conformity assessment according to the Measurement Instruments Directive (MID) 2014/32/EU [1]. The entire area of measuring instruments even including individual measurements regulated by either national or European legislation is referred to as legal metrology. According to estimations, about four to six percent of the gross national income in European countries is accounted for by legal metrology. In Germany alone, 130 million of such instruments are installed. These are responsible for economic transactions worth roughly 157 billion Euros each year. For a more detailed description of the role of legal metrology in general see [2].

In most cases, the conformity assessment is performed by so-called Notified Bodies, which have proven that they have at their disposal "(a) personnel with technical knowledge and sufficient and appropriate experience to perform the conformity assessment tasks, (b) descriptions of procedures in accordance with which conformity assessment is carried out, ensuring the transparency and the ability of reproduction of those procedures" [1]. One such Notified Body that performs conformity assessments is the Physikalisch-Technische Bundesanstalt (PTB), Germany's national metrology institute. The assessment itself is conducted according to a combination of modules (A to H1) which encompass different roles for manufacturers and Notified Bodies [1]. For most of these modules, a new general requirement has been introduced in 2014 concerning the submitted manufacturer's documentation. It states, "The documentation shall make it possible to assess the instrument's conformity to the relevant requirements, and shall include an adequate analysis and assessment of the risk(s)." Such a risk assessment does not only need to cover the physical measuring instrument itself but also the metrologically relevant software running on it. In this context, harmonization between European Notified Bodies obviously becomes necessary to ensure fair and comparable software risk assessment within the common trade zone. In this paper, an approach for software risk assessment is presented that

- makes use of established international standards as far as possible and
- identifies risks with reproducible numeric values to better ensure comparability between evaluation results.

The remainder of the paper is structured as follows: In Section II, a literature overview covering other methods in the field of software risk assessment is provided. In order to obtain reproducible analysis results, a clear definition of assets and threats to these assets is required. A derivation of such assets from the requirements of the MID is, therefore, provided in Section III. An algorithmic description of the risk assessment approach proposed here, may be found in Section IV. Afterwards, the new approach is compared with other existing methods based on two real-world examples in Section V. Section VI summarizes the paper and provides an overview of planned future work.

## II. Overview of existing methods

Before giving a list of reference approaches to software risk assessment, it is necessary to clearly identify what kind of risk assessment is required in the context of the MID. The directive establishes a common baseline by listing a number of essential requirements which all measuring instruments have to fulfill. Since the most important target of the MID is to ensure free and fair trade as well as to protect the consumer, these essential requirements are mainly targeted at protecting measuring results from accidental and intentional manipulation and to make both correct measuring results and detected manipulations traceable. Further details may be found in Section III. In this context, the term *risk* can be seen as the

978-83-60810-66-8/$25.00©2015, IEEE

product of the probability that the essential requirements are no longer met and the legal impact resulting from such a breach of the MID. It is important to note, that no financial loss needs to be associated with the risk, instead, the sole basis for the analysis are the essential requirements.

### A. *ISO/IEC 27005*

Probably most important to mention is the ISO/IEC 27000 family of standards which covers all aspects of an information security management system (ISMS). According to ISO/IEC 27005 [3], ”risk is a combination of the consequences that would follow from the occurrence of an unwanted event and the likelihood of the occurrence of the event.“ Thus, three different components are needed to calculate risk, namely

- a list of unwanted events,
- consequences resultig from such events,
- and the likelihood of occurence of individual events.

In order to derive all three components [3] gives details on generalized procedures to conduct risk assessment. The standard places risk assessment in a logical chain comprising context establishment, risk assessment, and risk treatment, where risk assessment consists of risk identification, risk estimation, and risk evaluation.

During the risk identification phase, assets are to be identified first. These are derived from a ”list of constituents with owners, location, function“, resulting in a list of assets to be managed. Afterwards, for each possible asset, threats are collected based on information from reviewed incidents, accounts from asset owners, and possibly external threat catalogs. These threats correspond to the ”unwanted events“ mentioned above. The next step consists of identifying existing risk control mechanisms which could, for instance, be determined from the provided documentation. Risk identification is completed by an identification of vulnerabilities which can be used to implement certain threats. In this context, a vulnerability can only cause harm if it can be used to realize a threat. Equivalently, threats without a corresponding exploitable vulnerability do not pose a risk.

The next part of ISO/IEC 27005, concerned with risk estimation, is likely the most relevant in the context of this paper. The standard considers both qualitative and quantitative approaches to calculate risk probability, where the quantitative approach ”uses a scale with numerical values for both consequences and likelihood, using data from a variety of sources.“ Such numerical values are a prerequisite for ensuring comparability among risk assessment results for different products conducted by different examiners. To derive at actual probabilities, ISO/IEC 27005 first assigns certain impacts or consequences to incidents that could result from realized threats by means of exploited vulnerabilities. Possible examples of impacts include loss of confidentiality of certain assets as well as a breach of asset integrity. In a final step, the probability, with which a threat is realized, is estimated. Important factors, in this context, are the frequency at which certain threats occur in real life and the difficulty of exploiting a vulnerability. For intentional exploitation of threats, ISO/IEC 27005 suggests a valuation of motivation and capabilities, resources available to the attackers as well as the perception of individual vulnerabilities. This approach will later be revisited in Section IV where certain aspects of ISO/IEC 27005 are reflected in the risk assessment approach presented here. Nevertheless, the standard does not prescribe a reference model to calculate individual numeric threat probabilities. The choice of such a model is instead left up to the user of the standard. One possible method to calculate risks quantitatively may, for instance, be found in [4], where the author proposes to define risks as probability functions that describe the likely gains or losses obtained from security incidents. The final components of risk assessment according to ISO/IEC 27005 are evaluation of the risk level and risk evaluation. The aim of these steps is to prioritize the identified risks according to the predetermined evaluation criteria.

### B. *WELMEC Guide 5.3*

In order to harmonize the work of Notified Bodies in Europe, a number of non-mandatory guides have been established within the European Legal Metrology Cooperation (WELMEC). The guide 5.3 ”Risk Assessment Guide for Market Surveillance: Weigh and Measuring Instruments“ deals with risk assessment from a market surveillance perspective and originates from Regulation 765/2008/EC [5]. Its main goal ”is to understand the impact the instrument will have on the end user/consumer“ [6]. The guide establishes a list of evaluation criteria which should help ”the market surveillance authority to define priorities and to determine the choice of strategies to achieve their goals.“ Since the guide is solely targeted at market surveillance authorities, the expected impact is not clearly defined but rather encompasses everything from ”economic implications, public health, consumer confidence [to] legal issues“ [6]. Even if only legal issues are considered, the spectrum of the guide is still too broad to objectively evaluate software in measuring instruments. Instead, the guide provides a clear rule to eventually calculate the risk associated with non-compliance but does not provide means for calculating individual threat probabilities.

### C. *Van Deursen et al. ”Source-Based Software Risk Assessment“*

One approach, that does not have this shortcoming, is the one by van Deursen et al. in [7]. There, risk assessment is defined as ”an independent assessment of the risks involved in building, operating or maintaining a software system.“ The method then calculates the risk based on so-called primary and secondary facts, where primary facts are data acquired through automatic source code analysis and secondary facts are obtained using user questionnaires. The primary facts are mainly needed to identify subsystems that show features not usually found in software systems. After the primary facts have been used to validate the secondary data, a final result can be computed. This method could readily be adapted for the use in legal metrology. However, source code is usually not a required part of the documentation for MID conformity assessment.

### *D. Foo et al. "Software Risk Assessment Model"*

Another method to objectively evaluate and compare risks associated with software was presented by Foo et al. in [8]. There the basic abstraction technique used by the authors is a software risk assessment model (SRAM) which is constructed based on an extensive questionnaire to be answered by the risk assessor of a software product. Within [8] the term risk is defined as "factors that may cause late delivery, cost overrun or low quality of a software product." Nevertheless, the authors list the productive level of the staff, flexibility of the delivery schedule and most importantly complexity of software as having a significant impact on the risk evaluation. In the context of a MID conformity assessment, the first two sources of information are of no importance since the MID is not concerned with business processes. The complexity of the software can again not be used due to lack of available information. Moreover, the risk assessor required by the SRAM approach needs to have access to resources such as source code and error statistics that are not available to a MID evaluator. For comparison, a description of a risk assessment approach for measuring instruments with a similar scope covering the entire life cycle of a device can be found in [9].

### *E. Sadiq et al. "Software Risk Assessment and Evaluation Process (SRAEP) using Model Based Approach"*

In [10] a different software risk assessment method was proposed that is also model-driven. Sadiq et al. therein describe the Software risk assessment and evaluation process (SRAEP) which is based on the software risk assessment and evaluation model (SRAEM). Their approach is targeted at highlighting threats to the success of a software project rather than threats to a finished software product. Nevertheless, a number of useful formalized steps are included in their method which shall be reused later. For this reason, the basic steps of the SRAEP will be revisited here.

According to the authors the motivation for using a model-based assessment strategy is two-fold:

- With the help of a model, precise descriptions of the target system, its context, and security features can be formulated. These are prerequisites for performing risk assessments.
- The modeling technology facilitates a more precise documentation of risk assessment results and of the assumptions on which their validity depends. This is expected to reduce maintenance costs by increasing the possibilities of reuse of the documentation.

The SRAEP itself can be divided into two steps: the identification of a context for the analysis and the identification of risks themselves. Sadiq et al. here split the context identification into an identification of areas of concern, a description and evaluation of assets, and, finally, an identification of security requirements. These three steps will be used again during asset derivation, see Section III, and in the risk assessment method that is proposed here as described in Section and IV. Before beginning with the risk analysis, the SRAEP requires an evaluator to acquire detailed knowledge of the analysis target. Based on this knowledge, all security issues related to software should be discussed making reference to common vulnerabilities or the results of tool-based vulnerability checks.

### *F. ISO/IEC 15408 (Common Criteria)*

An international standard for software security that explicitly does not address risk assessment is ISO/IEC 15408 also known as the "Common Criteria" (CC) [11]. In the CC, a set of functional security requirements is defined, which can be used to describe both product requirements in the form of Protection Profiles and product specifications in the form of Security Targets. An implemented Security Target, i.e. a product to be tested, is referred to as a Target of Evaluation (TOE). The standard also provides a list of assurance components, a chosen subset of which is also included in said Protection Profiles and Security Targets. These assurance components are then used to validate the design, the development, and finally the completed IT product itself. In which manner the assurance components are to be checked is not described in the CC themselves but rather in the Common Evaluation Methodology (CEM) [12] which accompanies the CC. Two building blocks from the CC with details provided in the CEM are of special interest here: Firstly, each Security Target includes a Security Problem Definition in which assets to be protected are identified. The CC initially list primary assets that represent objects or information of a given value whose authenticity, integrity or availability are to be protected. Certain aspects of an IT product can also become assets themselves when they play an integral role in realizing security functionality. These are referred to as secondary assets. Both types of assets are examined and listed in the security problem definition. Afterwards, possible threat agents and adverse actions that could be executed on the assets are investigated and described in a semi-formal manner. The combination of threat agent, asset, and adverse action is referred to as a threat. This construct will here again be used since it facilitates the implementation of reproducible risk assessment results.

Secondly, one part of validating a Security Target consists of a so-called vulnerability analysis which is specified in the CC's AVA_VAN class. The assurance components associated with this class allow an examiner to execute both white box and black box tests on the Security Target based on the knowledge acquired during the evaluation procedure. The vulnerability analysis uses a point score, where each adverse action to be executed on an asset is evaluated with respect to five different aspects ranging from the time required to the equipment needed to implement an attack, for details see Section IV. In each category mentioned a point score is determined. Based on the total sum of all points the TOE's resilience to an attack is checked. This is done with the aid of matrix, details on which will also be provided in Section IV. More information concerning the general mechanisms of the vulnerability analysis will be given there as well.

### G. ETSI TS 102 165-1

The European Telecommunications Standards Institute (ETSI) in an international non-profit organization that publishes industrial standards in the area of telecommunications systems. These are targeted at manufacturers of communication equipment and network operators. One of these standards is ETSI TS 102 165-1 ”Telecommunications and Internet converged Services and Protocols for Advanced Networking“ [13] (TISPAN), where in part 1 ”Method and proforma for Threat, Risk, Vulnerability Analysis“ are detailed. The so-called technical specification describes a risk assessment approach consisting of nine individual steps that comprise everything from a definition of the device to be examined (TOE) up to the establishment of risks and an identification of countermeasures. The method starts by defining clearly the boundaries of the TOE and by identifying its security functionalities using terminology from the common criteria. Afterwards, all assets are identified. In [13] these have to fall into one of the following categories: equipment, human assets or information stored. It will be shown in Section IV that this definition is to narrow for most applications outside the area of communication systems. Moreover, the standard does not describe a way to derive abstract assets resulting, for instance, from legal requirements. Next, possible ”attack interfaces“ are identified that a threat agent can use to implement a threat. In order to assess the likelihood of occurrence for an individual threat, elements from the CC's AVA_VAN class are used as described earlier. Details on the method may be found in Section IV. According to [13], ”threats to a telecommunications system are fairly restricted and fall into a small set of easily identified operations.“ Consequentially, [13] only lists a very small number of possible threats namely interception, manipulation, repudiation, and denial of service. While well suited to the area of telecommunication networks, these are to limited for general-purpose IT devices. The same is true for the definition of threat agents where [13] only allows a very small number of different roles. Finally, impact in the context of TISPAN is defined as a function of the intensity of an attack. For measuring instruments, as discussed here, a different definition is needed which will be given in Section III. Nevertheless, the method has certain properties which are of use to the scenario discussed here:

- calculation of the probability of an attack based on the AVA_VAN class from the CC,
- evaluation of impact and attack likelihood based on simple numeric scores (1 to 3 points),
- extension of the AVA_VAN class to account for multiple attacks being executed simultaneously.

## III. Formal derivation of security requirements from the Directive 2014/32/EU

Before beginning with the algorithmic description of the new risk assessment method for software itself, a specific set of assets and associated security properties for measuring instruments will be derived here. These will later be reused in the experimental evaluation. The derivation should be seen as an example on how to formalize legal or contractual requirements with respect to software. In application scenarios not related to the conformity assessment of measuring instruments, other assets such as human health or monetary values etc. would, of course, be used. The latest version of the MID lists several requirements relating to software as plain text, which will be formalized here.

### A. Exemplary Asset Derivation

The actual procedure of defining security requirements based on legal specifications will be highlighted with an example: Annex I of the MID lists so-called essential requirements for measuring instruments that have to be fulfilled before putting them on the European market. As an example clause 8.4 will be used here. It reads, ”Measurement data, software that is critical for measurement characteristics and metrologically important parameters stored or transmitted shall be adequately protected against accidental or intentional corruption.“ [1, L 96/173]

The requirement specifically mentions three asset candidates, namely measurement data, software that is critical for measurement characteristics, and metrologically important parameters stored or transmitted. All three assets are required to be protected against accidental or intentional corruption. Firstly, this can be interpreted as a requirement for guaranteeing integrity of these assets. Secondly, however, an intentional replacement of a parameter set also represents a viable way to invalidate parameter integrity. Thus, authenticity of said assets also appears to be required. This is not specifically mentioned in the MID but is common understanding among Notified Bodies [14]. Consequentially, the assets measurement data, software critical for measurement characteristics, and metrological parameters are associated with the security properties of integrity and authenticity. Availability of the software, however, is not mandatory since an instrument with no running measurement software cannot produce false measuring results.

Another requirement related to software can be found in Annex I of the MID, clause 7.6. It states, ”When a measuring instrument has associated software which provides other functions besides the measuring function, the software that is critical for the metrological characteristics shall be identifiable and shall not be inadmissibly influenced by the associated software.“[1, L 96/173] Again, two assets are specifically mentioned. The first is the identification of the software. The second one is an inadmissible influence by other software which is not a physical object or an IT object itself but rather is a property of the software. In the language of the CC, prohibiting external influence on the software can be expressed by stating, that the inadmissible influence shall be unavailable. This again enables the use of a fixed scheme to describe security functionality by identifying dedicated security properties associated with an asset.

### *B. Complete List of Assets*

An overview of all MID requirements for software in measuring instruments may be found in Table I. The two examples discussed here are listed there for completeness as well.

Apart from the measurement software itself, the assets to be protected include an identifier for the software, measurement, results and parameters that determine the behavior of the instrument. In addition, the presentation of the measurement result as well as the presentation of the identifier for the software have to meet special requirements. Details on how these requirements are usually fulfilled may be found in [14] where the paragraphs from the MID are translated into implementation specific requirements and into so-called acceptable (technical) solutions. Even though [14] is usually of great value for both software developers and software examiners, it will not be used here since the new risk assessment procedure (see Section IV) aims to be generic and independent from a limited number of established technical realizations.

## IV. Algorithmic Description of the Risk Assessment Procedure

The risk assessment method described in this paper follows the structure defined in [3] and consists of three main parts, namely identification of assets, identification of attack vectors, and calculating the probability of occurrence for an individual attack. Each of these will now in turn be described. A flowchart that links all three parts and incorporates details for each step may be found in Figure 1.

### *A. Identification of Assets*

As has been detailed in Section III, assets to be protected can be derived directly from the legal requirements for measuring instruments as laid down in the MID. This is in accordance with [3], which states that the risk evaluation process can take ”legal and regulatory requirements, and contractual obligations“ into account and should also consider the ”criticality of the information assets involved“. In addition to the asset definition, one or more attacker models are needed, see upper right corner of Figure 1. In the simplest case, a Notified Body will assume all market players to be untrustworthy with equal motivation to manipulate measurement results and measuring instruments. This includes manufacturers or distributors of such devices, users or maintainers, and customers. These only differ in their respective capabilities to implement an attack. Subsequently, the risk assessment procedure can use the market player with the most detailed knowledge and with the highest skills (normally manufacturer or user) as a representative model. The basic structure for both the attacker model and the formulation of adverse actions may be found in the Security Problem Definition as described in [11]. When examining an individual measuring instrument, it may make sense to individually differentiate between attackers/authenticated users with different levels of access. If the authentication data is available to any of the market operators mentioned, then the highest access rights granted to any of these will be allocated to the modeled attacker. The only entity to be considered trustworthy in this context is the market surveillance which may hold administrator level authentication data for certain devices.

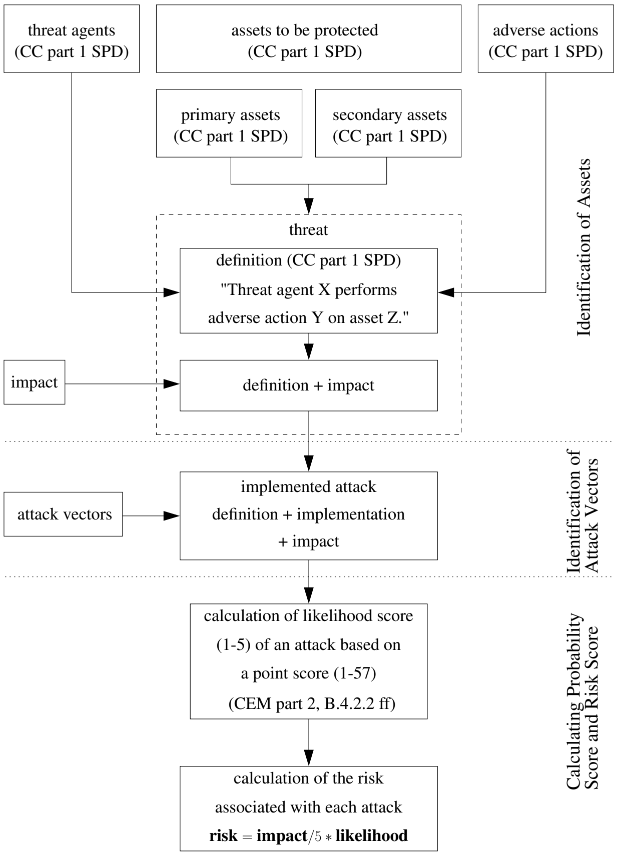


Fig. 1. Flowchart of the proposed risk assessment procedure. The notes on the right hand side indicate the division into the three main steps of the method.

Yet another action to be performed during the asset identification phase is the collection of certain adverse actions that can cause harm to the assets. Here again, a generic approach from the CC [11] can be used: Each modeled attacker may harm any of the identified assets by invalidating one or more of their security properties, i.e. availability, integrity or authenticity as applicable. Such an adverse action may then read for example, ”An attacker with the access rights of a local administrator manages to invalidate the availability of the proof of an intervention.“ The formulation of a complete threat may be found within the dashed box in Figure 1. The complete set of possible adverse actions derived from these basic combinatorics then only has to be checked for consistency and for possible duplicates. In a final step, the implemented attack, consisting of an adverse action and an attack vector, will be assigned an individual impact score between 1 and 5. Since all legal requirements are generally assumed to be equally important, the highest score (5) will usually be used. A smaller score will only be chosen if the attack only applies to a single measurement or can later be

TABLE I
REQUIREMENTS FROM THE MID [1] RELATING TO SOFTWARE AND THEIR FORMALIZATION AS ASSETS AND SECURITY PROPERTIES. THE NUMBERS IN BRACKETS AFTER EACH ASSET (A1 TO A10) REPRESENT THERE UNIQUE IDENTIFIER.

| **Requirement in the MID [1] Annex I** | **Asset** | **Security Property** |
|---|---|---|
| 7.6 ”When a measuring instrument has associated software which provides other functions besides the measuring function, the software that is critical for the metrological characteristics shall be identifiable and shall not be inadmissibly influenced by the associated software.“ | identification of the software (A9) | availability, integrity |
| | inadmissible influence on the software (A5) | unavailability |
| 8.1 ”The metrological characteristics of a measuring instrument shall not be influenced in any inadmissible way by the connection to it of another device, by any feature of the connected device itself or by any remote device that communicates with the measuring instrument.“ | inadmissible influence on the software (A5) | unavailability |
| 8.3 ”Software identification shall be easily provided by the measuring instrument.“ | presentation of the software identification (A10) | availability |
| 8.3 ”Evidence of an intervention shall be available for a reasonable period of time.“ | evidence of an intervention (A2) | availability, integrity |
| 8.4 ”Measurement data, software that is critical for measurement characteristics and metrologically important parameters stored or transmitted shall be adequately protected against accidental or intentional corruption.“ | measurement data (A3) | integrity, authenticity |
| | software critical for metrological characteristics (A1) | integrity, authenticity |
| | metrologically important parameters (A4) | integrity, authenticity |
| 10.1 ”Indication of the result shall be by means of a display or hard copy.“ | indication of the result (A6) | availability, integrity |
| 10.2 ”The indication of any result shall be clear and unambiguous and accompanied by such marks and inscriptions necessary to inform the user of the significance of the result.“ | marks and inscriptions (A7) accompanying the indication of a result | availability, integrity |
| 11.1 ”A measuring instrument other than a utility measuring instrument shall record by a durable means the measurement result accompanied by information to identify the particular transaction, when: the measurement is non-repeatable; and the measuring instrument is normally intended for use in the absence of one of the trading parties.“ | record of a measurement result (A8) | availability, integrity, authenticity |

detected by market surveillance.

### B. Identification of Attack Vectors

The second stage of the risk assessment phase is certainly the least formalized one. It begins with a careful study of the submitted manufacturer's documentation of the measuring instrument to be examined. This process is shown in the middle section of Figure 1. The evaluator then collects possible attack vectors consisting of actions to be performed, that would enable an attacker to realize any of the previously identified threats. This represents a clear difference to the TISPAN method detailed in Section II. Some of these attack vectors may be as simple as trying a number of password combinations on a keypad in order to gain a higher level of access. Others may comprise complex cross-site-scripting (XSS) attacks in conjunction with the preparation of a root kit to take over a device in the field and subsequently install unapproved software on the device. One relatively simple attack vector from this category is the execution of a denial-of-service (DoS) attack on a measuring instrument connected to the Internet. In many cases, such an attack will lead to the generation of an arbitrary number of error messages written to an audit log which is subject to legal control. Should the log be restricted in size, an earlier intervention may no longer be traceable later if the log is flooded with a huge amount of automatically generated errors. This would be a direct breach of the essential requirements as laid down in Section III.

### C. Calculating Probability Score and Risk Score

Once an adverse action with one or more associated attack vectors has been identified, it remains to calculate the likelihood with which the attack will actually be implemented, see lower part of Figure 1. A similar activity is in detail described in the vulnerability analysis (class AVA_VAN) in [12]. There, an evaluator estimates the resistance of an IT product (TOE in the language of the CC) to certain attacks. The evaluation in [12] is done based on five different scores describing the resources needed for the attack:

- Elapsed Time (0-19 points)
- Expertise (0-8 points)
- Knowledge of the TOE (0-11 points)
- Window of Opportunity (0-10 points)
- Equipment (0-9 points)

The score for elapsed time represents the amount of time required by the selected attacker to implement the chosen attack. A score of 0 usually signifies work of less than a day. Required work of less than a week would give a point score of 1, whereas a score of 19 represents an estimated work period of more than half a year. Further examples will be given in Section V. A table with details on all five score criteria and additional explanations for the choice of the criteria may be found in [12, p. 429].

The logarithmic progression of the scores ensures that with every additional point assigned to an attack, it becomes significantly more complex to implement. This also means that the score is more easily reproducible since evaluators

TABLE II
CALCULATING THE RESISTANCE OF A TOE. THE THIRD COLUMN MAPPING THE TOE RESISTANCE LEVEL TO THE APPROPRIATE PROBABILITY SCORE IS NOT PART OF THE ORIGINAL TABLE AS GIVEN IN [12].

| Sum of Points | TOE Resistance | Probability Score |
|---|---|---|
| 0-9 | No rating | 5 |
| 10-13 | Basic | 4 |
| 14-19 | Enhanced Basic | 3 |
| 20-24 | Moderate | 2 |
| >24 | High | 1 |

will certainly come up with estimated attack times of the same magnitude even if the actual times differ slightly. In the second category (expertise), between 0 and 8 points can be assigned, where 0 represents layman capabilities and 8 points are given when the attacker needs to be an expert in more than one field. The third category refers to the required knowledge concerning the attacked device. Again, a score of 0 is given if only publicly available knowledge is needed, such as information easily available from the web. 3 points represent restricted knowledge as might be found in the user documentation. The maximum of 11 points would stand for critical inside information only available to employees of the manufacturer. One very important score criterion is the window of opportunity available to the respective attacker. In the case of unlimited access, as would be usual for devices connected to the Internet, 0 points will be given, where 1 point signifies easy access. Should access, however, be difficult to obtain, 10 points can be assigned. In the ideal case, where access is impossible, no rating is done and the respective attack vector is removed from the list of candidates. At this point, there exists a simple way to include the motivation of threat agents into the score. Should an attacker lack the motivation to implement a threat even though he is able to realize the attack vector, the respective threat should be removed from the list.

When the assignment of score points has been done according to the five categories mentioned, the sum total of all scores is calculated. During a CC evaluation the so-called TOE resistance is then derived as indicated by Table II. A score between 10 and 13 points would, for instance, demonstrate a basic resistance to attacks, while a score above 24 indicates high resilience. In the context of the CC, the resistance to attacks would then be used to validate the selected evaluation assurance level (EAL). Here, however, the resistance rating is mapped to a probability score between 1 and 5, where 5 represents high probability of occurrence for an attack and 1 states that an attack is very unlikely to occur. The mapping of TOE resistance to probability score is also shown in Table II.

Calculating the risk associated with a threat subsequently consists of multiplying the impact score (between 1 and 5) for the given threat with the probability score of the most probable attack vector, that could realize the threat:

$$\text{risk score} = \frac{\text{impact score}}{5} \cdot \text{probability score} \qquad (1)$$

Dividing the impact score by 5 simply ensures, that the risk score is in the range between 1 and 5, too. As will be shown in the experimental evaluation, the risk score thus calculated can easily be used to rank risks associated with a single instrument or even to compare different instruments and their risks with one another.

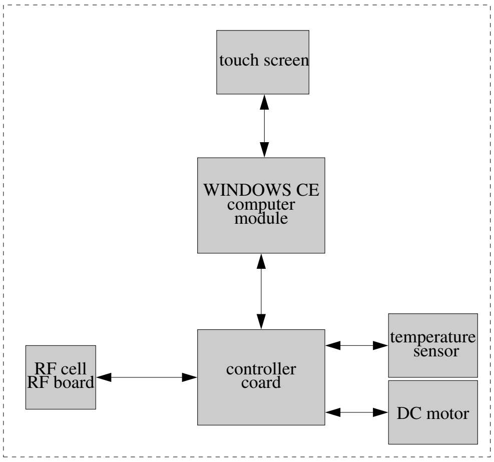


Fig. 2. High-level schematic for the grain moisture analyzer that was evaluated as an illustrative example. Parts lying physically inside the instrument are surrounded by the dotted line.

## V. EXPERIMENTAL EVALUATION AND COMPARISON WITH OTHER METHODS

### A. Grain Moisture Meter

The first measuring instrument that was examined during evaluation of the proposed software risk assessment method is a grain moisture analyzer. Such devices usually take a small sample of grain and calculate the moisture level within the sample by submitting it to infrared light and observing the absorbed wavelength spectrum. The relative moisture is economically important since it has a significant influence on the price of the grain. In this example, the measuring instrument is a stand-alone unit that is physically closed except for a touch screen, the sample inlet, as well as a serial and a USB port. As an operating system Windows CE is used. Certain types of grain can be selected via the touch screen, which is also used to start the measurement process and to show the current and past measurement results. In addition, the instrument contains a so-called audit log in which changes to both software and relevant measurement parameters are recorded. If an empty USB stick is plugged into the unit, it will write all available measurement results together with the respective date and time of the measurement to the stick. The measuring process can also be started via the serial port, which uses a proprietary protocol. Through this protocol measurement results can be read out, too. Access to the relevant system parameters and to the operating system are protected by a 6-digit password. A high-level schematic of the system may be found in Figure 2.

Based on the described system architecture and the documentation supplied by the manufacturer, a list of possible attack vectors can be compiled. The following list is just a short extract from the complete one and is used here for illustration purposes:

- **A_PASSWORD**: An attacker gains access to the administrator password by trying all 6-digit combinations.
- **A_SW_REPLACE**: An attacker retrieves the administrator password and replaces the legally relevant software.
- **A_INT_SERIAL**: An attacker exploits a vulnerability of the proprietary serial protocol and causes the instrument to malfunction.
- **A_INT_SERIAL_VALUE**: An attacker exploits a vulnerability of the proprietary serial protocol and manipulates a measurement value by interrupting the measurement.
- **A_INT_USB**: An attacker manages to install malicious code on the measuring instrument by disabling the USB-port's protection.

For each threat, as described above, an evaluator now has to go through the list of attack vectors and select those vectors that can realize the threat. In some cases, a combination of attack vectors might be necessary. An excerpt from the complete mapping between threat scenarios and attack vectors can be found in Table III. Each threat can then be rated individually. Both the point score for each aspect and its meaning are supplied in the table as well. The first threat (T1) here is a replacement of the legally relevant software by a local attacker. Since the only individual with adequate access to the measuring instrument is the operator of the device, he is also assumed to be the most likely attacker. This, of course, has an influence on the assigned point scores (see Table III). To realize T1, the attacker first has to retrieve the password for the operating system. In addition, a software needs to be written that mimics the behaviour of the approved one without raising suspicion from customers. The development of such a software is deemed to be very complex, giving it a time rating of more than half a year with an associated point score of 19. In addition, the attacker needs to be an expert in the area of software development or needs to have access to somebody who has such skills. The expertise score is therefore set to 6. Restricted knowledge of the device such as a description of the system behavior and its components is also required. Finally, the owner of the measuring instrument has unlimited access to it and to write software no special equipment aside from an off-the-shelf PC is required. The sum score for this scenario is 29, which even in the context of the CC is so high, that virtually no attack likelihood remains. Table II subsequently assigns the lowest probability score for this threat.

Subsequently, threats T2 to T5 are rated in the same manner. Threat T5 has a probability score of 2 which is identical to those of threats T3 and T4. Nevertheless, the associated risk score is only 1 since T5 will only affect one single measurement result and thus has a fairly low impact. For all other threats, there is no difference between risk and probability score since they were classified as having the highest possible impact score of 5.

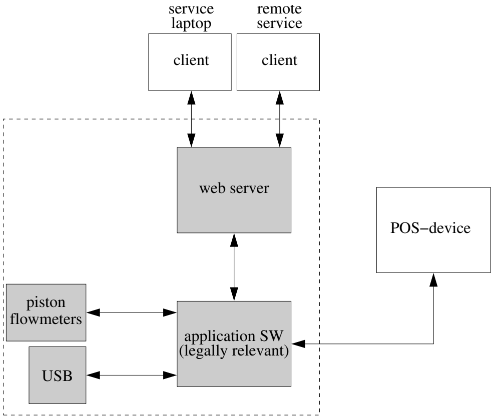


Fig. 3. High-level schematic for the fuel pump calculator that was evaluated as an illustrative example. Parts lying physically inside the instrument are surrounded by the dotted line.

### B. Fuel Pump

The second measuring instrument, that was evaluated according to the new scheme proposed here, is the calculator unit of a fuel pump. The device communicates externally with a point-of-sales (POS) device and reads data from a number of flow piston meters. As an operating system a common Linux distribution is used. Measurement results are displayed locally on a seven-segment-display and are also transmitted over a LAN to the POS device. The communication with the POS device is unidirectional. Parameters that influence the metrological behavior of the system and the operating system can be changed and accessed when a USB stick with a 32-bit key is plugged into the unit. This key is usually only in the possession of an authorized inspector. In addition, the instrument possesses a web server that can be accessed over the Internet. Through the web interface, the status of the machine can be queried and parameters can be set.

A rough schematic of the measuring instrument and its surroundings can be found in Figure 3. With reference to the documentation supplied by the manufacturer a list of possible attack scenarios can again be identified. The easiest way of deriving meaningful attack vectors is focusing on interfaces available to the outside world. Here, these include both the USB port and the communication with the POS device as well as the web interface. The communication with the POS device is physically sealed since it is also under legal control. The USB port is easily accessible for the owner of the pump, while the web interface is freely accessible for anybody in possession of the IP address. The web server in question, as a commonly used IT product, has several entries in the public CVE database [15] which is maintained by MITRE, a non-profit company operating multiple research and development centers financed by the US government. The database provides an extensive list of known vulnerabilities for virtually all software components

TABLE III
Evaluation of a small number of selected threats according to their attack vector for the example no. 1 ”grain moisture analyzer“

| **Threat** | **Description** | **Impact** | **Attack Vector** | **Elapsed Time** | **Expertise** | **Knowledge of the TOE** | **Window of Opportunity** | **Equipment** | **Sum** | **Score** | **Risk** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T1 | Local admin (S1) invalidates integrity or authenticity of the metrological software (O1). | 5 | A_SW_REPLACE | (>180d)<br>19 | (expert)<br>6 | (restricted)<br>3 | (unlimited)<br>0 | (standard)<br>0 | 28 | 1 | **1** |
| T2 | Local admin (S1) invalidates the availability of the evidence of an intervention (A2). | 5 | A_INT_SERIAL | (>30d)<br>4 | (proficient)<br>3 | (sensitive)<br>7 | (unlimited)<br>0 | (specialized)<br>4 | 18 | 3 | **3** |
| T3 | Local admin (S1) invalidates the integrity of the metrological parameters (A4). | 5 | A_INT_SERIAL_VALUE | (>60d)<br>7 | (expert)<br>6 | (sensitive)<br>7 | (unlimited)<br>0 | (specialized)<br>4 | 24 | 2 | **2** |
| T4 | Local admin (S2) invalidates the availability of the evidence of an intervention (A2) by deleting the evidence. | 5 | A_PASSWORD | (>180d)<br>19 | (layman)<br>0 | (restricted)<br>3 | (unlimited)<br>0 | (standard)<br>0 | 22 | 2 | **2** |
| T5 | Local admin (S2) invalidates integrity, authenticity or availability of a measurement result (A8). | 2 | A_INT_USB | (>60d)<br>7 | (expert)<br>6 | (restricted)<br>3 | (unlimited)<br>0 | (specialized)<br>4 | 20 | 2 | **1** |

publicly available. A short excerpt from the compiled list of possible attack vectors will be supplied here:

- **A_USB_SCRIPT**: An attacker fakes an authorized key on a USB stick thus gaining access to the operating system.
- **A_WEB_XSS**: An attacker utilizes CVE-2011-4273 for a XSS attack to execute arbitrary javascript code on the web server and to subsequently download a root kit to the system.
- **A_WEB_DOS**: An attacker exploits CVE-2009-5111, CVE-2003-1568 or CVE-2002-2429 by executing a DoS attack via partial HTTP requests.
- **A_WEB_SOCKET**: An attacker executes arbitrary malicious code while establishing a connection making use of CVE-2002-2431.

As was the case with the first example, all known threats are iteratively examined. For each of them, the evaluator has to decide whether there are any attack vectors that could be used to realize the respective threat. Afterwards, the combination of threat and attack vector is again evaluated using the point score from the CC. Here too, the threat T1 consists of a replacement of the legally relevant software after gaining access to the operating system. Since the password is entered via the USB port, there is the possibility to execute and automated brute-force attack on the authentication data. This will, however, require significant resources. In addition, a replacement software needs to be written that mimics the original one. Again, the time needed for implementing of T1 is very high, resulting in a point score of 19. Also, the attacker in question needs to be an expert software engineer (point score 6) with detailed knowledge of the measuring instrument (point score 3). Should the operator of the pump also be the attacker, unlimited access to the device is obviously given resulting in a respective score of 0. The sum total of 29 points is still so high, that the resulting probability score of 1 is negligible. This has to be seen in conjunction with the fact that the selected combination of attacker and capabilities is highly improbable in any case.

Much more likely appears a local attack on the web server (see threat T2). Since exploits for the servers vulnerabilities can be downloaded from the web, no significant implementation time (point score 4) is needed. The attacker still needs to be an expert (point score 6) in the field of software engineering to correctly use the exploit. A certain amount of detail with respect to the measuring instrument such as its IP address (point score 3) is also required. But since the attack is web-based, access is virtually unlimited. Given specialized equipment like a platform to test the attack, the resulting total point score of 17 is relatively low. This results in an medium threat probability. The evaluation of the remaining threats progresses in a similar manner. The evaluation results may be found in Table IV. Again, it is important to mention, that some threats (T5 and T6) have a low risk score despite a medium probability score since their impact is limited to one single measurement at a time.

### C. Comparison with WELMEC Guide 5.3

WELMEC Guide 5.3 uses the same definition of risk as the approach discussed here: a product of impact and probability of occurrence for a threat. Additionally, the guide proposes to calculate an average impact score based on economic

TABLE IV
EVALUATION OF A SMALL NUMBER OF SELECTED THREATS ACCORDING TO THEIR ATTACK VECTOR FOR THE EXAMPLE NO. 2 ”FUEL PUMP“

| **Threat** | **Description** | **Impact** | **Attack Vector** | **Elapsed Time** | **Expertise** | **Knowledge of the TOE** | **Window of Opportunity** | **Equipment** | **Sum** | **Score** | **Risk** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T1 | Local admin (S1) invalidates integrity or authenticity of the metrological software (A1). | 5 | A_USB_ SCRIPT | (>180d)<br>19 | (expert)<br>6 | (restricted)<br>3 | (unlimited)<br>0 | (standard)<br>0 | 28 | 1 | **1** |
| T2 | Local admin (S1) invalidates the integrity of the metrological parameters (A4). | 5 | A_WEB_ SOCKET | (>30d)<br>4 | (expert)<br>6 | (restricted)<br>3 | (unlimited)<br>0 | (specialized)<br>4 | 17 | 3 | **3** |
| T3 | Remote admin (S2) invalidates the availability of the evidence of an intervention (A2). | 5 | A_WEB_ SOCKET | (>30d)<br>4 | (expert)<br>6 | (restricted)<br>3 | (unlimited)<br>0 | (specialized)<br>4 | 17 | 3 | **3** |
| T4 | Remote admin (S2) invalidates the integrity or the authenticity of the metrological software (A1). | 5 | A_WEB_ DOS + A_WEB_ XSS | (>180d)<br>19 | (expert)<br>6 | (sensitive)<br>7 | (unlimited)<br>0 | (specialized)<br>4 | 36 | 1 | **1** |
| T5 | Local admin (S1) invalidates the availability or the integrity of the indication of the result. (A6) | 2 | A_USB SCRIPT | (>30d)<br>4 | (expert)<br>6 | (restricted)<br>3 | (easy)<br>1 | (standard)<br>0 | 14 | 3 | **1** |
| T6 | Remote admin (S2) invalidates availability or integrity of the indication of a result (A6). | 2 | A_WEB_ DOS | (>30d)<br>4 | (proficient)<br>3 | (restricted)<br>3 | (unlimited)<br>0 | (specialized)<br>4 | 20 | 3 | **1** |

implications, public health, consumer confidence, and legal issues. The assets identified in Section III, which fall into the latter category, could thus also be used in the context of the guide. However, the likelihood estimation in WELMEC Guide 5.3 clearly has the aim of assessing ”the probability of non-compliance“. It addresses both behavior of manufacturer and consumer as well as the production cycle of the instrument. The assessment is clearly focused on the manufacturer of the instrument. Unintended, implementation-based vulnerabilities of a measuring instrument are not within the scope of WELMEC Guide 5.3 since technical details with respect to the instrument's components are not taken into account at all. Instead, much more emphasis is placed on the perception of legal requirements and statistics concerning malfunctions observed in the field. While the latter may provide helpful hints, it cannot be used to assess risks associated with a new product in advance. The Guide is thus not able to produce comparable evaluation results.

### *D. Comparison with ISO/IEC 27005*

The most significant difference between the approach presented here and ISO/IEC 27005 [3] is the addition of the probability calculation based on the vulnerability analysis from the CC and the CEM. ISO/IEC 27005 explicitly states that likelihood estimation techniques should take into account ”the motivation and capabilities, which will change over time, and resources available to possible attackers“. Both motivation and capabilities (including equipment, required skills, and knowledge) can clearly be mapped to the CC-based probability estimation as detailed in Section IV. The new risk assessment approach can thus be seen as a practical realization of ISO/IEC 27005, which does itself not specify ways of calculating individual threat probabilities. The identification of assets as described in Section III is also clearly compatible with ISO/IEC 27005, since it follows the same three-step approach of risk identification, risk estimation and risk evaluation. Nevertheless, a number of additional hints can be found in the standard on how to improve the proposed method further:

- Motivation or resources of attackers may change over time, so a risk assessment for a specific measuring instrument may have to be conducted again after a certain time interval to keep it up to date.
- Even though a possible attacker may have access to a device, he might lack the motivation or the skills to carry out an attack. Thus, each implemented threat with an associated attack vector could be checked against a list of likely attackers. Unlikely combinations of skill and window of opportunity could then be removed from the evaluation table, resulting in a more clearly defined risk scenario.

## VI. SUMMARY AND FUTURE WORK

The method for software risk assessment for measuring instruments in legal metrology described here follows the guidelines of ISO/IEC 27005 [3]. In addition, elements from ISO/IEC 15408 [11] and ISO/IEC 18045 [12] were used to derive meaningful probability scores for certain threats. In order to make the evaluation process more objective, legal requirements for measuring instruments as laid down in [1] have been formalized, resulting in a list of assets to be protected and their respective security properties. With the aim of showing the feasibility of the approach, two real-world examples of measuring instruments were examined giving results that could be used as a feedback into the manufacturers design and production phase. Even though all application scenarios discussed here are from the sector of legal metrology, the approach may be of interest to evaluators of software in general. The formalization of assets and security requirements can, of course, be adapted to fit other legal or contractual obligations apart from the MID. The evaluation scheme can then be used in the same manner as was demonstrated here.

The two examples used for demonstrative purposes showed that the scheme can indeed provide meaningful results based on the information available to a Notified Body when assessing a manufacturer's design. If the source code also were available, the method from [7] could be applied to further validate the determined risk scenario. Even without additional information there are a number of steps that could be taken to improve the proposed approach:

In a first step, different evaluators will be asked to assess generic measuring instruments in a field test. To this end, the approach is currently being tested in a subgroup of WELMEC Working Group 7 ”Software“. During testing, the reproducibility of the assessment results can be investigated under realistic circumstances. Secondly, better developed attacker models will be incorporated to include more information about a measuring instrument's field of usage in the assessment. In this step, the motivation of certain attackers could also be added as another individual evaluation component. This change would also necessitate a modification of the point scores from the CC and will thus require very careful adjustments.

## References

[1] “Directive 2014/32/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of measuring instruments,” European Union, Council of the European Union ; European Parliament, Directive, February 2014.

[2] D. Peters, U. Grottker, F. Thiel, M. Peter, and J.-P. Seifert, “Achieving software security for measuring instruments under legal control,” in *Proceedings of the Federated Conference on Computer Science and Information Systems*, vol. 3, Warsaw Poland, September 2014, pp. 123–130, DOI: 10.15439/2014F460.

[3] “ISO/IEC 27005:2011(e) Information technology - Security techniques - Information security risk management,” International Organization for Standardization, Geneva, CH, Standard, June 2011.

[4] G. Geiger, “Ict Security Risk Management: Economic Perspectives,” in *Proceedings of the Federated Conference on Computer Science and Information Systems*, vol. 3, 2014, pp. 119–122, DOI: 10.15439/2014F439.

[5] “Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products,” European Union, Council of the European Union ; European Parliament, Regulation, July 2008.

[6] “WELMEC 5.3 Risk Assessment Guide for Market Surveillance: Weigh and Measuring Instrument,” European cooperation in legal metrology, WELMEC Secretariat, Ljubljana, Standard, May 2011.

[7] A. van Deursen and T. Kuipers, “Source-based software risk assessment,” in *Proceedings of the IEEE International Conference on Software Maintenance*. IEEE, September 2003, pp. 385–388, DOI: 10.1109/ICSM.2003.1235448.

[8] S.-W. Foo and A. Muruganantham, “Software risk assessment model,” in *Proceedings of the IEEE International Conference on Management of Innovation and Technology*, vol. 2. IEEE, November 2000, pp. 536–544, DOI: 10.1109/ICMIT.2000.916747.

[9] N. Greif and G. Parkin, “An international harmonised measurement software guide: the need and the concept,” in *Proceedings of the IMEKO World Congress Fundamental and Applied Metrology*, Lisbon, Portugal, September 2009, pp. 2440–2443.

[10] M. Sadiq, M. K. I. Rahmani, M. W. Ahmad, and S. Jung, “Software risk assessment and evaluation process (sraep) using model based approach,” in *Proceedings of the IEEE International Conference on Networking and Information Technology*. IEEE, June 2010, pp. 171–177, DOI: 10.1109/ICNIT.2010.5508535.

[11] “ISO/IEC 15408:2012 Common Criteria for Information Technology Security Evaluation,” International Organization for Standardization, Geneva, CH, Standard, September 2012, Version 3.1 Revision 4.

[12] “ISO/IEC 18045:2012 Common Methodology for Information Technology Security Evaluation,” International Organization for Standardization, Geneva, CH, Standard, September 2012, Version 3.1 Revision 4.

[13] “ETSI TS 102 165-1 Telecommunications and Internet converged Services and Protocols for Advanced Networking; Methods and protocols; Part 1: Method and proforma for Threat, Risk, Vulnerability Analysis,” European Telecommunications Standards Institute, Sophia Antipolis Cedex, FR, Standard, March 2011, v4.2.3.

[14] “WELMEC 7.2 Software Guide,” European cooperation in legal metrology, WELMEC Secretariat, Delft, Standard, March 2012.

[15] “CVE - Common Vulnerabilities and Exposures,” https://cve.mitre.org/, Accessed 04|17|2015.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1125–1134

DOI: 10.15439/2015F95
ACSIS, Vol. 5

# More Practical Application of Trust Management Credentials

Anna Felkner, Adam Kozakiewicz
NASK – Research and Academic Computer Network
Wawozowa 18, 02-796 Warsaw, Poland
Email: {anna.felkner, adam.kozakiewicz}@nask.pl

***Abstract*—Trust management is an approach to access control in distributed open systems, where access control decisions are based on policy statements made by multiple principals. The family of Role-based Trust management languages (RT) is an effective means for representing security policies and credentials in decentralized, distributed, large scale access control systems. It provides a set of role assignment credentials. A credential provides information about the privileges of users and the security policies issued by one or more trusted authorities.**

**The main purpose of this paper is to show how extensions can make the $RT$ family languages more useful in practice. It shows how security policies can be made more realistic by including timing information, maintaining the procedure or parameterizing the validity of credentials.**

## I. Introduction

THE modern human life is heavily reliant on trust. The complexity of tasks we are dealing with makes it practically impossible to succeed without delegating some of them. It is therefore not surprising that the issue of trust became a very active area of research quite long ago. The trust between humans has been thoroughly analysed both from the social sciences and economic point of view, as it is an important enabler of delegation. Attempts to transfer the concept of trust to different domains were made by computer scientists, including security, electronic commerce, semantic web or even social networks areas. In any network, be it computer or social one, trust remains an essential factor. The definition of trust is however not perfectly clear, different authors show slightly different definitions. Most often it is defined either on the basis of personal experience, reputation or recommendation.

The concept of trust is closely related to the notion of reputation – the opinion about a person held by others, e.g. the opinion about an Internet seller by his customers or the opinion about the behavior of a node in a wireless sensor network built by other nodes as a result of previous interactions. Reliability is another concept related to trust. Originally, this was a measure of the length of the period during which a machine can be considered trustworthy. In general, trust can be presented a a derivation of an entity's reputation.

Access control systems, based on traditional access control models (like Mandatory Access Control (MAC), Discretionary Access Control (DAC) and Role Based Access Control (RBAC)), are in essence identity based. Authorization decisions are based entirely on the role or – more directly – identity of the requesting party and can only be made if the requester is known to the owner of the resource. The decision is based on the relation of the only two entities involved – the owner of the protected resource, who is responsible for granting access, and the requester, who requires access.

In closed, centralized environments this approach is actually correct. As identity of the users of the system is established in advance, basing access decisions on it is a natural and easy to implement choice. Unfortunately this scenario does not scale well as the system becomes decentralized and highly distributed over a network. In such open systems the set of users is not only large, but may change dynamically, leading to entirely new challenges, such as the problem of propagation of information about the changes. A central database of user's identities is only a partial solution – although it enables the implementation of classic approaches to access control, it also introduces a single point of failure, where temporary lack of access to the central database makes any authorization decisions impossible system-wide. In absence of such a central identity repository it is no longer possible to assume that the identities of the requester and the resource owner are mutually recognized. A more flexible approach is needed, one which enables requesters to cross security domains and access resources owned by non-related entities. One such solution is called trust management.

Consider a simple example of a bookstore which offers special discounts to returning customers who are students. There is no obvious way to determine whether a given customer is eligible based only on his/her identity. Requesting a proof of identity will not resolve this problem. On the other hand, the problem can be solved efficiently using credentials, such as a bookstore card and a student card. The access rights are then determined not based on the users identity, but on the sufficiently documented information about the user's privileges assigned by other authorities and trust for those authorities. This requires a new, different approach to access control.

The rest of this paper is organized as follows. Trust management concept is shown in Section II and Role-based Trust management family of languages is shown in Section III. Section IV presents the Role-based Trust management language syntax with an example of $RT^T$ credentials. Inference system over $RT^T$ language is described in Section V. Section VI describes a few extensions of $RT^T$ language (time validity and determination of the order). Section VII shows an inference system over new $RT_+^T$ language time constraints. Final

978-83-60810-66-8/$25.00©2015, IEEE

remarks are given in the Conclusions.

## II. Trust Management

*Trust management* was first introduced as a term by Blaze et al. [2] in 1996. The term was defined as a unified approach to specify and interpret security policies, credentials and trust relationships. The privileges of an entity in such a system are based not on its identity, but on its attributes. Multiple principals have the right to issue credentials which are then used to demonstrate the entity's attributes. The definition of a *credential* is an attestation of authority, competence or qualification of an individual issued by a third party. The information contained in a credential includes privileges of a given user and/or security policies issued by trusted authorities. Real life examples of credentials are easy to propose – all sorts of academic diplomas, driver's licenses, identification documents, certificates or membership cards are clearly credentials. In fact, a real-life credential does not need to be a document in the traditional sense of the world – e.g. keys can be treated as a form of credential. In a computer system credentials may either be records available from one or more repositories, or – more flexibly – digitally signed documents which can be provided by the requester on demand. In the first case trust for the credential follows from trust for the repository and its own access control, in the second case it is provided by the signature, obviously requiring a trusted method of propagation or verification of public keys.

In literature, the earliest described example of a trust management application was PolicyMaker [3]. An assertion language defined in this system was capable of expressing locally trusted policy statements, as well as credentials requiring a digital signature using a private key. The second generation of trust management languages includes SPKI/SDSI [6], an enhanced version of PolicyMaker called KeyNote [4] and several other languages [7]. In these languages the privileges were still assigned directly to entities and delegation of permissions through credentials was performed only directly from the issuer to the subject. This generation still did not provide any mechanism enabling delegation to be separated from identity of the entity. Introduction of delegation based on atributes (as opposed to identity) was introduced in the next generation, represented by a family of Role-based Trust management (RT) languages [8], [16], [17]. Security policies are represented by defining a formalism using credentials to establish trust in distributed, decentralized access control systems.

## III. Role-based Trust Management Family Languages

The family of Role-based Trust management languages is used for representing security policies and credentials in decentralized, distributed access control systems. Several types of role assignment credentials are provided in $RT$ languages, depending on the language. $\mathbf{RT_0}$ [17] forms the core of the family, providing basic abilities – localization of authority for roles, delegation of that authority, role hierarchies and role intersections. These features are available in all $RT$ languages, who extend this set with new features. To represent relationships between entities, parametrized roles were introduced in $\mathbf{RT_1}$. To provide similar flexibility for resources as that provided for entities by roles, $\mathbf{RT_2}$ extends $RT_1$ with the notion of logical objects, enabling simple assignment of access rights for entire groups of logically related objects (resources). Note that both extentions presented so far do not actually change the expressive power of the language. They allow much more concise notation, but the same policy can in fact be expressed in $RT_0$, although with a much larger set of credentials, mapping each combination of parameters or each real instance of logical objects to a separate role. The first language actually adding new capabilities not present in other members of the family is the $\mathbf{RT^T}$ language, the main focus of this paper. The new capabilities include the ability to express agreement of multiple principals, even from disjoint sets, via manifold roles and separation of duties or threshold policies, via role-product operators.

A manifold role differs from a normal (singleton) role as instead of defining a set of principals, it defines a set of sets of principals. It is a wider term, since a singleton role can be expressed as a manifold role, whose principal sets are singletons, effectively meaning that cooperation is required from a group consisting of a single entity. Therefore, introduction of manifold roles does not affect the ability to express $RT_0$ credentials in the $RT^T$ language.

A threshold policy is used to specify a common occurrence, where agreement of multiple principals is required to initiate a given action. More formally, at least $k$ entities from a set satisfying certain conditions must agree on some fact. E.g. in banking certain transations require authorisation by two cashiers. Separation of duties policy is similar, but the agreeing entities fulfill different roles (e.g. in the banking example some transactions may also require authorisation by a controller). Both types of policies specify requirements that cannot be fulfilled by a single entity and therefore cannot be expressed in $RT_0$.

The $RT$ family includes one more important language, $\mathbf{RT^D}$, which provides mechanisms to describe delegation of role activations and selective use of role membership. However, this language is out of scope of this paper. More in-depth information about the $RT$ family of languages can be found in paper [16].

The main advantage of trust management approach is the ability to use *delegation*. A principal's authority over a resource may be transferred in a limited fashion to other principals by simple means of a credential. The notion of ownership is no longer central – the access control strategy and all decisions on who is authorized to use which resource is defined by a set of credentials. One upside of such approach is that the authority is easy to transfer over a network – as long as a credential can be transferred and trusted in another location, there is no need to involve the identity of the resource owner. However, although such decentralization of credential storage is very useful, it does present a variety of new problems.

In the years since its creation, the concept of trust management has evolved significantly, made applicable to new, broader contexts, such as assessment of reliability or trustworthiness of systems and individuals [15]. In this paper we restrict the meaning of trust management only to its original context of access control. A lot of different work connected with the possibility of using RT in different types of network, eg. Role based Trust management model for Peer-to-Peer networks [5], a trust management system for Ad-Hoc Networks [1], or Wireless Sensor Networks [11], and also Role-based Trust Management Model in Multi-domain Environment [18] have been studied recently.

Our approach is another type of extension of trust management languages. It shows how to make RT family languages more useful in practice by including time validity constraints and order of entities which can appear in the execution context.

## IV. The Syntax of $RT$ Family Languages

RT languages use a set of basic elements, such as entities, role names, roles and credentials. *Entities* are principals controlling access to resources by defining roles and issuing credentialsas well as requesters willing to access resources. The entity may be a person, but it might just as well be an application, identifying itself in a computer system by a user name or a public key. *Role names* represent permissions and can be issued to entities or groups of them by other entities. *Roles* represent sets of entities for which the access control policies grant particular permissions. *Credentials* define roles by appointing a new member of the role or by delegating authority to members of other roles.

$RT^T$ includes six different types of credentials, the first four in common with the less expressive $RT$ languages:

$A.r \leftarrow B$ – *simple membership*: entity $B$ is a member of role $A.r$.

$A.r \leftarrow B.s$ – *simple inclusion*: role $A.r$ includes (all members of) role $B.s$. This type of credential involves delegation of authority over role $r$, since by issuing new credentials defining $B.s$ $B$ may add new members to role $A.r$. This type of credential is also used to define role hierarchies.

$A.r \leftarrow B.s.t$ – *linking inclusion*: role $A.r$ includes role $C.t$ for each $C$, which is a member of role $B.s$. This is a delegation of authority from $A$ to all the members of the role $B.s$. The expression $B.s.t$ is called a *linked role*.

$A.r \leftarrow B.s \cap C.t$ – *intersection inclusion*: role $A.r$ includes all members of both roles $B.s$ and $C.t$. This is a partial delegation from $A$ to $B$ and $C$. The expression $B.s \cap C.t$ is known as an *intersection role*.

$A.r \leftarrow B.s \odot C.t$ – role $A.r$ can be satisfied by a union set containing one member of each of the two roles ($B.s$ and $C.t$). A single entity being a member of both roles suffices.

$A.r \leftarrow B.s \otimes C.t$ – role $A.r$ includes two *different* entities, one of which is a member of role $B.s$ and one a member of role $C.t$.

Since the models used in practice can be very complex, this paper uses some simplified examples. We focus on $RT^T$-specific credentials and intend to illustrate the basic notions and notation, not the full expressive power of the language.

*Example 4.1 (Example of $RT^T$ - subject):*

Suppose that a university will only activate a subject if at least two of four students apply and among the applicants at least one is a PhD student. $RT_0$ would require a long list of credentials listing all possible satisfactory combinations, changed whenever the list of students changes. $RT^T$ allows us to express this rule as just two policy credentials and a list of simple membership credentials, easy to manage along with the official list of students. The policy credentials are:

$$F.students \leftarrow F.student \otimes F.student \tag{1}$$

$$F.activeSubject \leftarrow F.students \odot F.phdStudent \tag{2}$$

Now, if the following membership credentials are added::

$$F.student \leftarrow \{Alex\} \tag{3}$$

$$F.student \leftarrow \{Betty\} \tag{4}$$

$$F.student \leftarrow \{David\} \tag{5}$$

$$F.student \leftarrow \{John\} \tag{6}$$

$$F.phdStudent \leftarrow \{John\} \tag{7}$$

$$F.phdStudent \leftarrow \{Emily\} \tag{8}$$

we can conclude that any pair of students from the set $\{Alex, Betty, David, John\}$ fulfills the role $F.students$ and that the subject can be activated if the pair includes John or if Emily is willing to attend.

*Example 4.2 (Example of $RT^T$ – signature):* Suppose that we have a situation in which we need to collect the signatures of the requester, accountant, his official superior, the manager of financial department and the director of a company to accept some transaction. Such a policy can be described using the following credentials:

$$\begin{gathered} Company.signature \leftarrow Company.requester \\ \odot\ Company.accountant \odot Company.superior \\ \odot\ Company.fdManager \odot Company.director \end{gathered} \tag{9}$$

Now suppose that we have such people, who play those roles, so the following credentials have been added to our security policy:

$$Company.requester \leftarrow \{Jacob\} \tag{10}$$

$$Company.accountant \leftarrow \{Jacob\} \tag{11}$$

$$Company.accountant \leftarrow \{Eliot\} \tag{12}$$

$$Company.accountant \leftarrow \{Alexander\} \tag{13}$$

$$Company.superior \leftarrow \{William\} \tag{14}$$

$$Company.superior \leftarrow \{Michael\} \tag{15}$$

$$Company.fdManager \leftarrow \{Jacob\} \tag{16}$$

$$Company.director \leftarrow \{William\} \tag{17}$$

As we can see, to complete the set of signatures we need just two people: $Jacob$, who can play a role of $requester$, $accountant$, $fdManager$ and $William$ who plays the role of $superior$ and $director$, but as may be required, groups of people $\{Jacob$, $Eliot$, $William\}$, $\{Jacob$, $Eliot$, $Michael$, $William\}$, $\{Jacob$, $Alexander$, $William\}$, and $\{Jacob$, $Alexander$, $Michael$, $William\}$ can also play a manifold role, and cooperatively complete the set of signatures.

## V. Inference System over $RT^T$ Credentials

$RT^T$ credentials define roles, which in turn are used to represent permissions. The set of member entities for a role is defined by a set $\mathcal{P}$ of $RT^T$ credentials. This set can be more conveniently calculated using an inference system, which defines an operational semantics of $RT^T$ language. The system consists of a set of inference rules used to derive credentials from existing ones and an initial set of formulae considered true.

Let $\mathcal{P}$ be a set of $RT^T$ credentials. The inference rules can be applied to create new credentials, derived from credentials of the set $\mathcal{P}$. A derived credential $c$ will be denoted using a formula $\mathcal{P} \succ c$, meaning that credential $c$ can be derived from a set of credentials $\mathcal{P}$.

The initial set of formulae of an inference system over a set $\mathcal{P}$ of $RT^T$ credentials are all the formulae: $c \in \mathcal{P}$ for each credential $c$ in $\mathcal{P}$. The inference rules of the system are the following:

$$\frac{c \in \mathcal{P}}{\mathcal{P} \succ c} \qquad (\mathbf{W_1})$$

$$\frac{\mathcal{P} \succ A.r \leftarrow B.s \qquad \mathcal{P} \succ B.s \leftarrow X}{\mathcal{P} \succ A.r \leftarrow X} \qquad (\mathbf{W_2})$$

$$\frac{\begin{array}{c}\mathcal{P} \succ A.r \leftarrow B.s.t \qquad \mathcal{P} \succ B.s \leftarrow C \\ \mathcal{P} \succ C.t \leftarrow X\end{array}}{\mathcal{P} \succ A.r \leftarrow X} \qquad (\mathbf{W_3})$$

$$\frac{\begin{array}{c}\mathcal{P} \succ A.r \leftarrow B.s \cap C.t \qquad \mathcal{P} \succ B.s \leftarrow X \\ \mathcal{P} \succ C.t \leftarrow X\end{array}}{\mathcal{P} \succ A.r \leftarrow X} \qquad (\mathbf{W_4})$$

$$\frac{\begin{array}{c}\mathcal{P} \succ A.r \leftarrow B.s \odot C.t \qquad \mathcal{P} \succ B.s \leftarrow X \\ \mathcal{P} \succ C.t \leftarrow Y\end{array}}{\mathcal{P} \succ A.r \leftarrow X \cup Y} \qquad (\mathbf{W_5})$$

$$\frac{\begin{array}{c}\mathcal{P} \succ A.r \leftarrow B.s \otimes C.t \qquad \mathcal{P} \succ B.s \leftarrow X \\ \mathcal{P} \succ C.t \leftarrow Y \qquad\qquad X \cap Y = \phi\end{array}}{\mathcal{P} \succ A.r \leftarrow X \cup Y} \qquad (\mathbf{W_6})$$

The inference systems of a language may not be unique – many different systems may be defined. There are however two properties required for the system to be useful in practice – the system must be sound an complete. Soundness guarantees that any formula derived by the system must be valid with respect to the semantics of the language, while completeness ensures that any valid formula is derivable.

All the credentials, which can be derived in the system, either belong to set $\mathcal{P}$ (rule $W_1$) or are of the type: $\mathcal{P} \succ A.r \leftarrow X$ (rules $W_2$ through $W_6$). Proof of soundness of the inference system involves showing that for each new formula $\mathcal{P} \succ A.r \leftarrow X$, the triple $(A, r, X)$ belongs to the semantics $S_{\mathcal{P}}$ of the set $\mathcal{P}$. Completeness is proved by showing that every formula $P \succ A.r \leftarrow X$ can be derived using inference rules for each element $(A, r, X) \in S_{\mathcal{P}}$. Both proofs can be found in [10], showing that the inference system is a valid alternative way of presenting the semantics of $RT^T$.

*Example 5.1 (Inference system for Example 4.1):* We will now derive the set of entities that can cooperate to activate a subject using an inference system, using a limited set of credentials for brevity: ((1), (2), (4), (6), and (7)). Using credentials (1), (2), (4), (6), and (7) according to rule $(W_1)$ we can infer:

$$\frac{F.students \leftarrow F.student \otimes F.student \in \mathcal{P}}{\mathcal{P} \succ F.students \leftarrow F.student \otimes F.student}$$

$$\frac{F.activeSubject \leftarrow F.students \odot F.phdStudent \in \mathcal{P}}{\mathcal{P} \succ F.activeSubject \leftarrow F.students \odot F.phdStudent}$$

$$\frac{F.student \leftarrow \{Betty\} \in \mathcal{P}}{\mathcal{P} \succ F.student \leftarrow \{Betty\}}$$

$$\frac{F.student \leftarrow \{John\} \in \mathcal{P}}{\mathcal{P} \succ F.student \leftarrow \{John\}}$$

$$\frac{F.phdStudent \leftarrow \{John\} \in \mathcal{P}}{\mathcal{P} \succ F.phdStudent \leftarrow \{John\}}$$

Then, using credentials (1), (6) and (4) and rule $(W_6)$ we infer:

$$\frac{\begin{array}{c}\mathcal{P} \succ F.students \leftarrow F.student \otimes F.student \\ \mathcal{P} \succ F.student \leftarrow \{John\} \\ \mathcal{P} \succ F.student \leftarrow \{Betty\} \\ \{John\} \cap \{Betty\} = \phi\end{array}}{\mathcal{P} \succ \mathbf{F.students} \leftarrow \{\mathbf{John}, \mathbf{Betty}\}}$$

These newly inferred credential and (2) and (7) with the rule $(W_5)$:

$$\frac{\begin{array}{c}\mathcal{P} \succ F.activeSubject \leftarrow F.students \odot F.phdStudent \\ \mathcal{P} \succ F.phdStudent \leftarrow \{John\} \\ \mathcal{P} \succ F.students \leftarrow \{John, Betty\}\end{array}}{\mathcal{P} \succ \mathbf{F.activeSubject} \leftarrow \{\mathbf{John}, \mathbf{Betty}\}},$$

we can show that the set of entities $\{John, Betty\}$ is sufficient to activate the subject.

*Example 5.2 (Inference system for Example 4.2):* We use the inference system to formally derive a set of entities who are essential to accept some transaction, i.e. the signatures of the requester, accountant, his official superior, the manager of financial department and the director of a company. To make the notation shorter, let us use $C$ instead of $Company$.

Using credentials (9)-(17) according to the rule $(W_1)$ we can infer:

$$\frac{\begin{array}{c} C.signature \leftarrow C.requester \\ \odot\ C.accountant\ \odot\ C.superior \\ \odot\ C.fdManager\ \odot\ C.director \in \mathcal{P} \end{array}}{\begin{array}{c} \mathcal{P} \succ C.signature \leftarrow C.requester \\ \odot\ C.accountant\ \odot\ C.superior \\ \odot\ C.fdManager\ \odot\ C.director \end{array}}$$

$$\frac{C.requester \leftarrow \{Jacob\} \in \mathcal{P}}{\mathcal{P} \succ C.requester \leftarrow \{Jacob\}}$$

$$\frac{C.accountant \leftarrow \{Jacob\} \in \mathcal{P}}{\mathcal{P} \succ C.accountant \leftarrow \{Jacob\}}$$

$$\frac{C.accountant \leftarrow \{Eliot\} \in \mathcal{P}}{\mathcal{P} \succ C.accountant \leftarrow \{Eliot\}}$$

$$\frac{C.accountant \leftarrow \{Alexander\} \in \mathcal{P}}{\mathcal{P} \succ C.accountant \leftarrow \{Alexander\}}$$

$$\frac{C.superior \leftarrow \{William\} \in \mathcal{P}}{\mathcal{P} \succ C.superior \leftarrow \{William\}}$$

$$\frac{C.superior \leftarrow \{Michael\} \in \mathcal{P}}{\mathcal{P} \succ C.superior \leftarrow \{Michael\}}$$

$$\frac{C.fdManager \leftarrow \{Jacob\} \in \mathcal{P}}{\mathcal{P} \succ C.fdManager \leftarrow \{Jacob\}}$$

$$\frac{C.director \leftarrow \{William\} \in \mathcal{P}}{\mathcal{P} \succ C.director \leftarrow \{William\}}$$

Then, using credentials (9), (10), (11), (14), (16) and (17) and rule $(W_5)$ we infer:

$$\frac{\begin{array}{c} \mathcal{P} \succ C.signature \leftarrow C.requester \\ \odot\ C.accountant\ \odot\ C.superior \\ \odot\ C.fdManager\ \odot\ C.director \\ \mathcal{P} \succ C.requester \leftarrow \{Jacob\} \\ \mathcal{P} \succ C.accountant \leftarrow \{Jacob\} \\ \mathcal{P} \succ C.superior \leftarrow \{William\} \\ \mathcal{P} \succ C.fdManager \leftarrow \{Jacob\} \\ \mathcal{P} \succ C.director \leftarrow \{William\} \end{array}}{\mathcal{P} \succ \mathbf{C.signature} \leftarrow \{\mathbf{Jacob}, \mathbf{William}\}}$$

showing that the set of entities $\{Jacob, William\}$ is sufficient to complete the set of signatures.

Or, using credentials (9), (10), (12), (14), (16) and (17) and rule $(W_5)$ we infer:

$$\frac{\begin{array}{c} \mathcal{P} \succ C.signature \leftarrow C.requester \\ \odot\ C.accountant\ \odot\ C.superior \\ \odot\ C.fdManager\ \odot\ C.director \\ \mathcal{P} \succ C.requester \leftarrow \{Jacob\} \\ \mathcal{P} \succ C.accountant \leftarrow \{Eliot\} \\ \mathcal{P} \succ C.superior \leftarrow \{William\} \\ \mathcal{P} \succ C.fdManager \leftarrow \{Jacob\} \\ \mathcal{P} \succ C.director \leftarrow \{William\} \end{array}}{\mathcal{P} \succ \mathbf{C.signature} \leftarrow \{\mathbf{Jacob}, \mathbf{Eliot}, \mathbf{William}\}}$$

showing that here we need more people to complete the set of signatures, i.e. $\{Jacob, Eliot, William\}$.

Or, using credentials (9), (10), (13), (15), (16) and (17) and rule $(W_5)$ we infer:

$$\frac{\begin{array}{c} \mathcal{P} \succ C.signature \leftarrow C.requester \\ \odot\ C.accountant\ \odot\ C.superior \\ \odot\ C.fdManager\ \odot\ C.director \\ \mathcal{P} \succ C.requester \leftarrow \{Jacob\} \\ \mathcal{P} \succ C.accountant \leftarrow \{Alexander\} \\ \mathcal{P} \succ C.superior \leftarrow \{Michael\} \\ \mathcal{P} \succ C.fdManager \leftarrow \{Jacob\} \\ \mathcal{P} \succ C.director \leftarrow \{William\} \end{array}}{\mathcal{P} \succ \mathbf{C.signature} \leftarrow \{\mathbf{Jacob}, \mathbf{Alexander}, \mathbf{Michael}, \mathbf{William}\}}$$

showing that here we need four people to complete the set of signatures, i.e. $\{Jacob, Alexander, Michael, William\}$.

Depending on which credentials at the moment we have (because not all the credentials are always available), we can determine the sets of people who can cooperatively sign the document.

## VI. Credential extensions

This section shows a few extensions of $RT^T$ languages which make it more useful in practice.

### A. Time Validity in $RT^T$

Real security policies involve time restrictions. Allowing the credentials to have limited time validity can make the $RT^T$ language more useful in practice. Inference rules with time validity for $RT_0$ were originally introduced in a slightly different way in [14]. In [13] we tried to extend the potential of $RT^T$ language by adding time validity constraints. Most permissions are in fact given for fixed periods of time, permanent permissions are less common. Time dependent credentials take the form: $c$ **in** $v$, meaning "the credential $c$ is available during the time $v$". Finite sets of time dependent credentials are denoted by $\mathcal{CP}$ and the new language is called $RT_+^T$. $c$ is used to denote "$c$ **in** $(-\infty,\ +\infty)$" to make notation lighter.

Most trust management languages are monotonic: adding new assertion to a query can never result in canceling an action, which was accepted before [9]. Therefore, each policy statement or credential added to the system may only increase the capabilities and privileges granted to others, making revocation of rights impossible. Introduction of time constraints does not invalidate the monotonicity of the system, but achieves some of the utility of negation.

Time validity can be denoted as follows:
$[\tau_1,\tau_2]; [\tau_1,\tau_2); (\tau_1,\tau_2]; (\tau_1,\tau_2); (-\infty,\tau]; (-\infty,\tau);$
$[\tau,+\infty); (\tau,+\infty); (-\infty,\ +\infty); v_1 \cup v_2; v_1 \cap v_2; v_1 \backslash v_2$
and $v_1$, $v_2$ of any form in this list, with $\tau$ ranging over time constants.

*Example 6.1 (Time validity for Example 4.1):*
Assuming that $Alex$, $Betty$, $David$ and $John$ in our scenario will not be student forever seems quite natural. $John$ and $Emily$'s PhD student status is similar. Thus, credentials (3)–(8) should be generalized to:

$$F.student \leftarrow \{Alex\} \textbf{ in } v_1 \tag{18}$$

$$F.student \leftarrow \{Betty\} \textbf{ in } v_2 \tag{19}$$

$$F.student \leftarrow \{David\} \textbf{ in } v_3 \tag{20}$$

$$F.student \leftarrow \{John\} \textbf{ in } v_4 \tag{21}$$

$$F.phdStudent \leftarrow \{John\} \textbf{ in } v_5 \tag{22}$$

$$F.phdStudent \leftarrow \{Emily\} \textbf{ in } v_6 \tag{23}$$

stating that (3) – (8) are only available during $v_1$, $v_2$, $v_3$, $v_4$, $v_5$, and during $v_6$, respectively. The policy itself, described by credentials (1) and (2) may however be permanent. By using (1), (2) and (18)–(23), we want to be able to derive that for example the set $\{Alex, Betty, John\}$ can cooperatively activate the subject during all of the period: $v_1 \cap v_2 \cap v_5$ or $\{Betty, John\}$ during the time $v_2 \cap v_4 \cap v_5$ or $\{Alex, David, Emily\}$ during the time intersection $v_1 \cap v_3 \cap v_6$. Another set of people can cooperatively activate the subject (depending of the time).

*Example 6.2 (Time validity for Example 4.2):* In our scenario, it is quite natural to assume that $Jacob$ are a requester only for a fixed period of time. The same with $Jacob$, $Eliot$ and $Alexander$ as a company accountants, also $Wiliam$ and $Michael$ as a superior, and $Jacob$ as a financial department manager, as well as $William$ as a director. Thus, credentials (10)–(17) should be generalized to:

$$Company.requester \leftarrow \{Jacob\} \textbf{ in } v_1 \tag{24}$$

$$Company.accountant \leftarrow \{Jacob\} \textbf{ in } v_2 \tag{25}$$

$$Company.accountant \leftarrow \{Eliot\} \textbf{ in } v_3 \tag{26}$$

$$Company.accountant \leftarrow \{Alexander\} \textbf{ in } v_4 \tag{27}$$

$$Company.superior \leftarrow \{William\} \textbf{ in } v_5 \tag{28}$$

$$Company.superior \leftarrow \{Michael\} \textbf{ in } v_6 \tag{29}$$

$$Company.fdManager \leftarrow \{Jacob\} \textbf{ in } v_7 \tag{30}$$

$$Company.director \leftarrow \{William\} \textbf{ in } v_8 \tag{31}$$

stating that (10) – (17) are only available during $v_1$, $v_2$, $v_3$, $v_4$, $v_5$, $v_6$, $v_7$, and during $v_8$, respectively. On the other hand, credential (9) can be always valid, as it expresses some time-independent fact. Now, by using (9) and (24)–(31), we want to be able to derive that for example the set $\{Jacob, William\}$ can cooperatively sign the document during all of the period: $v_1 \cap v_2 \cap v_5 \cap v_7 \cap v_8$, where $Jacob$ plays the role of *requester*, *accountant* and *financial department manager* and $William$ acts as a *superior* and *director of the company*. But during the time $v_1 \cap v_3 \cap v_6 \cap v_7 \cap v_8$ it has to be the set consisting of $\{Jacob, Eliot, Michael, William\}$.

While in both examples the policy defining credentials were assummed to be permanently valid, this is not required. Some policies are naturally time-limited (eg. seasonal sales).

### *B. Determination of the order*

Another powerful feature which would be useful to model more realistic policies is the ability to determine the order in which a member of a role or an entity (entities) can appear.

If we want to maintain a procedure, we have to add two new types of credentials at the syntax level. These are:

$A.r \leftarrow B.s \overset{\odot}{\rightarrow} C.t$ – role $A.r$ is satisfied by a union set of one member of role $B.s$ and one member of role $C.t$ in this exact order or by one entity satisfying the intersection role $B.s \cap C.t$.

$A.r \leftarrow B.s \overset{\otimes}{\rightarrow} C.t$ – role $A.r$ is satisfied by a set of two different entities: one member of role $B.s$ and one member of role $C.t$ in this order.

In our Example 4.1 we can want to have such situation:

$$F.activeSubject \leftarrow F.students \overset{\odot}{\rightarrow} F.phdStudent$$

which means that the order is important. First we need to have two students and just after that one PhD student.

That extension can be extremely useful in a large variety of situations. For example, if we have a situation, when one person is a member of a few roles, it can be useful to have some restrictions connected with appearing in particular roles during the execution context when the credential is used.

*Example 6.3 (The right order of signature):*
When we have a situation in which we need to collect the signatures of people who are essential to accept some transaction, we can imagine at least a few scenarios in our security policy.

Suppose that we need a signature of the requester, accountant, his official superior, the manager of financial department and the director of a company, in such order.

Now we can use the data from *Example* 4.2 and we can have three different scenarios:

1) The order is strictly obeyed and it is important that an accountant can give his signature after having

received the signature of a requester, and accountant's superior can give his signature after having received the signature of an accountant, even if it is one person. This means that in a first step $Jacob$ can sign the document as a $requester$, in a second step as an $accountant$, after that $William$ can give his signature as a $Jacob's$ official superior. In the next step $Jacob$ can sign the document as a financial department manager, and at the end $William$ can sign the document as a director of the company. Table I presents the signature order in our first scenario.

It can be important in some situation to strictly keep the order of signatures, but in a huge implementation it can be a little bit inefficient. That is why we can propose two other scenarios.

2) We can allow signing the document by one person who plays a few roles at once if the roles appear in credentials successively without any role between. In our example it can look like in the Table II (to make our example easier, we can use just credentials (10), (11), (14), (16), (17)).
   In such a simple example, we have one step less than in the previous scenario. It shows how such change can be useful in real large systems.
3) In our third scenario we can allow that one person, who plays more than one role, can give all the signatures at once. It can be very useful in an automatic implementation. Table III shows how it can look in our third scenario.
   That situation means that $Jacob$ accepts his signature as a financial department manager if $William$ signs the document as his official superior and $William$ accepts his signature as a $director$ if $Jacob$ signs the document as a financial department manager. We have to have a possibility to accept or not accept our signature, which is dependent on another person's signature.

If we want to mandate that the entity can appear in particular roles during the execution context exactly when the credential is used, we can put a new type of role denoted by underlined identifiers (e.g. $\underline{r}, \underline{s}, \underline{t}$). In such situation, when we change the credential:

$$\begin{gathered} Company.signature \leftarrow Company.requester \\ \overset{\odot}{\rightarrow} Company.accountant \overset{\odot}{\rightarrow} Company.superior \\ \overset{\odot}{\rightarrow} Company.fdManager \overset{\odot}{\rightarrow} Company.director \end{gathered}$$

into:

$$\begin{gathered} Company.signature \leftarrow Company.requester \\ \overset{\odot}{\rightarrow} Company.accountant \overset{\odot}{\rightarrow} Company.superior \\ \overset{\odot}{\rightarrow} Company.fdManager \overset{\odot}{\rightarrow} Company.\underline{director} \end{gathered}$$

in our third scenario we will have the situation described in Table IV, meaning that $William$ has to wait with his signature as a $director$ untill the time $Jacob$ approves his signature as $fdManager$.

All the semantics previously defined for $RT^T_+$, set-theoretic (which maps roles to a set of entity names), operational semantics (where credentials can be derived from the initial set of credentials using a set of inference rules [13]), and logic-programming (where credentials are translated into a logic program [12]), are still valid, meaning that proofs of the soundness and the completeness of that semantics are also valid.

This section shows how we can explore the potential of RT languages. It shows how security policies can be made more realistic by including timing information or maintaining the procedure.

## VII. Inference System for $RT^T$ Credentials with Time Validity

All the semantics previously defined for $RT$ languages are still valid for determined order, but they have to be changed for credentials with time validity. Because of that reason we have to take on it. This section is showing an inference system for $RT^T$ credentials with time validity.

We will now adapt the inference system over $RT^T$ credentials to respect time validity. Let $\mathcal{CP}$ be a set of $RT^T$ credentials, from which new credentials may be derived. A derived credential $c$ valid in time $\tau$ will be denoted using a formula $\mathcal{CP} \succ_\tau c$, meaning that the credential $c$ can be derived from a set of credentials $\mathcal{CP}$ during the time $\tau$. The initial set of formulae of an inference system over a set $\mathcal{CP}$ of $RT^T_+$ credentials are all the form: $c \ \mathbf{in} \ v \ \in \mathcal{CP}$ for each credential $c$ valid in time $v$ in $\mathcal{CP}$. The inference rules of the system are the following:

$$\frac{c \ \mathbf{in} \ v \in \mathcal{CP} \quad \tau \in v}{\mathcal{CP} \succ_\tau c} \qquad (\mathbf{CW_1})$$

$$\frac{\mathcal{CP} \succ_\tau A.r \leftarrow B.s \quad \mathcal{CP} \succ_\tau B.s \leftarrow X}{\mathcal{CP} \succ_\tau A.r \leftarrow X} \qquad (\mathbf{CW_2})$$

$$\frac{\begin{gathered}\mathcal{CP} \succ_\tau A.r \leftarrow B.s.t \qquad \mathcal{CP} \succ_\tau B.s \leftarrow C \\ \mathcal{CP} \succ_\tau C.t \leftarrow X\end{gathered}}{\mathcal{CP} \succ_\tau A.r \leftarrow X} \qquad (\mathbf{CW_3})$$

$$\frac{\begin{gathered}\mathcal{CP} \succ_\tau A.r \leftarrow B.s \cap C.t \qquad \mathcal{CP} \succ_\tau B.s \leftarrow X \\ \mathcal{CP} \succ_\tau C.t \leftarrow X\end{gathered}}{\mathcal{CP} \succ_\tau A.r \leftarrow X} \qquad (\mathbf{CW_4})$$

$$\frac{\begin{gathered}\mathcal{CP} \succ_\tau A.r \leftarrow B.s \odot C.t \qquad \mathcal{CP} \succ_\tau B.s \leftarrow X \\ \mathcal{CP} \succ_\tau C.t \leftarrow Y\end{gathered}}{\mathcal{CP} \succ_\tau A.r \leftarrow X \cup Y} \qquad (\mathbf{CW_5})$$

$$\frac{\begin{gathered}\mathcal{CP} \succ_\tau A.r \leftarrow B.s \otimes C.t \qquad \mathcal{CP} \succ_\tau B.s \leftarrow X \\ \mathcal{CP} \succ_\tau C.t \leftarrow Y \qquad X \cap Y = \phi\end{gathered}}{\mathcal{CP} \succ_\tau A.r \leftarrow X \cup Y} \qquad (\mathbf{CW_6})$$

All the derived credentials either belong to set $\mathcal{CP}$ (rule $CW_1$) or are of the type: $\mathcal{CP}_\tau \succ A.r \leftarrow X$ (rules $CW_2$ through $CW_6$). This new inference system is based on an extension of the inference rules from section V, where rules

TABLE I
SIGNATURE ORDER IN THE FIRST SCENARIO

| **Step** | **requester** | **accountant** | **superior** | **fdManager** | **director** |
|---|---|---|---|---|---|
| **1** | $Jacob$ | $\phi$ | $\phi$ | $\phi$ | $\phi$ |
| **2** | $Jacob$ | $Jacob, Eliot, Alexander$ | $\phi$ | $\phi$ | $\phi$ |
| **3** | $Jacob$ | $Jacob, Eliot, Alexander$ | $William, Michael$ | $\phi$ | $\phi$ |
| **4** | $Jacob$ | $Jacob, Eliot, Alexander$ | $William, Michael$ | $Jacob$ | $\phi$ |
| **5** | $Jacob$ | $Jacob, Eliot, Alexander$ | $William, Michael$ | $Jacob$ | $William$ |

TABLE II
SIGNATURE ORDER IN THE SECOND SCENARIO

| **Step** | **requester** | **accountant** | **superior** | **fdManager** | **director** |
|---|---|---|---|---|---|
| **1** | $Jacob$ | $Jacob$ | $\phi$ | $\phi$ | $\phi$ |
| **2** | $Jacob$ | $Jacob$ | $William$ | $\phi$ | $\phi$ |
| **3** | $Jacob$ | $Jacob$ | $William$ | $Jacob$ | $\phi$ |
| **4** | $Jacob$ | $Jacob$ | $William$ | $Jacob$ | $William$ |

TABLE III
SIGNATURE ORDER IN THE SECOND SCENARIO

| **Step** | **requester** | **accountant** | **superior** | **fdManager** | **director** |
|---|---|---|---|---|---|
| **1** | $Jacob$ | $Jacob$ | $\phi$ | $Jacob$ | $\phi$ |
| **2** | $Jacob$ | $Jacob$ | $William$ | $Jacob$ | $William$ |

TABLE IV
SIGNATURE ORDER IN THE THIRD "UNDERLINED" SCENARIO

| **Step** | **requester** | **accountant** | **superior** | **fdManager** | **director** |
|---|---|---|---|---|---|
| **1** | $Jacob$ | $Jacob$ | $\phi$ | $Jacob$ | $\phi$ |
| **2** | $Jacob$ | $Jacob$ | $William$ | $Jacob$ | $\phi$ |
| **3** | $Jacob$ | $Jacob$ | $William$ | $Jacob$ | $William$ |

$(W_i)$ are replaced with $(CW_i)$ and only valid time-dependent credentials from $\mathcal{CP}$ are considered.

The proof of soundness of the inference system requires showing that for each new formula $\mathcal{CP}_\tau \succ A.r \leftarrow X$, the triple $(A, r, X)$ belongs to the semantics $S_{\mathcal{CP}}$ of the set $\mathcal{CP}$. All the formulae $\mathcal{CP}_\tau \succ A.r \leftarrow X$, such that $A.r \leftarrow X \in \mathcal{CP}$ are sound, as shown in [12].

Completeness of the inference system over a set $\mathcal{CP}$ of $RT_+^T$ credentials can be proved by showing that a formula $\mathcal{CP} \succ A.r \leftarrow X$ can be derived using inference rules for each element $(A, r, X) \in S_{\mathcal{CP}}$. The proof is presented in [12].

### *A. Inferring time validity of credentials*

The proposed inference system can also derive the maximal time validity of a credential $c$ from $\mathcal{CP}$. Formula $\mathcal{CP} \succ_\tau c$ is modified to $\mathcal{CP} \succ\!\succ_v c$, meaning that at any time $\tau \in v$ in which $\mathcal{CP}$ has a semantics, it is possible to infer the credential $c$ from $\mathcal{CP}$. The inference rules of the system are the following:

$$\frac{c \textbf{ in } v \in \mathcal{CP}}{\mathcal{CP} \succ\!\succ_v c} \qquad (\mathbf{CWP_1})$$

$$\frac{\mathcal{CP} \succ\!\succ_{v_1} A.r \leftarrow B.s \quad \mathcal{CP} \succ\!\succ_{v_2} B.s \leftarrow X}{\mathcal{CP} \succ\!\succ_{v_1 \cap v_2} A.r \leftarrow X} \qquad (\mathbf{CWP_2})$$

$$\frac{\begin{array}{c}\mathcal{CP} \succ\!\succ_{v_1} A.r \leftarrow B.s.t \\ \mathcal{CP} \succ\!\succ_{v_2} B.s \leftarrow C \quad \mathcal{CP} \succ\!\succ_{v_3} C.t \leftarrow X\end{array}}{\mathcal{CP} \succ\!\succ_{v_1 \cap v_2 \cap v_3} A.r \leftarrow X} \qquad (\mathbf{CWP_3})$$

$$\frac{\begin{array}{c}\mathcal{CP} \succ\!\succ_{v_1} A.r \leftarrow B.s \cap C.t \\ \mathcal{CP} \succ\!\succ_{v_2} B.s \leftarrow X \quad \mathcal{CP} \succ\!\succ_{v_3} C.t \leftarrow X\end{array}}{\mathcal{CP} \succ\!\succ_{v_1 \cap v_2 \cap v_3} A.r \leftarrow X} \qquad (\mathbf{CWP_4})$$

$$\frac{\begin{array}{c}\mathcal{CP} \succ\!\succ_{v_1} A.r \leftarrow B.s \odot C.t \\ \mathcal{CP} \succ\!\succ_{v_2} B.s \leftarrow X \quad \mathcal{CP} \succ\!\succ_{v_3} C.t \leftarrow Y\end{array}}{\mathcal{CP} \succ\!\succ_{v_1 \cap v_2 \cap v_3} A.r \leftarrow X \cup Y} \qquad (\mathbf{CWP_5})$$

$$\frac{\begin{array}{c}\mathcal{CP} \succ\succ_{v_1} A.r \leftarrow B.s \otimes C.t\\ \mathcal{CP} \succ\succ_{v_2} B.s \leftarrow X \quad \mathcal{CP} \succ\succ_{v_3} C.t \leftarrow Y\\ X \cap Y = \phi\end{array}}{\mathcal{CP} \succ\succ_{v_1 \cap v_2 \cap v_3} A.r \leftarrow X \cup Y} \qquad (\mathbf{CWP_6})$$

$$\frac{\mathcal{CP} \succ\succ_{v_1} c \quad \mathcal{CP} \succ\succ_{v_2} c}{\mathcal{CP} \succ\succ_{v_1 \cup v_2} c} \qquad (\mathbf{CWP_7})$$

Rule is $(CWP_1)$ claims that $\mathcal{CP}$ can be used whenever it is valid. Rules $(CWP_2)$ - $(CWP_6)$ simply claim that inference rules can be used iff all their premises are true. Finally, the rule $(CWP_7)$ is used to join validity periods, meaning that if $c$ can be inferred both with validity $v_1$ and validity $v_2$, then it can also be inferred with validity $v_1 \cup v_2$. $\mathcal{CP} \succ\succ_v$ generalizes $\mathcal{CP} \succ_\tau$. They are both equivalent whenever $v = [\tau, \tau]$. Note that inferring a certain $c$ from $\mathcal{CP}$ may be possible in several different ways, resulting in different validity periods. Rule $(CWP_7)$ can then be used as many times as necessary to broaden $c$'s validity.

Maximal inference is the process, where in each step we infer with maximal time validity.

An inference terminating in $\mathcal{CP} \succ\succ_v c$ is called maximal if and only if:

1) there exists no $v' \supset v$ such that $\mathcal{CP} \succ\succ_{v'} c$, and
2) every its sub-inference terminating in $\mathcal{CP} \succ\succ_{v''} c'$, for $c' \neq c$ is maximal.

The first condition ensures that further use of rule $(CWP_7)$ will not extend the validity of $c$. The second condition ensures that this property is propagated through the whole inference tree. Maximal inferences guarantee that $v$ in $(CWP_1)$ is the maximal time validity for $A.r \leftarrow X$.

For these inferences we can prove soundness and completeness of $\mathcal{CP} \succ\succ_v$, as shown in [12].

*Example 7.1 (Time validity in inference system for Example 4.1):* Let us get back to our example and to make long example shorter, let us use less credentials: (1), (2), (19), (21), and (22). According to rule $(CWP_1)$ we can infer:

$$\frac{F.students \leftarrow F.student \otimes F.student \in \mathcal{CP}}{\mathcal{CP} \succ\succ F.students \leftarrow F.student \otimes F.student}$$

$$\frac{F.activeSubject \leftarrow F.students \odot F.phdStudent \in \mathcal{CP}}{\mathcal{CP} \succ\succ F.activeSubject \leftarrow F.students \odot F.phdStudent}$$

$$\frac{F.student \leftarrow \{Betty\} \textbf{ in } v_2 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_2} F.student \leftarrow \{Betty\}}$$

$$\frac{F.student \leftarrow \{John\} \textbf{ in } v_4 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_4} F.student \leftarrow \{John\}}$$

$$\frac{F.phdStudent \leftarrow \{John\} \textbf{ in } v_5 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_5} F.phdStudent \leftarrow \{John\}}$$

When we want to check when two different students can cooperate, from credentials (1), (19), (21) and rule $(CWP_6)$ we infer:

$$\frac{\begin{array}{c}\mathcal{CP} \succ\succ F.students \leftarrow F.student \otimes F.student\\ \mathcal{CP} \succ\succ_{v_2} F.student \leftarrow \{Betty\}\\ \mathcal{CP} \succ\succ_{v_4} F.student \leftarrow \{John\}\\ \{Betty\} \cap \{John\} = \phi\end{array}}{\mathcal{CP} \succ\succ_{v_2 \cap v_4} \mathbf{F.students} \leftarrow \{\mathbf{Betty}, \mathbf{John}\}}$$

In next step we use it and additionally credentials (2), (22) and rule $(CWP_5)$:

$$\frac{\begin{array}{c}\mathcal{CP} \succ\succ F.activeSubject \leftarrow F.students \odot F.phdStudent\\ \mathcal{CP} \succ\succ_{v_5} F.phdStudent \leftarrow \{John\}\\ \mathcal{CP} \succ\succ_{v_2 \cap v_4} F.students \leftarrow \{Betty, John\}\end{array}}{\mathcal{CP} \succ\succ_{v_2 \cap v_4 \cap v_5} \mathbf{F.activeSubject} \leftarrow \{\mathbf{Betty}, \mathbf{John}\}}$$

showing that the set of entities that can cooperatively activate a subject is: $\{Betty, John\}$ during the time: $v_2 \cap v_4 \cap v_5$.

*Example 7.2 (Time validity in inference system for Example 4.2):* Let us get back to our example and use credentials (9) and (24)–(31) (to make the notation shorter, let us use $C$ instead of $Company$). According to rule $(CWP_1)$ we can infer:

$$\frac{\begin{array}{c}C.signature \leftarrow C.requester\\ \odot\ C.accountant\ \odot\ C.superior\\ \odot\ C.fdManager\ \odot\ C.director \in \mathcal{CP}\end{array}}{\begin{array}{c}\mathcal{CP} \succ\succ C.signature \leftarrow C.requester\\ \odot\ C.accountant\ \odot\ C.superior\\ \odot\ C.fdManager\ \odot\ C.director\end{array}}$$

$$\frac{C.requester \leftarrow \{Jacob\} \textbf{ in } v_1 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_1} C.requester \leftarrow \{Jacob\}}$$

$$\frac{C.accountant \leftarrow \{Jacob\} \textbf{ in } v_2 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_2} C.accountant \leftarrow \{Jacob\}}$$

$$\frac{C.accountant \leftarrow \{Eliot\} \textbf{ in } v_3 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_3} C.accountant \leftarrow \{Eliot\}}$$

$$\frac{C.accountant \leftarrow \{Alexander\} \textbf{ in } v_4 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_4} C.accountant \leftarrow \{Alexander\}}$$

$$\frac{C.superior \leftarrow \{William\} \textbf{ in } v_5 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_5} C.superior \leftarrow \{William\}}$$

$$\frac{C.superior \leftarrow \{Michael\} \textbf{ in } v_6 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_6} C.superior \leftarrow \{Michael\}}$$

$$\frac{C.fdManager \leftarrow \{Jacob\} \textbf{ in } v_7 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_7} C.fdManager \leftarrow \{Jacob\}}$$

$$\frac{C.director \leftarrow \{William\} \textbf{ in } v_8 \in \mathcal{CP}}{\mathcal{CP} \succ\succ_{v_8} C.director \leftarrow \{William\}}$$

Now, when we want to check when $Jacob$ and $William$ are the only people, who are necessary to cooperatively sign the document we use credentials (9), (24), (25), (28), (30), (17) and rule $(CWP_5)$:

$$\frac{\begin{array}{c}\mathcal{CP} \succ\succ C.signature \leftarrow C.requester\\ \odot\ C.accountant\ \odot\ C.superior\\ \odot\ C.fdManager\ \odot\ C.director\\ \mathcal{CP} \succ\succ_{v_1} C.requester \leftarrow \{Jacob\}\\ \mathcal{CP} \succ\succ_{v_2} C.accountant \leftarrow \{Jacob\}\\ \mathcal{CP} \succ\succ_{v_5} C.superior \leftarrow \{William\}\\ \mathcal{CP} \succ\succ_{v_7} C.fdManager \leftarrow \{Jacob\}\\ \mathcal{CP} \succ\succ_{v_8} C.director \leftarrow \{William\}\end{array}}{\begin{array}{c}\mathcal{CP} \succ\succ_{v_1\ \cap\ v_2\ \cap\ v_5\ \cap\ v_7\ \cap\ v_8} \mathbf{C.signature}\\ \leftarrow \{\mathbf{Jacob}, \mathbf{William}\}\end{array}}$$

showing that the set of entities $\{Jacob, William\}$ is sufficient to complete the set of signatures during the time $v_1 \cap v_2 \cap v_5 \cap v_7 \cap v_8$.

Or, using credentials (9), (24), (26), (28), (30) and (31) and rule $(CWP_5)$ we infer:

$$\begin{array}{c}
\mathcal{CP} \succ\!\succ C.signature \leftarrow C.requester \\
\odot\ C.accountant\ \odot\ C.superior \\
\odot\ C.fdManager\ \odot\ C.director \\
\mathcal{CP} \succ\!\succ_{v_1} C.requester \leftarrow \{Jacob\} \\
\mathcal{CP} \succ\!\succ_{v_3} C.accountant \leftarrow \{Eliot\} \\
\mathcal{CP} \succ\!\succ_{v_5} C.superior \leftarrow \{William\} \\
\mathcal{CP} \succ\!\succ_{v_7} C.fdManager \leftarrow \{Jacob\} \\
\mathcal{CP} \succ\!\succ_{v_8} C.director \leftarrow \{William\} \\
\hline
\mathcal{CP} \succ\!\succ_{v_1 \cap v_3 \cap v_5 \cap v_7 \cap v_8} \mathbf{C.signature} \\
\leftarrow \{\mathbf{Jacob}, \mathbf{Eliot}, \mathbf{William}\}
\end{array}$$

showing that here we need more people to complete the set of signatures, i.e. $\{Jacob, Eliot, William\}$ during the time $v_1 \cap v_3 \cap v_5 \cap v_7 \cap v_8$.

Or, using credentials (9), (24), (27), (29), (30) and (31) and rule $(CWP_5)$ we infer:

$$\begin{array}{c}
\mathcal{CP} \succ\!\succ C.signature \leftarrow C.requester \\
\odot\ C.accountant\ \odot\ C.superior \\
\odot\ C.fdManager\ \odot\ C.director \\
\mathcal{CP} \succ\!\succ_{v_1} C.requester \leftarrow \{Jacob\} \\
\mathcal{CP} \succ\!\succ_{v_4} C.accountant \leftarrow \{Alexander\} \\
\mathcal{CP} \succ\!\succ_{v_6} C.superior \leftarrow \{Michael\} \\
\mathcal{CP} \succ\!\succ_{v_7} C.fdManager \leftarrow \{Jacob\} \\
\mathcal{CP} \succ\!\succ_{v_8} C.director \leftarrow \{William\} \\
\hline
\mathcal{CP} \succ\!\succ_{v_1 \cap v_4 \cap v_6 \cap v_7 \cap v_8} \mathbf{C.signature} \\
\leftarrow \{\mathbf{Jacob}, \mathbf{Alexander}, \mathbf{Michael}, \mathbf{William}\}
\end{array}$$

showing that here we need four people to complete the set of signatures, i.e. $\{Jacob, Alexander, Michael, William\}$ during the time $v_1 \cap v_4 \cap v_6 \cap v_7 \cap v_8$.

## VIII. Conclusions

In the paper we model the use of trust management systems in decentralized and distributed environments. The modelling framework is a family of Role-based Trust management language $RT^T$. The core part of the paper is introduction of time validity constraints and especially maintaining the procedure – modifications aimed at making the $RT^T$ language more realistic. While the inference systems presented in this paper are simple, they are well-founded theoretically. The utility of the proposed extentions is most visible in large-scale distributed systems, where users have only partial view of their execution context.

## References

[1] R. Akbani, T. Korkmaz, G.V.S. Raju, "Mobile Ad-Hoc Networks Security", Z. Qian et al. (Eds.): *Recent Advances in in Computer Science and Information Engineering*, Springer-Verlag Berlin Heidelberg 2012, pp. 659–666. http://dx.doi.org/10.1007/978-3-642-25769-8_92

[2] M. Blaze, J. Feigenbaum, J. Lacy, "Decentralized Trust Management", *Proc. 17th IEEE Symposium on Security and Privacy,* Oakland CA, 1996, pp. 164–173. http://dx.doi.org/10.1109/SECPRI.1996.502679

[3] M. Blaze, J. Feigenbaum, and M. Strauss, "Compliance checking in the PolicyMaker trust management system", in Proc. 2nd Int. Conf. Financial Cryptogr., London, UK, 1998, pp. 254–274. http://dx.doi.org/10.1007/BFb0055488

[4] M. Blaze, J. Feigenbaum, and A. D. Keromytis, "The role of trust management in distributed systems security" in Secure Internet Programming, J. Vitek, C. Damsgaard Jensen, Eds. London: Springer, 1999, pp. 185–210. http://dx.doi.org/10.1007/3-540-48749-2_8

[5] S. Chithra, "A Role Based Trust Model for Peer to Peer Systems Using Credential Trees", *International Journal of Computer Theory and Engineering*, Vol.3, No.2, April 2011, ISSN: 1793-8201, pp. 234–239. http://dx.doi.org/10.7763/IJCTE.2011.V3.310

[6] D. Clarke et al., "Certificate chain discovery in SPKI/SDSI", J. Comp. Secur., vol. 9, pp. 285–322, 2001.

[7] P. Chapin, C. Skalka, and X. S. Wang, "Authorization in trust management: Features and foundations", ACM Comput. Surv., vol. 3, pp. 1–48, 2008. http://dx.doi.org/10.1145/1380584.1380587

[8] M. R. Czenko, S. Etalle, D. Li, and W. H. Winsborough, "An Introduction to the Role Based Trust Management Framework RT", Tech. Rep. TR-CTIT-07-34, Centre for Telematics and Information Technology University of Twente, Enschede, The Netherlands, 2007. http://dx.doi.org/10.1007/978-3-540-74810-6_9

[9] M. R. Czenko et al., "Nonmonotonic Trust Management for P2P Applications", in Proc. 1st Int. Worksh. Secur. Trust Manag. STM 2005, Milan, Italy, 2005. http://dx.doi.org/10.1016/j.entcs.2005.09.037

[10] A. Felkner, K. Sacha, "Deriving $RT^T$ Credentials for Role-Based Trust Management", *e-Informatica Software Engineering Journal*, Volume 4, No 1, 2010, pp. 9–19.

[11] A. Felkner, "How the Role-based Trust Management Can be Applied to Wireless Sensor Nnetworks", *Journal of Telecommunications and Information Technology,* Volume 4, 2012, pp.70–77.

[12] A. Felkner, A. Kozakiewicz, "$RT^T_+$-Time Validity Constraints in $RT^T$ Language", *Journal of Telecommunications and Information Technology,* Volume 2, 2012, pp. 74–82.

[13] A. Felkner, A. Kozakiewicz, "Time Validity in Role-based Trust Management Inference System", *Secure and Trust Computing, Data Management, and Applications Communications in Computer and Information Science,* Volume 187, 2011, pp. 7–15. http://dx.doi.org/10.1007/978-3-642-22365-5_2

[14] D. Gorla, M. Hennessy, V. Sassone, "Inferring Dynamic Credentials for Role-Based Trust Management", *Proc. 8th Conference on Principles and Practice of Declarative Programming*, ACM, 2006, pp. 213–224. http://dx.doi.org/10.1145/1140335.1140361

[15] W. M. Grudzewski, I.K. Hejduk, A.Sankowska, M. Wańtuchowicz, "Trust Management in Virtual Work Environments: A Human Factors Perspective", *CRC Press Taylor & Francis Group*, 2008.

[16] N. Li, J. Mitchell, W. Winsborough, "Design of a Role-Based Trust-Management Framework". Proc. IEEE Symposium on Security and Privacy. IEEE Computer Society Press, Oakland CA (2002), pp. 114–130. http://dx.doi.org/10.1109/SECPRI.2002.1004366

[17] N. Li, W. Winsborough, J. Mitchell, "Distributed Credential Chain Discovery in Trust Management". J. Comput. Secur. 1 (2003), pp. 35–86.

[18] H. Liu, Q. Zhang, J. Zheng, X. Guo, "Role-based Trust Management Model in Multi-domain Environment", *TELKOMNIKA Indonesian Journal of Electrical Engineering*, Vol 11, No 1: January 2013, pp. 417–424

DOI: 10.15439/2015F359
ACSIS, Vol. 5

# A generic framework to support participatory surveillance through crowdsensing

Apostolos Malatras, Laurent Beslay
European Commission, Joint Research Centre (JRC), Institute for the Protection and Security of the Citizen
Email: apostolos.malatras@jrc.ec.europa.eu, laurent.beslay@jrc.ec.europa.eu

***Abstract*—Harnessing the power and popularity of participatory or opportunistic sensing for the purpose of providing added value security and surveillance services is a promising research direction. However, challenges such as increased privacy concerns, as well as technological issues related to the reliable processing and meaningful analysis of the collected data, hinder the widespread deployment of participatory surveillance applications. We present here our work on addressing some of the aforementioned concerns through our related participatory application that focuses on crisis management and in particular buildings' evacuation. We discuss the technical aspects of our work, the viability and practicality of which is validated by means of a real experiment comprising 14 users in the context of an emergency evacuation exercise.**

## I. Introduction

RECENT developments regarding the capabilities of smartphones that are increasingly equipped with middle- to high-end sensors and their widespread penetration in modern society have spawn a novel paradigm of information generation and sharing, that of participatory sensing [1]. In this bottom-up paradigm illustrated in Figure 1, users of smartphones take advantage of the capabilities of the devices that they are carrying in terms of sensing and collect data regarding their surrounding environment and themselves, e.g. acceleration, temperature, light, sound, etc. They then proceed with sharing this information with other users either by uploading it to a common repository accessible to everyone (perhaps in the form of a map service where the location of the collected data is also pinpointed), or by sending their data to a centralized entity that provides them with related services [2].

The paradigm of participatory sensing is applicable to a wide range of application domains. Of particular interest is its consideration in light of security applications, in which case a new research domain, i.e. that of participatory surveillance, emerges [3]. Participatory surveillance refers to the use of principles from participatory sensing in order to monitor, control, and assess a variety of events for the purpose of security [4]. For example, the evacuation of a building could be enhanced when having access to data collected from the evacuees through their smartphones or in case of criminal activities the legal and police authorities could have a wealth of data coming from smartphones of nearby people to support their investigations [5], even in the absence of operational networking infrastructures. The latter data yield no significant information as such, but subject to processing using machine learning techniques they could be used to deduce useful knowledge about the activities the users were conducting at the time of data collection.

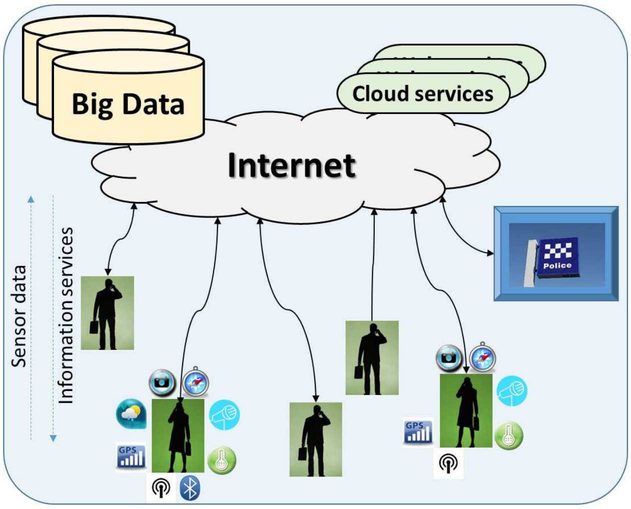


Fig. 1. Principles of operation of participatory surveillance.

To examine the viability of such scenarios we built a prototype participatory surveillance application and staged an evacuation exercise to validate its viability and practicality in real settings. We also designed a participatory surveillance framework and an experimental methodology to collect and analyze the data gathered throughout the exercise. One of the main goal of this research work was to examine the potential knowledge that can be derived from raw data stemming from the participatory sensing tasks. The driving objective was our ability to infer the different types of activities that the users were engaged in during the experiments, using only raw sensor data as input. We were also interested in examining how much information can be extracted from a minimal set of data and the results discussed later were quite interesting in terms of privacy.

We report here on our findings utilizing artificial intelligence and in particular machine learning algorithms to preprocess and analyze the collected data in order to infer the type of activities the users were conducting when the data was being collected. The results are very promising, with various configurations of our framework being able to identify up to

978-83-60810-66-8/$25.00©2015, IEEE

99% of users' activities (walking, standing still, climbing stairs or descending stairs) on the actual data collected from the evacuation exercise.

The remaining of this paper is structured as follows. After this brief introduction, Section II reviews related work in the area of participatory sensing in the context of security. Section III discusses the design of our participatory surveillance application together with potential scenarios of its use, whereas Section IV outlines privacy concerns and considerations regarding participatory surveillance. Section V presents the proposed generic framework to support analysis of data collected through such applications, evaluation of the accuracy and performance of which is the subject of Section VI. The paper concludes with Section VII where the limitations of this work are underlined and opportunities for further research in the domain are pinpointed.

## II. Related Work

Sensors on mobile phones can be used to infer different types of information regarding the users of the phones, as well as the surrounding environment. Accordingly, an extensive review is presented in [6]. Accelerometers, gyroscopes, and other sensors have been used to detect human activities with typical cases presented in [7], [8], [9]. Such works are significant for security reasons, because they can reveal the state of users, e.g. running or laying still, and thus in combination with other contextual information they can hint on possibly suspicious actions of users, e.g. running away from a crime scene, while everybody else around the user is walking. Moreover, considering a smartphone's microphone as input, audio recordings regarding noteworthy events could be recorded in an inconspicuous manner, whereas the phone camera could serve as an additional channel of information reporting, as discussed in [10].

The prominence and ubiquity of current smartphones that are equipped with a variety of embedded sensors has spurred applications related to user-centric sensing and monitoring of the surroundings, namely the paradigm of participatory sensing [1]. Such applications have found applicability in environmental monitoring [11], green vehicle routing [12], noise mapping in urban environments [13] and lately in security and surveillance operations [4]. Despite the fact that there are several potential shortcomings from such an approach, i.e. regarding privacy [14] and the quality and accuracy of data collection [15], its benefits nonetheless are far from negligible.

In the context of security, participatory sensing applications can provide great data collection services at high granularity (spatial and temporal) and at a low cost. Current surveillance practices rely mostly on video monitoring, e.g. CCTV, which has several shortcomings in person identification [16] and cannot inherently cover extended areas. While there has been work on using sensors as information side channels for security operations, e.g. microphone [17], accelerometers and PIR [18], and magnetometers [19], such techniques have however not been envisaged at a large scale. Limited number of sensors were deployed in previous works and such approaches therefore also suffer from poor range and poor data records. Participatory sensing builds on the use of sensors on smartphones that are nowadays pervasive and ubiquitous and can thus provide information for large geographical areas, assuming a large volunteer user base. The associated costs are minimal compared to the installation and deployment of infrastructure-based solutions, while additionally the ease of deployment is great since nowadays users carry their phones with them for the major party of the day and it is always on, collecting sensor data. Another benefit of participatory sensing applications for security purposes can be found in the straightforward identification of people that is supported through the cell IDs of the phones and allows the association of users with the data related to them. The *Cell-All* project from the US Department of Homeland Security on the use of chemical sensors on mobile phones to detect chemical attack related emergency situations was one of the first efforts towards crowdsensing being put in use for security purposes [20].

Furthermore, one of the biggest concerns in participatory sensing is ensuring that users are actively contributing and sharing their data [21], [22] since it could eventually lead to poor performance due to the lack of accurate and informative representations. This problem has been considered in the context of noise mapping, where a persuasive, motivating game was considered in [23] to stimulate user data collection and sharing. In terms of participatory surveillance, we postulate that the citizens' sense of engagement and contribution in securing their environments will be the driving factor for their engagement. Nevertheless, in order to promote such engagement the fundamental issue raised by users, i.e. privacy, should be addressed. In this respect, there has been significant research work on the anonymization of shared user data with prominent examples being reported in [24] and [25].

## III. Participatory Surveillance Scenarios

Participatory surveillance as a concept aims at utilizing the notions of participatory sensing and the ubiquity of smartphones equipped with a wealth of sensors in order to provide services related to surveillance and public security. This paradigm shift aims at empowering both the citizen and the police authorities and raising public awareness and citizen engagement [3]. Citizens on one hand feel more empowered since they are contributing in securing their neighborhoods and acquire thus a more active role in their society. Police authorities on the other hand gain from the wealth of data coming from citizens' smartphones and other monitoring means. It is an inexpensive and efficient way of enriching the data coming from traditional sources of surveillance, e.g. CCTV, as well as reaching areas where deployment of CCTV-like systems is not possible or allowed. Moreover, data coming from smartphones is not only limited to pictures or videos, but it can also include data from the embedded sensors such as accelerometer, gyroscope, pressure, magnetometer, etc. This kind of data supports the police authorities in gaining a better understanding of potentially noteworthy incidents [4].

In this respect, we considered a scenario that involved a variety of actors in order to collect experimental datasets from smartphone sensors and also to test the feasibility of the notion of participatory surveillance. We chose to avoid scenarios that would seem invasive and that might be considered as threatening citizens' privacy, e.g. continuous localization using GSM signals or recording the audio and video signals surrounding citizens at all times. The scenario we considered involved crisis management and in particular an evacuation exercise of an office building in case an emergency occurs, e.g. fire. People inside the building, namely employees, safety staff, building delegates (in charge of enforcing standard evacuation procedures) and visitors, were assumed to have a smartphone equipped with a custom application that monitored the values registered from their sensors and reported the data back to a centralized control room.

In terms of participatory surveillance, the principal goal was the utilization of the collected sensor datasets to provide the remotely located control room supervisors with useful knowledge regarding the progress of the evacuation, e.g. bottlenecks in exits indicated by users standing still, running in stairs indicating panic, users in peculiar situations (lying down or falling). Accordingly, the collected datasets that contained raw sensor data, e.g. from accelerometers or gyroscopes, had to be processed and analyzed in a manner that would allow us to extract useful information about user activities related to the collected data. The motivation behind this exercise was to examine whether data from smartphones alone would be sufficient to support surveillance tasks during an emergency when infrastructure surveillance mechanisms such as CCTV would be unavailable, namely to study the potential use of smartphones and participatory applications as a backup channel for surveillance.

The evacuation exercise scenario is in our view typical of prospective participatory surveillance applications, since it exhibits the majority of desired characteristics. In particular, it considers the use of a variety of sensors and a large number of people; it is privacy-friendly since it allows people to decide when and what type of data they wish to share; it is easily extensible to include further features and data sources; it provides solid motivation for the use of smartphones for security operations, since in such a case the lack of infrastructure surveillance network would be detrimental to police operations. Undoubtedly, the major concern of user privacy is present in this scenario as well as in every other participatory sensing application, albeit at a smaller scale. In the following, we discuss relevant privacy concerns and describe our approach in alleviating them.

## IV. Privacy Considerations

While the benefits stemming from participatory surveillance applications are evident, the privacy risks involved are not clear and need to be carefully considered. The mere concept of participatory surveillance comes along with a series of potential privacy risks. Users are required to share personal data coming from the sensors embedded on their phones, in order to support and improve security operations and promote the communal sense of safety. We illustrate in Section VI that even with the use of data coming from just the accelerometer, it is possible to infer the activities that the user was conducting at the time of data collection. Use of additional sensors could exacerbate this risk, providing more detailed information on user activities. Indicative of relevant privacy risks, is the recent work presented in [26] that considered RF-sensing to infer the state of device-free individuals without their cooperation. Taking into account the capabilities of modern smartphones and the wealth of data available to participatory surveillance applications, it becomes clear that proper privacy enhancing technologies need to be put in place.

A major concern refers to the fact that user data can easily be traced back to their owner, because of the nature of cellular networks and the uniqueness of phone identifiers. Since data can be used to expose potential private user information, anonymization techniques need to be utilized to hinder such exposure. A comprehensive review of related solutions can be found in [24], with the most prominent approaches considering techniques such as k-anonymity and l-diversity [25]. Moreover, the entire space of sensors on smartphones needs to be carefully examined. The latter are nowadays carrying a large number of sensors and the knowledge that can be extracted from them (by processing the corresponding sensed data) is still not fully chartered. Studies like the one presented in this paper, expose the privacy risks related to the accelerometer, however the need to perform similar studies for the entirety of available sensors is paramount. Accordingly, guidelines could be provided to the end users to instruct them on the potential risks involved in sharing data coming from diverse sensors. This is particularly important since users are quite sensitive about sharing photos or location data for example, but are unaware of the risks involved in sharing data from low level sensors that could be equally detrimental to their privacy [14].

Access to this data should also be protected, so as not to allow its unauthorized viewing and processing. Participatory surveillance is the context in which this data is being collected and therefore police authorities should access this data under this particular context. The role of national Data Protection Authorities (DPAs) naturally emerges as a safepoint of supervisory control regarding participatory data protection. In addition to any such effort, participatory surveillance applications should also be designed with privacy in mind, in which case they should record any access to data and the processing method invoked on them to assist in prospective inquiries. Furthermore, who will have ownership of this data and for how long it should be kept and processed remains an open issue that could trigger conflicting situations. One could for example postulate that data should be retained by police authorities only when a criminal activity took place and then stored indefinitely. However, the exposure of such an activity cannot be foreseen and therefore data should be kept to ensure its availability if needed. Another conflict that might arise involves the user wishing to remove his shared data (right to be forgotten), whereas the police authorities might not permit this due to ongoing or forthcoming investigations. There is no panacea to resolve such conflicts, with application- and context-dependent solutions usually being the norm. Undoubtedly, appropriate legislation and rule systems should be introduced to regulate this newly established field and thus promote its prosperity.

## V. Data Analysis Framework Design

The notion of participatory surveillance refers to the collection of data from a variety of smartphones and other mobile devices and in particular from the on-board sensors carried by such devices. Processing and analyzing this wealth of data could lead to the inference of interesting information and knowledge regarding various surveillance aspects, namely the identification of distinct human activities, the location of a user and his/her surroundings (physical and social) and the occurrence of abnormal conditions, e.g. extreme sound levels potentially attributed to screaming or intense physical stress possibly attribute to user falling. The goal of the data analysis part of any effective participatory surveillance system is therefore to deduce such useful information. In the following we present a generic framework that has been devised for such purposes of data analysis and discuss its various aspects.

Analyzing data to extract useful patterns and accordingly use these to identify human activities and distinguish between them has been a very active research domain over the years [27]. The goal is to have computing systems capable of inferring knowledge by themselves using only raw data as input. The main elements of machine learning for the purposes of activity recognition using smartphones include the following:

- Data collection: collection of data using sensors located on the users' smartphone.
- Data training: the collected raw data need to be processed in order to deduce some useful information features and characteristics that will assist in its classification.
- Data classification: use of the aforementioned features in conjunction with machine learning classification algorithms to classify data, i.e. assign classes to data instances.

Figure 2 depicts these 3 different steps in the machine learning process. We applied these steps on both the collected reference data and the test data for our experiments and appropriately configured the classification process to enhance its performance in respect to our requirements. During the training phase the most appropriate and appropriately configured classifier is selected, so as to be applied in the testing phase over new data and classify them accordingly.

As previously discussed, sensor data can be exploited to detect human activities and thus provide insight on the actions and whereabouts of the smartphone users in a non-intrusive manner. We decided to use a systematic approach to tackle this problem and for this reason introduced a generic framework for the analysis of data coming from participatory surveillance activities. The main goal is to build a comprehensive dataset for training statistical classifier and applying this to actual, i.e. test, data to establish possible patterns/matches and thus identify human activities for the purposes of surveillance. The reason why we chose to define a generic framework is to establish a methodological framework of doing similar experiments, as well as to facilitate further developments in the domain. The framework is presented in Figure 3, while its elements are detailed in the following:

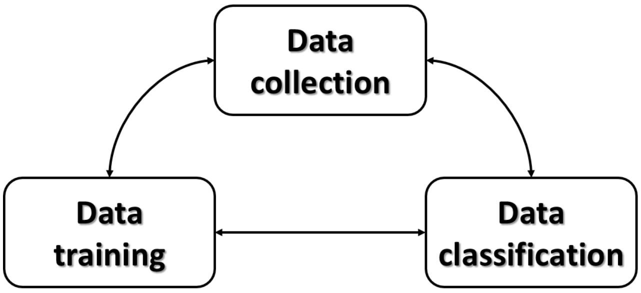


Fig. 2. Machine learning stages for activity recognition applied on both training and testing phases.

- Define problem space: the particulars of the participatory surveillance tasks need to be carefully defined at a high level, namely what needs to be achieved. They will serve as the requirements that will drive the rest of the analysis process.
- Define activities to be identified: not all scenarios for participatory surveillance rely on the identification of the same set of activities. However, the selection of the interesting activities is important at an early stage since it drives the definition of the required data and sensors to monitor these activities.
- Define sensors to be used: having defined the activities we are interested in, the next step concerns the selection of the most appropriate sensors to support the identification of these activities.
- Plan and conduct training experiments: the training phase is the first phase in the machine learning process and it involves a set of base experiments to collect reference data for the activities in question. The planning of these experiments is therefore of paramount importance, so as no configuration parameters or testing conditions become neglected during the following phase.
- Collect datasets for training: collect training data referring to elementary activities related to participatory surveillance as defined in previous steps of the process. The data should refer to more than one repetitions of the activity over a span of time.
- Pre-process training data: the collected data is in most cases noisy and needs to be pre-processed prior to being used by machine learning classifiers. Raw data usually has very fine granularity making it difficult to discern relevant statistical properties. Therefore it needs to be processed in order to extract statistical features over time, as well as in the frequency domain. The pre-processing tasks involve the removal of outliers from the original dataset, the annotation of the data for classification pur-

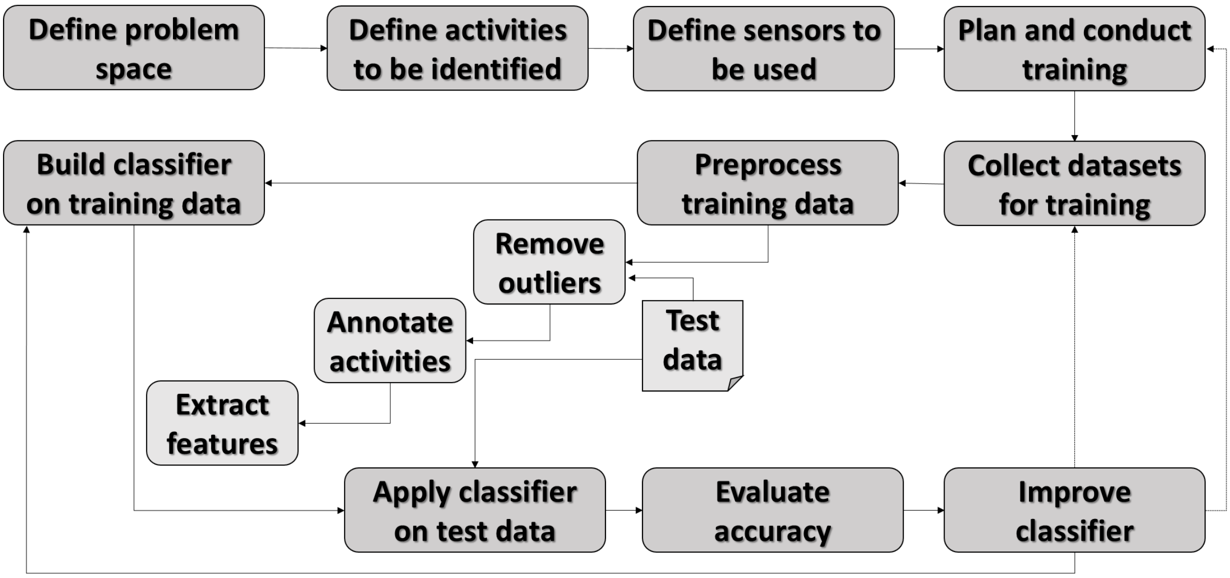

Define problem space
Define activities to be identified
Define sensors to be used
Plan and conduct training
Build classifier on training data
Preprocess training data
Collect datasets for training
Remove outliers
Annotate activities
Test data
Extract features
Apply classifier on test data
Evaluate accuracy
Improve classifier

Fig. 3. Generic framework for the analysis of participatory surveillance data.

poses (applies only to training data, since the classifier will ¿predictř the class of the test data) and the extraction of statistical features.

- Build classifier based on training data: amongst the large number of classification algorithm available in the related literature, the optimal one for the particular type of collected data and extracted features should be selected. Each classifier has a set of configuration parameters and a sensitivity analysis of each of them and their influence on the accuracy of the classifier needs to be performed in order to conclude on the most appropriate classifier for the considered experimental settings.
- Apply classifier on test data: having decided on the optimal classifier, it needs to be applied on the collected test data (a posteriori or at runtime depending on the experimental settings). The classifier should be able to identify the class to which each of the test data belongs and decide upon that.
- Evaluate accuracy: the accuracy of the classifier is evaluated against the ground truth, hence the need to properly and accurately annotate both the training and test data.
- Improve classifier: test and training data are of the same type but a lot of irregularities might appear on the test data that might not have been present in the training data. There are many reasons for this, most important of which is the fact that training data users are rarely the same as test data users and thus do not have the same physiological patterns. Therefore, the classifier might not predict the test data as efficiently as expected and further modifications need to be applied on it and accordingly the experiments might need to be performed again.

In what follows we elaborate on the elements of our proposed framework in the context of our participatory surveillance evaluation case-study, i.e. the evacuation exercise.

## VI. Evaluation

To validate the feasibility of participatory surveillance and evaluate its efficiency we conducted an evacuation exercise experiment, where the main goal was to establish whether the use of smartphones' sensors as a backup channel for information collection can yield information about users' activities. In this respect, we built a comprehensive dataset for training a statistical classifier and applied this to the test data collected during the experiment to establish possible pattern matches and subsequently analyze the results to augment the design of the assumed participatory surveillance system.

### A. *Implementation*

In order to collect data from the sensors embedded on the smartphones of the users we experimented initially with the Funf open sensing framework [28] running on Android platforms, but we finally opted for a custom built application to avoid the unnecessary complexity in configuring Funf. In particular, Funf allows the developer to set preferences in regards to data collection, such as the types of sensors to be monitored, the duration of the monitoring and recording phase and the interval between two consecutive recording phases. A limitation of this framework is the fact that the actual frequency of data collection, while seemingly subject to a user's preferences, is actually a compromise between the value set by the user and the frequency that Android itself and the corresponding sensors can actually support. Android allows for 4 different rates for data collection from sensors, i.e. *normal, UI, game, fastest*. However, the actual values for these frequencies differ between different sensors, which further complicates matters when it comes to data collected using Funf. For these reasons we built our custom application to collect sensor data, the main distinguishing feature of which is that it allows explicit setting of data collection frequency per sensor.

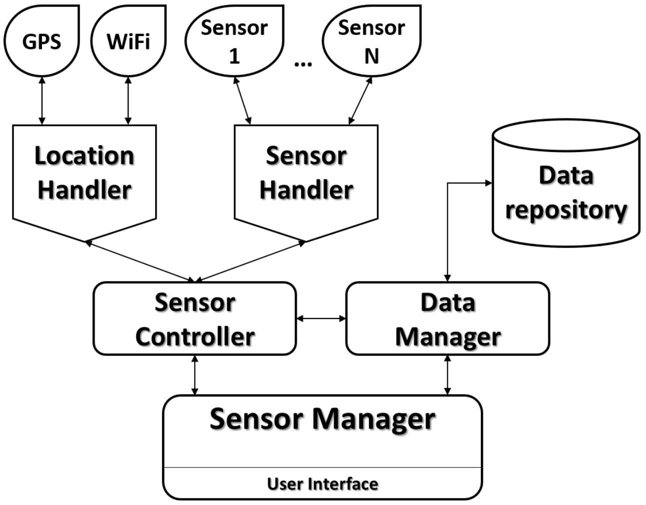


Fig. 4. Architecture of sensor data collection application.

The application we built has a modular and extensible architecture that is shown in Figure 4. The *Sensor Manager* is the main entry point for the app, currently running on Android phones, by means of a dedicated GUI that allows the user to select the sensor the values of which she is interested in recording. Two main elements of the architecture are the *Sensor Controller* and the *Data Manager*. The former interacts with the *Location Handler* to retrieve current location and the *Sensor Handler* to get low level system access to the embedded sensors and retrieve their values, while the latter parses this information into an appropriate data format and stores it to a local *Data Repository*. Moreover, the *Data Manager* supports management of this repository, i.e. search, update, delete data, applying access control policies to avoid unauthorized access to private data and logging every request for data.

### B. Experimental setup

The evacuation exercise experiment was conducted in a public building at the JRC premises and it comprised both floors of the building, as well as the parking space, where the actual evacuation meeting points is located. The participatory surveillance exercise was part of a larger experiment that aimed amongst others at evaluating additional techniques such as indoor localization using smartphones and facial recognition through the cameras on the smartphones. For the participatory surveillance tasks, 14 actors were involved: 1 building delegate and 13 regular users carrying their smartphones. The latter had our custom application deployed in order to collect sensor data, e.g. accelerometer, and the users were also instructed at times (via SMS from the control room) to use their smartphones to record video of their surroundings. Users and the building delegate were equipped with different types of smartphones (Samsung Galaxy Nexus, Sony Xperia S, HTC One) to account for the diversity of existing platforms and embedded sensors.

Upon completion of the evacuation exercise we had a total of 13 datasets from a corresponding number of phones that had collected sensor data through our application. Unfortunately, the instability of the application and of the smartphones' platform led to not all phones having recorded data (only 6 out of 13 phones reported worthwhile data). It has to be clarified that the reasons for the erroneous, wrongly timed and limited data collection cannot be pinpointed to a particular event. They can be attributed to users not having activated all services, e.g. location reporting services, networking hindrances due to obstacles or other reasons that collectively prevented timely data reporting, smartphone being overloaded, smartphone battery having been depleted, etc. The major problem that we encountered was the concurrent and synchronized collection of sensor and localization data. This was necessary in order to be able to reason about the sensor data and to pinpoint the location of interesting events. In the future a larger user test base should be considered and an extensive preparation phase prior to the experiments should take place to ensure proper operation. Moreover, test users should not be left to freely interact with the considered applications and services, but instead they should follow a strict script/set of actions in order to ensure that the results we obtain will not be biased by the individual users' attitudes.

The frequency of data collection from the sensors was set to 500Hz (every 2ms), which intuitively is rather high to allow for distinguishing between differences in human activities. Moreover, it reduces the battery level significantly since the device is constantly operational and collecting data. In future experiments we plan to perform a thorough sensitivity analysis of this variable, as well as of other variables in our configuration. The sensors involved in data collection include the triaxial accelerometer, microphone, battery, gravity sensor, gyroscope, light sensor, magnetometer, orientation, WiFi and Bluetooth wireless interfaces and the proximity sensor. While data was collected for all the aforementioned sensors, we focus our analysis here on those collected by accelerometers. The reason behind this decision lies in the fact that in this evacuation exercise, identification of human activities, e.g. walking or running, was the main goal and in this respect the accelerometer has been the sensor most widely used for this purpose in related research works [29].

Moreover, due to the established limitations of the sensing capabilities of modern smartphones and their issues with calibration (see for example [30], [31]), sensors such as the gyroscope, magnetometer and orientation are considered to be of limited accuracy and this placed constraints on their use in our experimental design. An additional reason for not using these sensors was the fact that they are dependent on many external features, e.g. the way the phone is being held, specific rooms/configurations where the experiments took place, so it would have been harder to deduce any useful context from such sensors considering the difficulties in repeating experimental settings and conditions. We therefore opted to place emphasis on data collection by the triaxial accelerometer present in all current smartphones for the analysis of the data collected in the participatory surveillance project. Evidently, the training dataset can be easily expanded to include data from the other sensors in order to establish whether useful knowledge about the users' status and activities can be extracted from them.

### *C. Training phase*

The focal objective of building a reference training dataset is to establish a solid and wide ground truth to be used for the evaluation of the participatory surveillance experiments, i.e. the identification of distinct human activities. In this respect, we collected training data with a specific focus on data that corresponds to the activities expected to take place during the evacuation exercise; walking, walking up stairs, walking down stairs, standing still. It should be clarified that the more activities we consider, the more accurate the prediction will be since there will be more details and more patterns derived for each of these activities. However, the increase in the number of considered activities brings a corresponding increase in the complexity of the data pre-processing and classification processes. Considering this trade-off we opted for a training dataset containing accelerometer data corresponding to the 4 basic activities mentioned before.

The training dataset contains 5 minutes worth of data collected for each of the aforementioned activities, which were conducted independently and annotated immediately after completion to limit possible annotation errors. Users were asked to retain constant gait and velocity as much as possible. We performed 10 recording sub-sessions, i.e. 40 second sub-sessions, for each activity to eliminate possible user fatigue that would influence data collection. In addition, only 30 seconds were considered out of each sub-session; we trimmed the first and last 5 seconds to omit outliers during initialization and finalization of the activity. Lastly, we used the exact same settings in terms of data collection frequency, smartphone devices and phone placement as in the actual evacuation exercise.

### *D. Data preprocessing*

The collected training data refers to 3 streams of measurements, one for each axis of the accelerometer. They represent continuous discrete samples that need to be pre-processed prior to making any analysis based on them. In accordance with typical activity recognition algorithms, we commence by cleaning the raw data and removing the outliers, followed by the application of a windowing technique to extract groups of data that could potentially expose repetitive activities or tasks with common characteristics, both of which are representative of the majority of human activities. We built a set of custom Java tools that facilitate the automation and efficient preprocessing of the raw data and hence allow to invest more time on the data analysis tasks. All the data are stored as CSV files, the first line of which indicates the type of data.

Removing outliers is a very important task in data preprocessing, since these values could influence the outcome of the analysis process as they are not conforming to the rest of the data and have thus been potentially generated by side activities to the one currently under examination. The most common technique we use to clean the dataset is to remove the influence stemming from the initialization and finalization of activities. We are interested in the execution of the actual activity and therefore we trim the dataset on both ends accordingly. This is similar to the technique used in [7] and we also take into account the risk of significantly reducing the size of the dataset; this is the reason why we collected additional data as previously described. More advanced techniques, e.g. mean, Kalman, particle filters, could also be applied, but since the application of these filters is subject to intense processing requirements, we opted against applying any such technique on the collected accelerometer data.

Subsequently we applied a windowing algorithm to create logical instances, i.e. windows, of the original dataset. The windows are used to reduce the problem space on one hand, but also to assist in grouping similar samples on the other hand. Generally speaking there exist three different windowing techniques, namely sliding, event-based and activity-based windows, applied to raw data for the purpose of recognition of human activities, as discussed in [32]. We used the sliding window technique with overlap 50% over the entire training data population, for a total of over 60,000 samples. Table I illustrates the different values for the window size that we experimented with and the corresponding duration of the window and generated instances in the dataset based on the assumed data collection frequency. We should note that the same techniques for data preprocessing are applied in both the training data as well as the test data during the actual classification process.

TABLE I
WINDOW SIZE, DURATION AND SIZE OF INSTANCES IN DATASETS

| Window size | Duration | Instances in dataset |
|---|---|---|
| 64 | 1.28s | 1875 |
| 128 | 2.56s | 937 |
| 256 | 5.12s | 468 |
| 512 | 10.24s | 232 |

The sampling frequency of the original training dataset is too high for individual samples to exhibit any interesting properties, especially in regards to human activities that normally have a duration lasting at least a few seconds. To alleviate this concern, we extract features over the different instances (windows) of the data. The notion of features refers to statistical properties of the windows of data and provides some qualitative or quantitative information on them. We examined time domain and frequency domain features in the related literature and we selected a total of 105 features to compute for all the data contained in each of the windows in the dataset. Out of these 105 features, 75 were in the time domain and 30 in the frequency domain. The latter necessitated that we first apply a Fast Fourier Transform (FFT) over the windows' data and then calculate the corresponding features. Table II summarizes the selected features that were applied on the accelerometer values for each of the three axes, the magnitude of acceleration and the tilt on all three axes.

TABLE II
TIME AND FREQUENCY DOMAIN FEATURES FOR CLASSIFICATION

| Time domain | Frequency domain |
|---|---|
| Mean | Mean |
| Median | Median |
| Minimum | Minimum |
| Maximum | Maximum |
| Harmonic mean | Spectral energy |
| Geometric mean | Entropy |
| Pearson's correlation | Pearson's correlation |
| Variance | |
| Standard deviation | |
| Root mean square | |
| Covariance | |

Cyclic patterns of human activities can be easier observed using frequency domain features and this is the main reason for their significance. Indicatively, Figure 5 depicts the mean FFT for the 3 axes of acceleration; cyclic, identifiable patterns can be clearly seen in the activities involving stairs as expected.

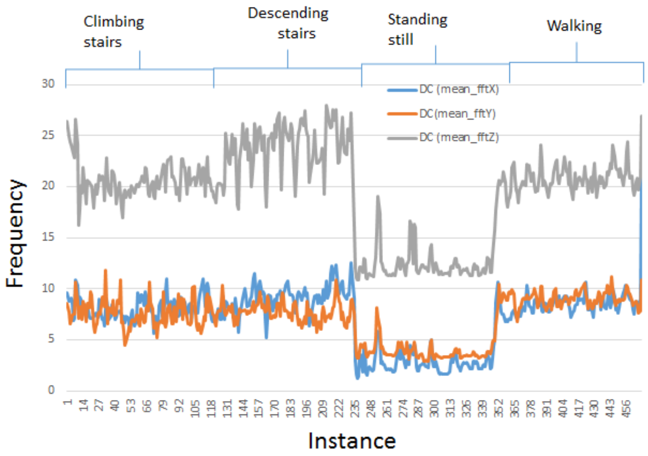


Fig. 5. Mean FFT of acceleration in the 3 axes and the activities that it corresponds to.

### E. Data classification

Having completed the aforementioned procedures and in accordance to our generic framework, the next step in the process involves training the classifier in order for it to be able to correctly analyze the collected data, but also to be able to respond to prospective queries on how new data should be classified, i.e. annotated, in line with previous data. We used the Weka toolkit [33] to experiment with a variety of machine learning algorithms, namely KStar, C4.5 decision tree, Bayes (Bayes Network with K2 search algorithm), Support Vector Machines (SVM), Sequential Minimal Optimization (SMO), k-NN, MultiClass (meta-classifier) and MultiLayer Perceptron Neural Network(MLP).

We indicatively present in the following results on the training of the classifiers based on training data coming from a Sony Ericsson phone, where the data collection frequency was set to be 500, the window size 256 with 50% overlap, the user conducted all 4 activities and she did so holding the phone in her hand at the waist level. In order to test the accuracy of the various classifiers, we used Weka and performed a comparative analysis of the aforementioned 8 algorithms using cross-validation with 10 folds and 10 iterations using all the training data. The annotated test data were used for the evaluation and that is where the corresponding results refer to. To gain a level of statistical confidence in the results obtained by Weka, we applied the well-known statistical hypothesis test Students' T-Test, with a requested confidence of 0.05 (indicates statistical difference threshold when performing pairwise comparison between schemes), and managed to acquire measurements for a variety of aspects regarding accuracy and the performance of the classifiers accordingly. Results regarding these classifiers are presented in Table III and discussed in the following.

All classifiers exhibited very high accuracy rates, reaching up to 99.5%. It becomes therefore evident that the use of just the accelerometer in detecting and distinguishing between human activities yields very promising results. Additionally, it exposes the privacy risks involved in participatory sensing, since it illustrates the level of knowledge that can be gained with the use of a non-intrusive sensor on board a smartphone.

### F. Analysis of results

Table III clearly highlights the optimal performance of the SVM classifier over all others, with the C4.5 one a close second. SVM has an accuracy level of 99.5%, whereas C4.5 98.8% and the worst performing classifiers are the Bayes Network and KStar with accuracies of 85% and 86.6%, respectively. It is interesting to examine this observation in light of the mean absolute and root mean squared error metrics. A simplistic analysis would expect the Bayes Network and KStar classifiers to have the largest degrees of errors, due to their low accuracy. However, while Bayes Networks and KStar do not indeed fare well, it is actually the SMO and MLP classifiers that exhibit the largest mean errors and root mean squared errors despite high accuracy levels of 97.01% and 97% respectively. The reason for this is the fact that while the latter classifiers managed to correctly classify a larger number of instances, the misclassifications were so big that they succeeded in significantly increasing the mean errors. This observation is consistent with our belief that evaluation of a machine learning classifier is not a simple process of computing a couple of metrics, but rather an extensive procedure where a series of quantitative metrics should be taken into account and considered in parallel, while additionally one should not neglect qualitative analysis, as discussed later.

Two very important metrics in assessing classifiers are the precision and the recall (borrowed from the field of Information Retrieval - IR). Precision measures successful assignments to a class over all assignments to that class (including incorrect ones) and in this respect it refers to the

TABLE III
TIME AND FREQUENCY DOMAIN FEATURES FOR CLASSIFICATION

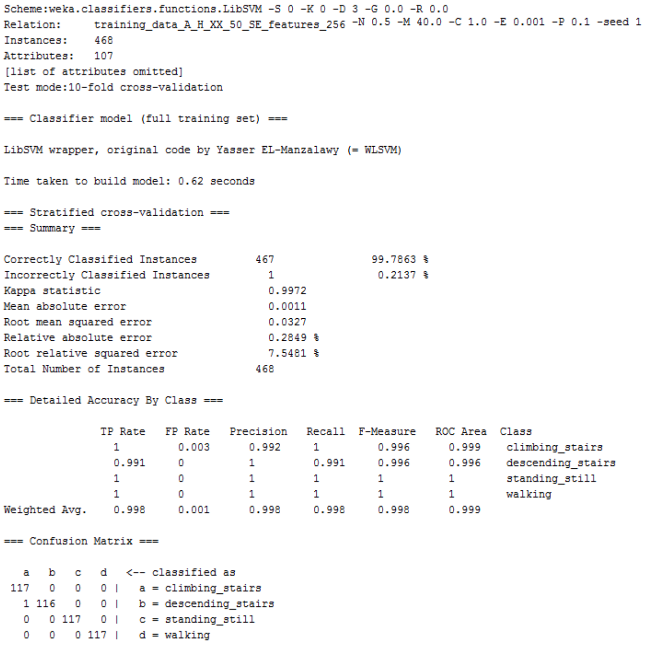
| **Classifier** | **C4.5** | **Bayes** | **KStar** | **SVM** | **SMO** | **kNN** | **MLP** | **MultiClass** |
|---|---|---|---|---|---|---|---|---|
| **Evaluation metric** | | | | | | | | |
| **Accuracy** | 98.8054 | 85.0457 | 86.5528 | 99.4976 | 97.1061 | 91.2616 | 97.0416 | 96.8560 |
| **Incorrect classifications** | 1.1946% | 14.9543% | 13.4472% | 0.5024% | 2.8939% | 8.7384% | 2.9584% | 3.1440% |
| **Mean absolute error** | 0.0060 | 0.0748 | 0.0659 | 0.0025 | 0.2529 | 0.0673 | 0.3044 | 0.0208 |
| **Root mean squared error** | 0.0638 | 0.2642 | 0.2472 | 0.0345 | 0.3159 | 0.1814 | 0.3523 | 0.1164 |
| **IR Precision** | 1.000 | 0.7704 | 0.7720 | 0.9976 | 0.9516 | 0.8294 | 0.9574 | 0.9842 |
| **IR Recall** | 0.9850 | 0.7762 | 0.7814 | 0.9974 | 0.9799 | 0.8767 | 0.9698 | 0.9497 |
| **F-measure** | 0.9923 | 0.7715 | 0.7745 | 0.9975 | 0.9652 | 0.8501 | 0.9625 | 0.9659 |
| **Area under ROC** | 0.9925 | 0.9461 | 0.9585 | 0.9983 | 0.9879 | 0.9741 | 0.9999 | 0.9942 |
| **KB mean information** | 1.9712 | 1.6417 | 1.6807 | 1.9879 | 0.9803 | 1.7244 | 0.6418 | 1.9052 |
| **Kappa statistic** | 0.9841 | 0.8006 | 0.8207 | 0.9933 | 0.9614 | 0.8835 | 0.9606 | 0.9581 |
| **Elapsed time training** | 0.0902s | 0.1299s | 0.0004s | 1.1515s | 0.1447s | 0.0006s | 176.3556s | 0.4553s |
| **Elapsed time testing** | 0.0069s | 0.0244s | 10.3873s | 0.1052s | 0.0079s | 0.1009s | 0.1494s | 0.0578s |

fraction of classified instances that are relevant, i.e. correct. Conversely, recall refers to the fraction of relevant instances that have been classified. Therefore, high recall values indicate that the classifier was successful in classifying correctly most of the instances, whereas high precision means that the classifier performed more correct classifications than incorrect ones. In our experiments, it is the C4.5 classifier that has the best precision with a value of 1.0, followed by the SVM with a value of 0.99, while the Bayes Network and KStar classifiers have the lowest precision with value of 0.77 for both of them. In general, precision follows the same pattern as accuracy, which was to be expected since these two metrics are conceptually close. Similar results (top two algorithms being SVM and C4.5 and lower two Bayes Network and KStar) can be seen for the recall metric, albeit with more distinguish values. Interestingly enough, while the MultiClass classifier had the third best precision at 0.98, it is the SMO classifier that has the third best recall at 0.98. This indicates a different performance between these two, where MultiClass is better at classifying more instances correctly than incorrectly and SMO outperforms MultiClass in performing more correct classifications. Nevertheless, SMO had a much higher root mean squared error than MultiClass, so in general one can expect better performance of the latter.

Another very important metric is the F-measure that combines both precision and recall (also known as F1 score or F1 measure since precision and recall are evenly weighted). It is actually the harmonic mean of precision and recall and is used to express the accuracy of the classification process and is widely considered to be more useful than the percent of correct classifications as expressed by the accuracy metric. According to the F-measure, the SVM and C4.5 perform the best, while the Bayes Network and KStar the worst. Moreover, the area under ROC (receiver operating characteristic) metric has also proven to be extremely useful in evaluating classification algorithms [34], although its value has been recently heavily criticized and thus undermined, e.g. in [35]. The area under the ROC curve is equal to 1 for a perfect classification and drops as classification quality drops. In our experiments MLP performs the best in terms of the area under ROC with a value very close to 1, followed by SVM and MultiClass, while Bayes Networks and KStar perform the worst. It has to be clarified nonetheless that even the worst performing classifier, i.e. Bayes Network, has a value equal to 0.95 that broadly speaking is very good for classification purposes.

Other metrics that we considered in order to compare the performance of the considered classifiers include the Kappa statistic and the KB mean information. Kappa statistic is used to indicate the agreement of prediction compared to the ground truth and is important since it is a probabilistic value that takes into account not only the comparison to the ground truth, but also the probability that a correct assignment to a class was by chance. As before the SVM and C4.5 classifiers were the ones with higher Kappa statistic value at 0.99 and 0.98 respectively (the higher the value, the better matching the agreement), while KStar and Bayes Network had the lowest values, 0.82 and 0.8 respectively. Kononenko and Bratko [36] introduced an information-based evaluation criterion for each classifier's performance, which excludes prior class probabilities and thus assesses better the performance of the classifier under uncertain conditions. Once again SVM and C4.5 had the highest performance in regards to this metric, but surprisingly MLP did not perform well enough. In our view, the reason is based on the construction of the MLP construction network that inherently requires knowledge of prior class probabilities (back-propagation error correction is at its core), so when these probabilities are excluded its performance is bound to be reduced.

The last metric that we considered was the aspect of time. In particular, we examined the time required to train the classifiers and the time required for them to perform classifications over the test data. Since the classifiers were trained and tested on the same sets of data the values obtained for these metrics are directly comparable. MLP, which is the most complex of the considered classifiers, requires the most time for training, averaging 176.35 seconds. This is however not reflected in the testing phase that only takes 0.15 seconds. The fastest classifier to train is KStar followed closely by k-NN, whereas apart from the MLP classifier, the SVM one at 1.15 seconds is also relatively slow to train. Conversely, when it comes to testing times SVM is quite fast at 0.1 seconds but the fastest ones are C4.5 and SMO. The slowest classifier for testing was KStar with the remaining algorithms exhibiting small variance in their values.

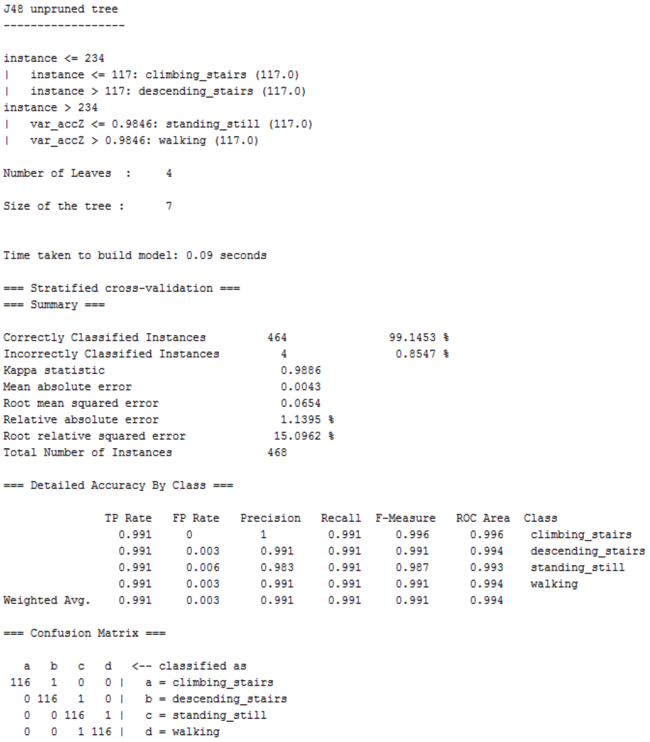


```
Scheme:weka.classifiers.functions.LibSVM -S 0 -K 0 -D 3 -G 0.0 -R 0.0
Relation:     training_data_A_H_XX_50_SE_features_256 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1
Instances:    468
Attributes:   107
[list of attributes omitted]
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 0.62 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         467               99.7863 %
Incorrectly Classified Instances         1                0.2137 %
Kappa statistic                          0.9972
Mean absolute error                      0.0011
Root mean squared error                  0.0327
Relative absolute error                  0.2849 %
Root relative squared error              7.5481 %
Total Number of Instances              468

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.003     0.992     1         0.996      0.999    climbing_stairs
                 0.991     0         1         0.991     0.996      0.996    descending_stairs
                 1         0         1         1         1          1        standing_still
                 1         0         1         1         1          1        walking
Weighted Avg.    0.998     0.001     0.998     0.998     0.998      0.999

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 117   0   0   0 |   a = climbing_stairs
   1 116   0   0 |   b = descending_stairs
   0   0 117   0 |   c = standing_still
   0   0   0 117 |   d = walking
```

Fig. 6. Confusion matrix for the SVM classifier.

It is clear that there are tradeoffs to be considered when choosing the optimal classifier for participatory surveillance needs such as identifying human activities. We need to consider accuracy, precision, recall, as well as the overhead in terms of time for each of the classifiers since they will need to be considered in real time operation. The training phase will only occur once, so long timespans for this phase can be sidestepped, but long times for testing can be used to exclude certain classifiers from our candidates' list. Evidently, quantitative results as those previously presented are important, since they provide a thorough evaluation of the performance of the different classifiers in regards to a variety of aspects. It is however equally important to be able to qualitatively analyze the classification process and in particular to be able to analyze why classification errors occur. The best way to do this is by checking the confusion matrix (also known as contingency table) of the classification process that represent the classification results versus the ground truth. Indicatively, Figure 6 shows the confusion matrix for the SVM classifier and Figure 7 for the C4.5 one. Confusion matrices are important because they allow us to diagnose which classes were confused to each other and therefore be able to draw conclusions as to why this occurred in the first place.

Based on the aforementioned extensive analysis and evaluation of the considered classifiers we came to the conclusion that the most suitable ones for participatory surveillance needs include the SVM and C4.5 one. They exhibited the optimal balance between performance, accuracy and quality of classification.



```
J48 unpruned tree
------------------

instance <= 234
|   instance <= 117: climbing_stairs (117.0)
|   instance > 117: descending_stairs (117.0)
instance > 234
|   var_accZ <= 0.9846: standing_still (117.0)
|   var_accZ > 0.9846: walking (117.0)

Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         464               99.1453 %
Incorrectly Classified Instances         4                0.8547 %
Kappa statistic                          0.9886
Mean absolute error                      0.0043
Root mean squared error                  0.0654
Relative absolute error                  1.1395 %
Root relative squared error             15.0962 %
Total Number of Instances              468

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.991     0         1         0.991     0.996      0.996    climbing_stairs
                 0.991     0.003     0.991     0.991     0.991      0.994    descending_stairs
                 0.991     0.006     0.983     0.991     0.987      0.993    standing_still
                 0.991     0.003     0.991     0.991     0.991      0.994    walking
Weighted Avg.    0.991     0.003     0.991     0.991     0.991      0.994

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 116   1   0   0 |   a = climbing_stairs
   0 116   1   0 |   b = descending_stairs
   0   0 116   1 |   c = standing_still
   0   0   1 116 |   d = walking
```

Fig. 7. Confusion matrix for the C4.5 classifier.

## VII. Conclusion

In this paper we presented our work on designing and developing a solution for participatory surveillance. We aim at involving end users in the tasks related to security and surveillance and thus on one hand assist and promote the overall perceived level of safety, while on the other hand promoting users' sense of contribution and participation in the society and hence their awareness. By utilizing the numerous sensors on smartphones that are nowadays ubiquitous we postulate that significant information regarding critical, security-related events can be inferred. As a proof of concept, we built a system to collect such data from users in the context of an emergency evacuation exercise and we presented here relevant results on the use of this data. By using just one sensor, namely the accelerometer, very high levels of accuracy in predicting users' activities were reached. In our view, this validates the great potential that exists in the field of participatory surveillance, in particular for the management of emergency/crisis events. Even with quite a few limitations that we encountered in our study, e.g. sensors accuracy or user participation, and with the limited amount of collected data, intelligence on the different activities of users was deducible.

Based on the collected results and our analysis, we are confident that with the integration of additional sensors, as well as with the collection of a far more detailed reference dataset, we would definitely be able to discern between distinct human activities at a much greater level of detail and with quantifiable assessment metrics. These aspects are among the ones we plan to investigate further in the future. To test the feasibility of our machine learning approach on participatory surveillance data we did not check against all possible testing conditions; this extensive sensitivity analysis is nonetheless the focus of our ongoing work. We are also planning work on examining the potential benefits that might arise from exploiting additional sensors such as the magnetometer, gyroscope, etc.

The usefulness of participatory surveillance is extremely high if one considers the fact that such a framework could for example allow groups of rescuers to gain access to information about the current and ongoing status and activities of people inside a building, e.g. static user for a long time or user suffering a physical shock. The analysis of the results and the possibility of detecting with high accuracy the class of previously unclassified data has highlighted the great potential of participatory surveillance systems. However, it has also exposed the great privacy risks regarding users sharing data from their smartphones from such systems. We believe that the use of additional sensors and the information fusion emerging from the use of multiple sensors will exacerbate these privacy risks and allow for more accurate detection of the users' activities, as well as the context of her surroundings. Typical examples of such risks reported in the literature include the possibility to infer the PIN of users on smartphones or the password that they type using accelerometers and gyroscopes [37]. We therefore also plan to examine the risks involved from the potential sharing of data from a variety of sensors and not only the accelerometer.

## References

[1] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," in *In: Workshop on World-Sensor-Web (WSWd'06): Mobile Device Centric Sensor Networks and Applications*, 2006, pp. 117–134.

[2] D. Estrin, "Participatory sensing: applications and architecture [internet predictions]," *Internet Computing, IEEE*, vol. 14, no. 1, pp. 12–42, 2010. doi: 10.1109/MIC.2010.12

[3] K. Shilton, "Participatory sensing: Building empowering surveillance," *Surveillance & Society*, vol. 8, no. 2, pp. 131–150, 2010.

[4] Z. Dong, B. Lu, L. He, P. Cheng, Y. Gu, and L. Fang, "Exploring smartphone-based participatory computing to improve pervasive surveillance," in *11th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '13. ACM, 2013. doi: 10.1145/2517351.2517388. ISBN 978-1-4503-2027-6 pp. 69:1–69:2. [Online]. Available: http://doi.acm.org/10.1145/2517351.2517388

[5] F. Coudert, M. Gemo, L. Beslay, and F. Andritsos, "Pervasive monitoring: Appreciating citizen's surveillance as digital evidence in legal proceedings," in *Imaging for Crime Detection and Prevention 2011 (ICDP 2011), 4th Intl Conference on*, 2011. doi: 10.1049/ic.2011.0130 pp. 1–6.

[6] N. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell, "A survey of mobile phone sensing," *Communications Magazine, IEEE*, vol. 48, no. 9, pp. 140–150, 2010. doi: 10.1109/MCOM.2010.5560598

[7] L. Bao and S. Intille, "Activity recognition from user-annotated acceleration data," in *Pervasive Computing*, ser. LNCS, A. Ferscha and F. Mattern, Eds. Springer, 2004, vol. 3001, pp. 1–17. ISBN 978-3-540-21835-7. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-24646-6_1

[8] T. Huynh and B. Schiele, "Analyzing features for activity recognition," in *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies*, ser. sOc-EUSAI '05. New York, NY, USA: ACM, 2005. doi: 10.1145/1107548.1107591. ISBN 1-59593-304-2 pp. 159–163. [Online]. Available: http://doi.acm.org/10.1145/1107548.1107591

[9] T. Brezmes, J.-L. Gorricho, and J. Cotrina, "Activity recognition from accelerometer data on a mobile phone," in *Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part II: Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, ser. IWANN '09. Berlin, Heidelberg: Springer-Verlag, 2009. doi: 10.1007/978-3-642-02481-8_120. ISBN 978-3-642-02480-1 pp. 796–799. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-02481-8_120

[10] M.-R. Ra, B. Liu, T. F. La Porta, and R. Govindan, "Medusa: A programming framework for crowd-sensing applications," in *10th Intl Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '12. ACM, 2012. doi: 10.1145/2307636.2307668. ISBN 978-1-4503-1301-8 pp. 337–350. [Online]. Available: http://doi.acm.org/10.1145/2307636.2307668

[11] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda, "Peir, the personal environmental impact report, as a platform for participatory sensing systems research," in *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '09. New York, NY, USA: ACM, 2009. doi: 10.1145/1555816.1555823. ISBN 978-1-60558-566-6 pp. 55–68. [Online]. Available: http://doi.acm.org/10.1145/1555816.1555823

[12] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher, "Greengps: A participatory sensing fuel-efficient maps application," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '10. New York, NY, USA: ACM, 2010. doi: 10.1145/1814433.1814450. ISBN 978-1-60558-985-5 pp. 151–164. [Online]. Available: http://doi.acm.org/10.1145/1814433.1814450

[13] M. Wisniewski, G. Demartini, A. Malatras, and P. Cudré-Mauroux, "Noizcrowd: A crowd-based data gathering and management system for noise level data," in *Mobile Web Information Systems*, ser. LNCS, F. Daniel, G. Papadopoulos, and P. Thiran, Eds. Springer, 2013, vol. 8093, pp. 172–186. ISBN 978-3-642-40275-3

[14] K. Shilton, "Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection," *Comm. of the ACM*, vol. 52, no. 11, pp. 48–53, Nov. 2009. doi: 10.1145/1592761.1592778. [Online]. Available: http://doi.acm.org/10.1145/1592761.1592778

[15] R. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," *Communications Magazine, IEEE*, vol. 49, no. 11, pp. 32–39, 2011. doi: 10.1109/MCOM.2011.6069707

[16] H. Keval and M. A. Sasse, "To catch a thief – you need at least 8 frames per second: The impact of frame rates on user performance in a cctv detection task," in *Proceedings of the 16th ACM International Conference on Multimedia*, ser. MM '08. New York, NY, USA: ACM, 2008. doi: 10.1145/1459359.1459527. ISBN 978-1-60558-303-7 pp. 941–944. [Online]. Available: http://doi.acm.org/10.1145/1459359.1459527

[17] A. Ito, A. Aiba, A. Ito, and S. Makino, "Detection of abnormal sound using multi-stage gmm for surveillance microphone," in *Information Assurance and Security, 2009. IAS '09. Fifth International Conference on*, vol. 1, Aug 2009. doi: 10.1109/IAS.2009.160 pp. 733–736.

[18] J. A. Hanson, K. L. McLaughlin, and T. J. Sereno, "A flexible data fusion architecture for persistent surveillance using ultra-low-power wireless sensor networks," pp. 80470M–80470M–12, 2011. [Online]. Available: http://dx.doi.org/10.1117/12.883280

[19] T. He, S. Krishnamurthy, J. A. Stankovic, T. Abdelzaher, L. Luo, R. Stoleru, T. Yan, L. Gu, J. Hui, and B. Krogh, "Energy-efficient surveillance system using wireless sensor networks," in *Proceedings of the 2Nd International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '04. New York, NY, USA: ACM, 2004. doi: 10.1145/990064.990096. ISBN 1-58113-793-1 pp. 270–283. [Online]. Available: http://doi.acm.org/10.1145/990064.990096

[20] T. Monahan and J. T. Mokos, "Crowdsourcing urban surveillance: The development of homeland security markets for environmental sensor networks," *Geoforum*, vol. 49, no. 0, pp. 279 – 288, 2013. doi: http://dx.doi.org/10.1016/j.geoforum.2013.02.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0016718513000341

[21] S. Reddy, D. Estrin, M. Hansen, and M. Srivastava, "Examining micro-payments for participatory sensing data collections," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, ser. Ubicomp '10. New York, NY, USA: ACM, 2010. doi: 10.1145/1864349.1864355. ISBN 978-1-60558-843-8 pp. 33–36. [Online]. Available: http://doi.acm.org/10.1145/1864349.1864355

[22] S. Reddy, D. Estrin, and M. Srivastava, “Recruitment framework for participatory sensing data collections,” in *Pervasive Computing*, ser. Lecture Notes in Computer Science, P. Floréen, A. Krüger, and M. Spasojevic, Eds. Springer Berlin Heidelberg, 2010, vol. 6030, pp. 138–155. ISBN 978-3-642-12653-6. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12654-3_9

[23] I. Martí, L. Rodríguez, M. Benedito, S. Trilles, A. Beltrán, L. Díaz, and J. Huerta, “Mobile application for noise pollution monitoring through gamification techniques,” in *Entertainment Computing - ICEC 2012*, ser. Lecture Notes in Computer Science, M. Herrlich, R. Malaka, and M. Masuch, Eds. Springer Berlin Heidelberg, 2012, vol. 7522, pp. 562–571. ISBN 978-3-642-33541-9. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33542-6_74

[24] D. Christin, A. Reinhardt, S. S. Kanhere, and M. Hollick, “A survey on privacy in mobile participatory sensing applications,” *Journal of Systems and Software*, vol. 84, no. 11, pp. 1928–1946, Nov. 2011. doi: 10.1016/j.jss.2011.06.073. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2011.06.073

[25] K. L. Huang, S. S. Kanhere, and W. Hu, “Preserving privacy in participatory sensing systems,” *Computer Communications*, vol. 33, no. 11, pp. 1266 – 1280, 2010. doi: http://dx.doi.org/10.1016/j.comcom.2009.08.012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140366409002448

[26] S. Sigg, M. Scholz, S. Shi, Y. Ji, and M. Beigl, “Rf-sensing of activities from non-cooperative subjects in device-free recognition systems using ambient and local signals,” *Mobile Computing, IEEE Transactions on*, vol. 13, no. 4, pp. 907–920, April 2014. doi: 10.1109/TMC.2013.28

[27] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387310738

[28] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, “Social fmri: Investigating and shaping social mechanisms in the real world,” *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 643 – 659, 2011. doi: http://dx.doi.org/10.1016/j.pmcj.2011.09.004 The Ninth Annual {IEEE} International Conference on Pervasive Computing and Communications (PerCom 2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574119211001246

[29] A. Bulling, U. Blanke, and B. Schiele, “A tutorial on human activity recognition using body-worn inertial sensors,” *ACM Comput. Surv.*, vol. 46, no. 3, pp. 33:1–33:33, Jan. 2014. doi: 10.1145/2499621. [Online]. Available: http://doi.acm.org/10.1145/2499621

[30] C. Barthold, K. Subbu, and R. Dantu, “Evaluation of gyroscope-embedded mobile phones,” in *Systems, Man, and Cybernetics (SMC), 2011 IEEE Intl Conference on*, Oct 2011. doi: 10.1109/ICSMC.2011.6083905. ISSN 1062-922X pp. 1632–1638.

[31] Z. Wu, Y. Wu, X. Hu, and M. Wu, “Calibration of three-axis magnetometer using stretching particle swarm optimization algorithm,” *Instrumentation and Measurement, IEEE Transactions on*, vol. 62, no. 2, pp. 281–292, Feb 2013. doi: 10.1109/TIM.2012.2214951

[32] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, D. Howard, K. Meijer, and R. Crompton, “Activity identification using body-mounted sensorsŮa review of classification techniques,” *Physiological Measurement*, vol. 30, no. 4, p. R1, 2009. [Online]. Available: http://stacks.iop.org/0967-3334/30/i=4/a=R01

[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: An update,” *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. doi: 10.1145/1656274.1656278. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[34] K. A. Spackman, “Signal detection theory: Valuable tools for evaluating inductive learning,” in *Proceedings of the Sixth International Workshop on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989. ISBN 1-55860-036-1 pp. 160–163. [Online]. Available: http://dl.acm.org/citation.cfm?id=102118.102172

[35] D. J. Hand, “Measuring classifier performance: A coherent alternative to the area under the roc curve,” *Mach. Learn.*, vol. 77, no. 1, pp. 103–123, Oct. 2009. doi: 10.1007/s10994-009-5119-5. [Online]. Available: http://dx.doi.org/10.1007/s10994-009-5119-5

[36] I. Kononenko and I. Bratko, “Information-based evaluation criterion for classifier’s performance,” *Machine Learning*, vol. 6, no. 1, pp. 67–80, Jan. 1991. doi: 10.1023/A:1022642017308. [Online]. Available: http://dx.doi.org/10.1023/A:1022642017308

[37] Z. Xu, K. Bai, and S. Zhu, “Taplogger: Inferring user inputs on smartphone touchscreens using on-board motion sensors,” in *5th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, ser. WISEC ’12. ACM, 2012. doi: 10.1145/2185448.2185465. ISBN 978-1-4503-1265-3 pp. 113–124. [Online]. Available: http://doi.acm.org/10.1145/2185448.2185465

# Frontiers in Network Applications, Network Systems and Web Services

SYMPOSIUM SoFAST-WS focuses on modern challenges and solutions in network systems, applications and service computing.  The Symposium builds upon the success of Frontiers in Network Applications and Network Systems (FINANS'2012) and 4th International Symposium on Web Services (WSS' 2012) held in 2012 in Wroclaw, Poland. These two events are now integrated into one event to fully exploit the synergy of topics and cooperation of research groups.

The topics discussed during the symposium include different aspects of network systems, applications and service computing. The primary objective of the symposium is to bring together researchers and practitioners analyzing, developing and administering network systems, with particular emphasis on Internet systems. Authors are invited to submit their papers in English, presenting the results of original research or innovative practical applications in the field.

### Topics

Topics include (but are not limited to):

-
- Architecture, scalability and security of Open API solutions,
- Technical and social aspects of Open API and open data,
- Service delivery platforms - architecture and applications,
- Telecommunication operators API exposition in Telco 2.0 model,
- The applications of intelligent techniques in network systems,
- Mobile applications,
- Network-based computing systems,
- Network and mobile GIS platforms and applications,
- Computer forensic,
- Network security,
- Anomaly and intrusion detection,
- Traffic classification algorithms and techniques,
- Network traffic engineering,
- High-speed network traffic processing,
- Heterogeneous cellular networks,
- Wireless communications,
- Security issues in Cloud Computing,
- Network aspects of Cloud Computing,
- Control of networks,
- Standards for Web services,
- Semantic Web services,
- Context-aware Web services,
- Composition approaches for Web services,
- Security of Web services,
- Software agents for Web services composition,
- Supporting SWS Deployment,
- Architectures for SWS Deployment,
- Applications of SWS to E-business and E-government,
- Supporting Enterprise Application Integration with SWS,
- SWS Conversational Protocols and Choreography,
- Ontologies and Languages for Service Description,
- Ontologies and Languages for Process Modeling,
- Foundations of Reasoning about Services and/or Processes,
- Composition of Semantic Web Services,
- Innovative network applications, systems and services.

### Event Chairs

**Furtak, Janusz,** Military University of Technology, Poland

**Grzenda, Maciej,** Orange Labs Poland and Warsaw University of Technology, Poland

**Legierski, Jarosław,** Orange Labs Poland, Poland

**Luckner, Marcin,** Warsaw University of Technology, Poland

**Szmit, Maciej,** Orange Labs Poland, Poland

### Program Committee

**Benslimane, Sidi Mohammed,** University of Sidi Bel-Abbès, Algeria

**Chojnacki, Andrzej,** Military University of Technology, Poland

**Cocucci, Osvaldo,** Orange Labs Products & Services, France

**Fernández, Alberto,** Universidad Rey Juan Carlos, Spain

**García-Domínguez,** Antonio, University of York, United Kingdom

**Gibert, Philippe,** Orange Labs Products and Services, France

**Kaczmarski, Krzysztof,** Warsaw University of Technology, Poland

**Katakis, Ioannis,** National and Kapodistrian University of Athens, Greece

**Kiedrowicz, Maciej,** Military University of Technology, Poland

**Korbel, Piotr,** Lodz University of Technology, Poland

**Kowalczyk, Emil,** Orange Labs, Poland

**Kowalski, Andrzej,** Orange Labs, Poland

**López Nores, Martín,** University of Vigo, Spain

**Maamar, Zakaria,** Zayed University, United Arab Emirates

**Macukow, Bohdan,** Warsaw University of Technology, Poland

**Misztal, Michal,** Military University of Technology, Poland

**Nowicki, Tadeusz,** Military University of Technology, Poland

**Richomme, Morgan,** Orange Labs, France

**Wrona, Konrad,** NATO Consultation, Netherlands

**Zieliński, Zbigniew,** Military University of Technology, Poland

**Żorski, Witold,** Military University of Technology, Poland

DOI: 10.15439/2015F284
ACSIS, Vol. 5

# Measurement methodology of TCP performance bottlenecks

Andrzej Bąk and Piotr Gajowniczek
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
email: bak@tele.pw.edu.pl

Michał Zagożdżon
Orange Labs
Orange Polska S.A.
Obrzeżna 7, 02-679 Warsaw, Poland
email: michal.zagozdzon@orange.com

***Abstract*—Transmission Control Protocol (TCP) is still used by vast majority of Internet applications. However, the huge increase in bandwidth availability and consumption during the last decade has stimulated the evolution of TCP and introduction of new versions that are more suited for high speed networks. Many factors can influence the performance of TCP protocol, starting from scarcity of network resources, through client or server misconfiguration, to internal limitations of applications. Proper identification of the TCP performance bottlenecks is therefore an important challenge for network operators. In the paper we proposed the methodology for finding root causes of througput degradation in TCP connections based on passive measurements. This methodology was verified by experiments conducted in a live network with 4G wireless Internet access.**

## I. Introduction

Since the foundation of the Internet the vast majority of network data is transmitted using Transmission Control Protocol (TCP). TCP underlies many ‘traditional’ Internet applications such as web browsing, email, bulk data transfer etc., but also the relatively new ones, such as HTTP adaptive streaming that is quickly becoming the preferred method for Over-The-Top video delivery. All this makes the TCP performance analysis one of the most important research areas of the Internet networking. Since the beginning of TCP’s public use in 1989 a lot of research effort was devoted to improve its performance, and the protocol itself has evolved significantly.

The early TCP version used RTO (Retransmission Time-Out) timer to recover from packet loss which was inefficient even on low speed links. The TCP Reno/NewReno version [1] introduced fast retransmit & recovery mechanism that improved the TCP performance in presence of packet loss. For a long time the TCP Reno was a de-facto standard widely deployed in the Internet. However, the significant increase in network capacity observed during the last decade has stimulated introduction of new TCP congestion control algorithms that are more suited for high speed links, such as Fast TCP [2], BIC [3], STCP [4], CUBIC [5] [6] [7], HTCP [8] [9] [10], HSTCP [11] [12], Compound TCP [13], TCP Westwood [14] etc. The new versions of Linux operating system do even allow switching between different congestion control algorithms without the need to recompile the kernel.

This paper presents the methodology of finding the root causes of throuput degradation in TCP connections on the base of passive measurements obtained from probes capturing traffic on the network links. This methodology combines the detection of TCP source application type (greedy vs non-greedy) [15] with estimation of coefficients related to transmission effectiveness that are based on the RFC 6349 [16]. The proposed approach is supported by results obtained from measurements conducted in the live 4G mobile network of Orange Poland.

## II. Sources of TCP performance bottlenecks

TCP uses congestion and flow control mechanisms to control the transmission rate of the sender process by limiting the amount of data that can be transmitted without waiting for acknowledgement (called the *window size*). Changing transmission rate in response to receiver’s limitation in processing incoming data (*flow control*) is based on the current size of the receiver’s window ($awnd$). This value is advertised to TCP sender process in segments that are sent as acknowledgements to the received data. Too small values of the receiver’s window can however negatively affect the performance of the TCP protocol. Therefore, in newer implementations it can be adapted algorithmically depending on the characteristics of the transmission path (such as throughput and delay).

Congestion control is done by algorithms that aim to ‘sense’ the bottleneck throughput on the transmission path and adapt the transmission rate to this limit. The sender process keeps the state variable called the congestion window ($cwnd$) that works in a similar way as advertised window, except that its value is set by an algorithm running on the sender side. Usually, the sender starts with a small value of $cwnd$ and tries to increase it each time when an acknowledgement for the previously sent data segment is received. The initial phase of aggressive $cwnd$ increase is called a *slow start* - the $cwnd$ is increased by 1 segment after each acknowledgement which leads to exponential growth in the amount of transmitted data. After encountering data loss the $cwnd$ shrinks; the following increase is usually slower and TCP sender enters the phase called *congestion avoidance*. There are many different congestion control algorithms and their variants - for an excellent review see [17]. However, they all share the same purpose - to maximize the usage of capacity available on the transmission path while also minimizing the probability of data loss.

978-83-60810-66-8/$25.00©2015, IEEE

For the TCP sender process the actual window size is the minimum of the advertised receiver's window ($awnd$) and its own congestion window ($cwnd$). After sending full window of data, TCP must stop transmission and wait for acknowledgement. The acknowledgement related to the earliest outstanding segment that was transmitted will start to arrive after the RTT (Round Trip Time) between the sender and the receiver. Each arriving acknowledgment will trigger transmission of the next segment of data awaiting in the output buffer. Hence, the TCP process can send at most $min(cwnd, awnd)$ of data per round trip time cycle and the instantaneous TCP throughput can be roughly estimated as:

$$TCP_{th} = \frac{min(cwnd, awnd)}{RTT} \tag{1}$$

TCP window that is too small may severely limit the performance of TCP connection. In order to obtain high throughput, TCP must be able to fill the network pipeline with data that will keep the network busy. Therefore, the TCP sender's window size must be greater than the bandwidth delay product:

$$min(cwnd, awnd) \geq C * RTT \tag{2}$$

where $C$ denotes the capacity available for TCP connection on the tranmission path.

There are various ways to set the values of $awnd$ and $cwnd$. As it was noted earlier, especially for $cwnd$ there is a number of different congestion control algorithms that react differently to the potential congestion detected either by the Retransmission Time-Out (a timer on sender's side) or by receiving a duplicate acknowledgement (Dup-ACK) from the receiver. Both events lead to segment retransmission and $cwnd$ window scaling, and appear in result of changes in the network environment, such as increased load, change in traffic mix, change in link parameters etc.

Another factor limiting the TCP performance is related to sizing TCP socket buffers on both sides of the connection. The problem of buffers being too small is especially visible in the networks with high bandwidth delay product (the maximum buffer space for TCP sockets depends on the operating system in case of typical Internet hosts). The receiver's socket buffer size can significantly influence the performance of TCP connection (receiver's buffer can limit the sending rate of the TCP source). Therefore, proper configuration of the sockets' buffers is very important to assure high TCP throughput [18]. Modern operating systems introduce automated algorithms for tuning the TCP buffers [19] [20] [21] [22] [23].

Similar case is related to sizing the buffers of network devices [25]. TCP sender can emit data in bursts (up to the current $cwnd$ window size). If network buffers are too small, the inevitable data loss will prevent the congestion window from growing and TCP connection will not be able to ramp up the transmission rate to available capacity. It is generally advised that network buffers should be at least twice the size of the network bandwidth delay product to assure high TCP throughput.

Another factor influencing the achievable TCP throughput is related to packet reordering [26] [27] that can be introduced for example by parallel packet processing in network devices. Receiving out-of-order segments can result in duplicate acknowledgements being sent and interpreted as data loss. This may in turn lead to unnecessary retransmissions, $cwnd$ reduction and throughput degradation. On the receiver's side frequent segment reordering may lead to extensive buffering and potential reduction in receiver's window size.

Finally, the throughput limitation can lie within the application itself. For example, in adaptive HTTP streaming the client requests chunks of video file from the server with frequency related to the encoding rate, even if the available capacity would allow transmitting data faster. In this context, TCP sources can be divided into greedy (always trying to utilize the most of available transmission capacity) and non-greedy (where rate is limited by internal behavior of the source). This classification is utilized in the methodology discussed in section III-B.

## III. Detection of the root causes of TCP performance degradation

In this section we describe the algorithm for detecting the root causes of the TCP performance bottlenecks using passive TCP measurements.

### A. Network measurements

Following the recommendations from RFC 6349, it is advised to perform the MTU (Maximum Transmission Unit) discovery procedure (see [28] for reference) before starting measurements, to avoid unwanted packet fragmentation.

We assume that the TCP traffic is monitored near the sender (at the client or at the server, depending on the direction of the transmission). The monitoring point must be close enough to the sender in order to precisely estimate the RTT parameter which is required by the proposed bottleneck detection algorithm. The monitored network traffic is saved by the probes in *.pcap* format for further processing.

The throughput of the TCP connection $TCP_{th}$ can be estimated directly from data captured by passive probes as a ratio of data sent and acknowledged during a measurement period to the length $t$ of this period:

$$TCP_{th} = \frac{ACK(t)}{t} \tag{3}$$

$ACK(t)$ denotes the highest acknowlegement sequence number observed up to time $t$. It can be obtained directly from the headers of the captured TCP segments.

Due to the nature of congestion and flow control mechanisms, the TCP sender needs some time before it can reach the desired transmission speed. This time may vary from few seconds to even hours depending on the network RTT, bandwidth, TCP congestion and flow control algorithms etc. For example, during congestion avoidance phase the TCP source needs approximately 30 seconds to increase the transmssion rate by 10 Mbps if the network RTT is 200 ms. Therefore, for proper estimation of the TCP throughput the measurement time should be long enough. The following approach is suggested to assure that. Assuming some interval $\Delta t$ and threshold $c$, seek for time instant $t$ that satisfies the following condition:

$$\left| \frac{TCP_{th}(t + \Delta t) - TCP_{th}(t)}{\Delta t} \right| < c \tag{4}$$

The above formula approximates the derivative of TCP rate estimator. The measurement time should be long enough to assure that the TCP rate estimator does not significantly change over time. In the experiments presented further in this paper we assumed $\Delta t = 1s$ and $c = 100$ KB/$s^2$.

In addition to the typical traffic traces captured at measurement points, the proposed methodology requires running some additional measurements to calculate certain TCP performance indicators (described in section III-D). The first measurement is related to estimation of reference (bottleneck) bandwidth $C_{REF}$. This can be achieved by probing the network bottleneck with UDP traffic. There are many variants of this approach - for examples see [29] or [30]. In our measurements we have used the latter: a train of 50 UDP packets was sent to the receiver and the available capacity was measured simply by dividing the total length of the received UDP packets by the total reception time (under the condition of no packet loss).

To verify if the buffers are properly dimensioned, the *back to back frames* test should be also performed. This test consists of sending the specified number of UDP packets with the maximum possible rate and repeating it while increasing the number of transmitted packets in each trial. The maximum batch size that can be sent without observing packet loss is an indirect measure of buffer size on the transmission path of the stream.

### B. Categorizaton of TCP sources

TCP throughput depends on the amount of data the TCP source emits during a single RTT period, as this value is controlled by the congestion and flow control mechanism. At any time instant the amount of outstanding data (sent but not acknowledged) is limited to $min(cwnd, awnd)$. As was explained in section II the TCP source can send at most $min(cwnd, awnd)$ bytes per RTT period. Therefore, the amount of outstanding data in relation to the RTT is an indicator of instantaneous TCP performance.

If the amount of outstanding data is less than what $cwnd$ and $awnd$ parameters allow, it means that the sender is not fully exploiting the available transmission capacity. The cause may be related to internal sender faults (such as application software or hardware issues, CPU overload etc.), but more often is a result of the consent behavior of TCP source (that may not require more throughput, as it is in case of typical streaming applications where transmission speed is related to the bitrate of the video stream). In the opposite case (if the outstanding data is close to the $cwnd$ or $awnd$), the bottlenecks are introduced either by the network or by the receiver.

Summing up the above discussion, the TCP source may fall into one of the following categories:

- *Internally limited*
  Non-greedy source i.e. a TCP connection that is not fully exploiting the capacity available in the network; the amount of outstanding data is significantly lower than the $cwnd$ and $awnd$ windows would allow.
- *Receiver limited*
  TCP source whose transmission rate is limited by the receiver; the amount of outstanding data is close to the $awnd$ and also lower than the $cwnd$.
- *Network limited*
  TCP source whose transmission rate is limited by the network, i.e. by the available capacity, packet loss rate or network RTT; the amount of outstanding data is close to the $cwnd$.

In order to classify the TCP source into one of the above categories we need the following parameters: outstanding data, RTT, receiver's window size and congestion window size. The first three parameters can be easily obtained from the packet traces captured by the passive probes. However, the congestion window is not directly measurable as it is an internal parameter of the TCP stack at the sender and cannot be directly inferred from the TCP traces. In order to cope with this problem we follow the approach of [15]. The TCP connection state is emulated using the recorded TCP traces to recover the $cwnd$ parameter. We also recover the value of RTO to distinguish between retransmissions induced by the fast retransmit phase and those due to the timer expiration. This is required to precisely track the changes in the $cwnd$ parameter.

### C. Emulation of TCP connection state

The internal state of TCP congestion control mechanism is defined by three main parameters: size of congestion window ($cwnd$), threshold for switching between slow start and congestion avoidance phase ($ssthr$), and the retransmission timer (RTO). These parameters are essential for emulation of the TCP connection state.

After a 3-way handshake procedure, the TCP process enters the established state and the TCP sender starts to transmit data. The sender sets its $cwnd$ parameter to some initial value and begins transmitting in the slow start mode. While in slow start, TCP adds one segment to the $cwnd$ for each acknowledged segment, doubling its $cwnd$ with every RTT period. Therefore, during slow start TCP throughput grows exponentially. The aim of this phase is to quickly probe the network capacity and to estimate the optimum window size without heavily overloading the network.

Slow start phase ends when either the $ssthr$ is reached or the segment loss is detected. TCP detects segment loss by two mechanisms: expiration of RTO timer or reception of duplicate acknowledgements (Dup-ACKs). In the first case the TCP sender retransmits all outstanding data and enters the slow start mode again. In the second case, after receiving 3 consecutive Dup-ACKs the TCP sender enters the recovery phase and employs fast retransmit & recovery mechanism to recover the lost segment. In contrary to the RTO mechanism, only one segment is retransmitted. The assumption behind this approach is that in this case only one segment is most likely lost and there is no need to follow the go-back-N protocol and retransmit all outstanding data.

The sender sets the $ssthr$ to the half of the $cwnd$ window before the segment loss, sets the $cwnd$ to $ssthr$+3 segments and retransmits the segment pointed by Dup-ACKs. Each time another Dup-ACK arrives, the sender adds one segment to the $cwnd$ (inflating the congestion window). The aim of this is to sustain the TCP throughput (as Dup-ACK indicates that the network is still able to deliver packets).

In the recovery phase, the TCP sender is allowed to transmit new data as indicated by $cwnd$. The recovery phase ends when all outstanding data from the beginning of this phase is acknowledged. When leaving the recovery phase the TCP sets the $cwnd$ back to the $ssthr$ and enters the congestion avoidance mode.

In TCP Reno/NewReno versions, during congestion avoidance phase one segment is added to the $cwnd$ in each RTT period. This means that the $cwnd$ grows linearly over time, increasing TCP throughput more conservatively then in slow start phase. However, it may take long time to recover TCP throughput in the network with high bandwidth delay product. Therefore, new congestion control mechanisms introduce more aggressive approaches for increasing the $cwnd$ during congestion avoidance phase.

In order to track the sender's $cwnd$ we emulate the behavior of the TCP protocol. To obtain high accuracy of the emulation we used the original source code of the Linux kernel version 3.18 [24]. The H-TCP congestion control algorithm code was used as this protocol was employed in our test setup. The code was taken from the following TCP modules:

- `tcp_input.c` - estimation of the RTO algorithm; the following function was used:
  - $tcp\_rtt\_estimator$
- `htcp.c` - estimation of the H-TCP congestion control algorithm; the following functions were used:
  - $measure\_achieved\_throughput$
  - $htcp\_cong\_avoid$
  - $tcp\_slow\_start$
  - $htcp\_alpha\_update$
  - $htcp\_beta\_update$
  - $htcp\_recalc\_ssthresh$

Each time the data or acknowledgment segment is observed, we run an appropriate piece of the Linux kernel code. When the acknowledgement segment with higher sequence number is observed we calculate the RTT sample (the time difference between reception of acknowledgement segment and observation of data segment for the given sequence number at the monitoring point) and run the $tcp\_rtt\_estimator()$ function to update the value of the RTO timer.

Next, the $measure\_achieved\_throughput()$ and $htcp\_cong\_avoid()$ functions of the H-TCP algorithm are called to update the internal state of the congestion control algorithm. The $htcp\_cong\_avoid()$ function runs the algorithm for slow start or congestion avoidance phase (depending on whether the $cwnd$ is less or greater then $ssthr$) and updates the $cwnd$ value accordingly.

When 3 duplicate acknowledgements are observed we assume that TCP enters the recovery phase of the fast retransmit & recovery mechanism. We call the $htcp\_recalc\_ssthresh()$ to update the $ssthr$ value according to the H-TCP algorithm. Each time next DupACK segment is observed, we inflate the $cwnd$ by one segment (as specified by NewReno algorithm).

When all outstanding data at the beginning of the recovery phase is acknowledged, the code for regular acknowledgments is executed (emulating congestion avoidance phase). When out-of-sequence data packet is observed that was not acknowledged and the time since transmission of the original segment is greater than the RTO, we assume retransmission due to the timer expiration. The $cwnd$ is reset to the initial value and the slow start code is executed by the $htcp\_cong\_avoid()$ function. When the $cwnd$ exceeds $ssthr$, the $htcp\_cong\_avoid()$ function executes the H-TCP congestion avoidance code again for each observed acknowledgement segment.

To validate the implemented TCP state tracking one can compare the outstanding data calculated from measurements with estimated values of the $cwnd$ (as the amount of outstanding data can approximate the $cwnd$, especially for greedy TCP sources).

For automated detection whether the TCP connection is network, receiver or internally limited, we have implemented the algorithm proposed and described in detail in [15]. Unlike in the original algorithm, we do not however divide the TCP flow into chunks and categorize each chunk individually, but rather do the categorization for the TCP connection as a whole.

An example based on network measurements is shown in Fig. 1 and Fig. 2 for H-TCP-based source. Fig. 1 shows the comparison of outstanding data with the value of $cwnd$ estimated by emulation of the H-TCP congestion control algorithm using Linux kernel source code.

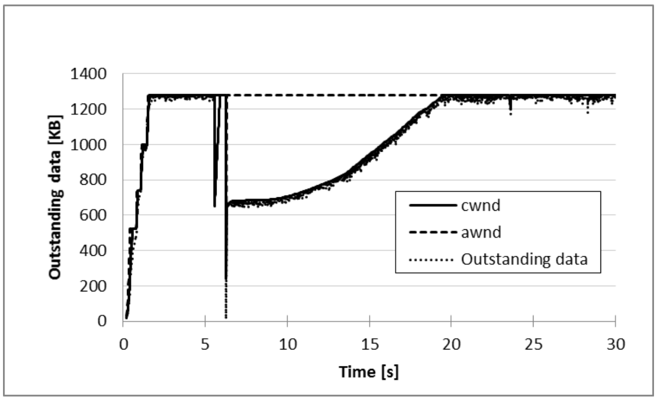


Fig. 1. H-TCP $cwnd$ emulation

The estimated $cwnd$ almost exactly matches the amount of measured outstanding data. Similar results are obtained for the estimation of RTO parameter for the same source (Fig. 2).

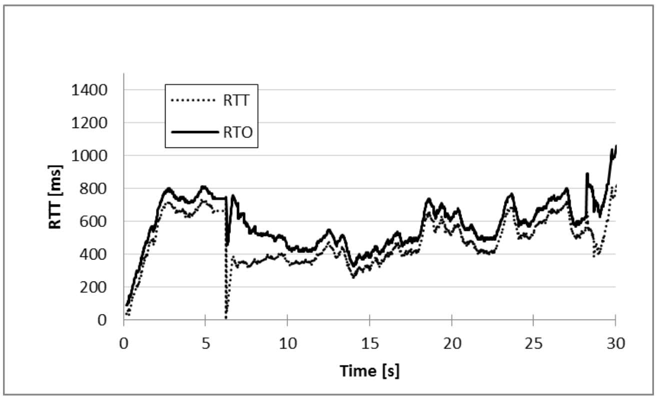


Fig. 2. H-TCP RTO estimation

The following figures present analogous results for the TCP Reno/NewReno based source. The accuracy of $cwnd$ emulation is shown in Fig. 3, RTO estimation in Fig. 4.

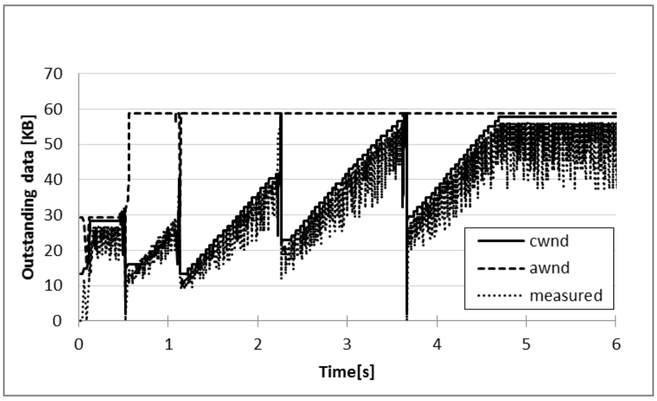


Fig. 3. TCP Reno $cwnd$ emulation

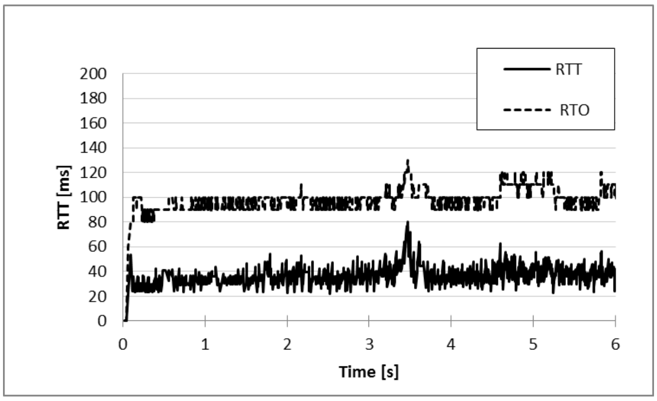


Fig. 4. TCP Reno RTO estimation

### D. TCP performance metrics

Based on RFC 6349, the following metrics are recommended to test the TCP effectiveness.

- *TCP Throughput Ratio* $W$. This metric is calculated as a percentage ratio of achieved throughput $TCP_{th}$ to the reference throughput $C_{REF}$ and should approach 100% for good connections.

$$W = \frac{TCP_{th}}{C_{REF}} * 100 \tag{5}$$

- *TCP transmission effectiveness* $E$. It is a percentage ratio of non-retransmitted data to the total amount of data sent during the measurement period and should also approach 100% for effective connections.

$$E = \frac{D - D_{RET}}{D} * 100 \tag{6}$$

$D_{RET}$ denotes the amount of data retransmitted during the measurement period.

- *Buffer Delay* $T$. To calculate this parameter one needs the reference delay $RTT_{MIN}$ calculated beforehand from measurements taken when the network load is minimal. The *tcptrace* tool can be used for this task. Alternatively, $RTT_{MIN}$ may be approximated by the minimal RTT observed during the actual measurement period. Denoting an average RTT observed within the measurement period as $RTT_{AVG}$, the Buffer Delay can be calculated as:

$$T = \frac{RTT_{AVG} - RTT_{MIN}}{RTT_{MIN}} * 100 \tag{7}$$

As the name implies, this parameter is related to buffer size in the network nodes and can be interpreted as a measure of buffer load imposed by the measured TCP connection (mostly related to the buffer at the bottleneck link). If we assume that buffer size $B$ conforms to the following formula:

$$B > 2 * C_{REF} * RTT_{MIN} \tag{8}$$

then the Buffer Delay should be greater than 200%.

### E. Root cause analysis

For the root cause analysis we use the emulation of the TCP sender state derived from passive measurements and the metrics of TCP connection performance calculated on the base of measurements. The general algorithm is depicted in Fig. 5.

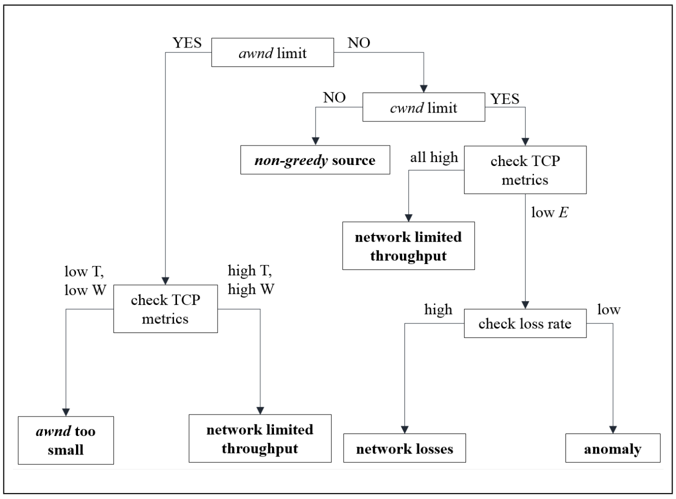


Fig. 5. General algorithm for root cause analysis

In the proposed approach the first task is to check if the throughput is not limited by $awnd$. This can be done by analyzing the behavior of the outstanding data using emulation of the TCP sender state. If it is the case, then it is advised to check the TCP connection metrics. Low $T$ (low buffering) and low $W$ (low bandwidth utilization) together with large $T$ (lack of retransmissions) support the hypothesis that the advertised $awnd$ value is indeed limiting the sender's performance. If however the $T$ and $W$ are relatively large, the true limitation may lie in the network itself and the $awnd$ value reached by the sender is large enough for high connection efectiveness.

If neither $awnd$ nor $cwnd$ (estimated from emulation of TCP sender state) is the limiting factor then the achieved throughput results from the internal sender constraints (non-greedy source). Low value of Buffer Delay may additionally support this hypothesis.

If the $cwnd$ imposes the limit on achieved throughput, it is advised to check TCP connection metrics. High effectiveness of transmission together with large Buffer Delay confirm that TCP throughput is limited by a bottleneck link in the network. However, if the level of observed retransmissions is high (low $E$), then the reason behind low throughput may lie in excessive packet loss in the network resulting e.g. from faults, bad conditions on wireless access link etc. It has to be noted that TCP retransmissions occur naturally in result of congestion control algorithm continuous attempts to fit the transmission rate to the bottleneck bandwidth, but the excessive level of retransmissions is suspicious and has to be checked further.

Finally, the case when transmission effectiveness is low but the packet loss is also low has to be treated as an anomaly that requires further investigation.

### *F. Validation of the proposed approach*

The proposed approach was validated by conducting measurements in the real network. The measurement setup is depicted in Fig. 6.

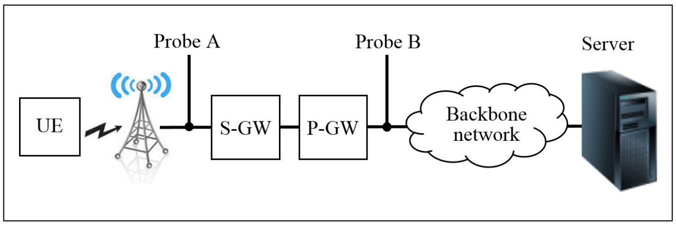


Fig. 6. Measurement setup

Measurements were done in commercial 4G mobile network of Orange Poland with real user traffic served in the background. The setup consisted of the UNIX-based web server connected to the backbone network. The TCP traffic can be monitored at the server and/or at the mobile device (with *tcpdump*). Additionally, two hardware monitoring probes were installed in the mobile access network. The monitored TCP traffic was saved in *.pcap* format for further processing.

We ran a number of tests based on downloading files from the server to the mobile device. Two types of experiments were carried out. In the first case the files were downloaded from the server in a greedy mode. The server was configured to transmit data with maximum possible rate so that the network available capacity was the only limiting factor for TCP throughput.

In the second type of the experiment the socket buffer size at the server was limited below the bandwidth delay product of the network which is approximately 100 KB (40 $ms\ RTT * 20\ Mbps\ C_{REF}$). As can be seen from Fig. 7, the tested TCP connection begins in slow start and within few seconds the $awnd$ and $cwnd$ parameters reach their maximum sizes. This is possible due to large network buffers that can accommodate thousands of packets. After the next few seconds there is a packet drop (signalled by 3 Dup-ACK segments), TCP connection retransmits the lost segment and enters the congestion avoidance phase.

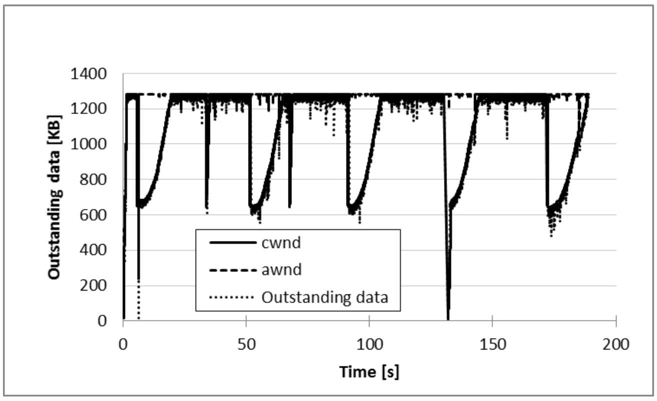


Fig. 7. Network limited TCP connection

While in congestion avoidance the $cwnd$ follows the H-TCP congestion control algorithms. At 130 sec. time instant we observe a retransmission due to the RTO expiration. TCP falls back to the slow start mode and after reaching $ssthr$ (set to 0.5 of the $cwnd$ before segment loss) it switches again to congestion avoidance phase. Notice that the estimated $cwnd$ follows the amount of measured outstanding data very accurately.

**Network limited TCP connection**

min rtt [ms] = 23.5
avg rtt [ms] = 462
avg out data [B] = 1103400
avg awnd [B] = 1275910
avg cwnd [B] = 1114465
measured throughput [Mbps] = 15.8
outstanding data/rtt [Mbps] = 19
fast retransmits = 6
RTO expirations = 5
TCP efficiency (E) [%] = 99.98
buffer delay (T) [%] = 1866

According to the proposed TCP throughput measurement methodology, the tested TCP connection is clearly network limited. The outstanding data follows the $cwnd$ closely while the TCP efficiency $E$ is high which means no excessive packet loss inside the network. The buffer delay $T$ is also very high indicating that TCP connection is transmitting a lot of data to the network (see text in the relevant frame). The network capacity $C_{REF}$, measured with the UDP protocol immediately before starting the test transfers (in the same conditions that influenced maximum throughput achievable in the location during the experiment), is about 18 Mbps. Therefore, the TCP connection in this case utilizes almost 90% of available capacity (TCP throughput ratio $W$ is also high).

In the second experiment the socket buffer of the receiver was limited to about 60 KB. In this case the outstanding data follows the $awnd$ (see Fig. 8) indicating that the TCP connection is limited by the client.

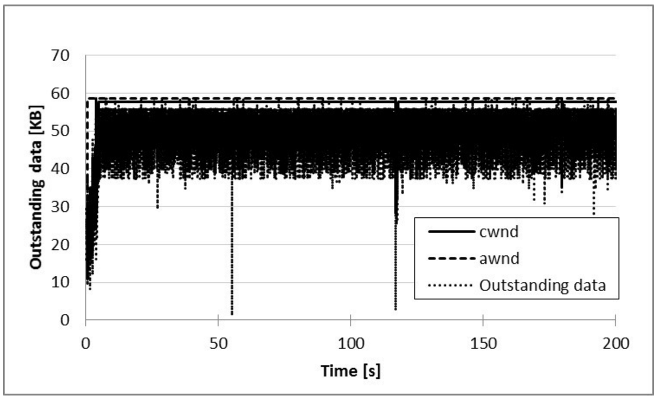


Fig. 8. Receiver limited TCP connection

This reasoning is also justified by low value of buffer delay $T$ which is now below 100%, meaning that TCP is not filling the network with data (see frame). The achieved TCP throughput is about 12.7 Mbps.

**Receiver limited TCP connection**
min rtt [ms] = 17.5
avg rtt [ms] = 32.4
avg out data [B] = 50956
avg awnd [B] = 58616
avg cwnd [B] = 57615
measured throughput [Mbps] = 12.7
outstanding data/rtt [kbps] = 12.5
fast retransmits = 5
RTO expirations = 0
TCP efficiency (E) [%] = 99.99
buffer delay (T) [%] = 85

## IV. Conclusions

The paper presents the methodology for identifying the root cause of the TCP connection performance bottlenecks. We have used the Linux kernel source code to implement the algorithm for estimation of the internal TCP connection state. Such approach allows to infer the dynamics of the TCP congestion window which is otherwise unavailable from passive TCP monitoring. The knowledge of the internal TCP state ($cwnd$, $ssthr$, RTO) is essential in understanding the observed behavior of the TCP connection and allows identifying the source of the TCP throughput limitations. In our approach it is used together with analysis of the TCP performance metrics proposed in RFC 6349. Such combined approach, complimented with additional active measurements (probing available capacity, measuring bottleneck buffers) can be helpful in tracing down network problems related to TCP-based applications.

## References

[1] T. Henderson, S. Floyd, A. Gurtov, Y. Nishida, *RFC 6582: The NewReno Modification to TCP's Fast Recovery Algorithm*

[2] D.X. Wei, Cheng Jin; S.H. Low, S. Hegde, *FAST TCP: Motivation, Architecture, Algorithms, Performance* IEEE/ACM Transactions on Networking, vol.14, no.6, pp.1246-1259, Dec. 2006, doi: 10.1109/TNET.2006.886335

[3] L. Xu, K. Harfoush, I. Rhee, *Binary increase congestion control for fast, long distance networks* Proc. of IEEE INFOCOM, vol. 4, pp. 2514–2524, March 2004

[4] T. Kelly, *Scalable TCP: improving performance in highspeed wide area networks* Computer Communications Review, vol. 32, no. 2, April 2003.

[5] H. Jamal, K. Sultan, *Performance Analysis of TCP Congestion Control Algorithms* Int. Journal of Computers and Comm., Issue 1, vol. 2, 2008

[6] S. Ha, I. Rhee, L. Xu, *CUBIC: a new TCP-friendly high-speed TCP variant* SIGOPS Oper. Syst. Rev. 42, 5 (July 2008), 64-74, doi: 10.1145/1400097.1400105

[7] D.J. Leith, R.N. Shorten, G. McCullagh, *Experimental evaluation of Cubic-TCP* Proc. of PFLDnet, 2008

[8] G. Armitage, L. Stewart, M. Welzl, J. Healy, *An Independent H-TCP Implementation under FreeBSD 7.0 – Description and Observed Behaviour* ACM SIGCOMM Computer Communication Review, vol. 38, no. 3, July 2008

[9] D. Leith, R. Shorten, *H-TCP: TCP for high-speed and long-distance networks* Proc. of PFLDnet, 2004

[10] D.J. Leith, R.N. Shorten, Y. Lee, *H-TCP: A framework for congestion control in high-speed and long-distance networks* Proc. of PFLDnet, 2005

[11] S. Floyd, *RFC 3649: HighSpeed TCP for large congestion windows*

[12] S. Floyd, *RFC 3742: Limited Slow-Start for TCP with Large Congestion Windows*

[13] K. Tan, J. Song, Q. Zhang, M. Sridharan, *A compound TCP approach for high-speed and long distance networks* Proc. of INFOCOM 2006, pp.1-12, 2006, doi: 10.1109/INFOCOM.2006.188

[14] S. Mascolo, C. Casetti, M. Gerla, M.Y. Sanadidi, R. Wang, *TCP Westwood: Bandwidth estimation for enhanced transport over wireless links* Proc. of ACM MOBICOM, 2001, pp. 287–297

[15] M. Schiavone, P. Romirer-Maierhofer, F. Ricciato, A. Baiocchi, *Towards Bottleneck Identification in Cellular Networks via Passive TCP Monitoring* Lecture Notes in Computer Science, vol. 8487, pp. 72-85, 2014

[16] B. Constantine, G. Forget, R. Geib, R. Schrage, *RFC 6349: Framework for TCP Throughput Testing*

[17] A. Afanasyev, N. Tilley, P. Reiher, L. Kleinrock, *Host-to-Host Congestion Control for TCP*, IEEE Communication Surveys and Tutorials, vol. 12, no. 3, pp. 304–342, July 2010

[18] R.S. Prasad, M. Jain, C. Dovrolis, *Socket Buffer Auto-Sizing for High-Performance Data Transfers* Journal of Grid Computing, 2003, vol. 1, Issue 4, pp 361-376

[19] J. Semke, M. Mathis Mahdavi, *Automatic TCP Buffer Tuning*, Computer Communication Review, ACM SIGCOMM, vol. 28, no. 4, October 1998

[20] M.K. Gardner, W.-C. Feng, M. Fisk, *Dynamic Right-Sizing in FTP (drsFTP): Enhancing Grid Performance in User-Space* Proc. of IEEE Symposium on High-Performance Distributed Computing, July 2002

[21] M. Mathis, R. Reddy, *Enabling High Performance Data Transfers* Jan. 2003; available at: http://www.psc.edu/networking/perf tune.html

[22] M. Fisk, W. Feng, *Dynamic Right-Sizing: TCP Flow-Control Adaptation* Proc. of the 14th Annual ACM/IEEE SC2001 Conf., November 2001

[23] E. Weigle, W. Feng, *A Comparison of TCP Automatic Tuning Techniques for Distributed Computing* Proc. of the 11th IEEE International Symposium on High Performance Distributed Computing, 2002

[24] Linux kernel 3.18, https://www.kernel.org/

[25] M. Hirabaru, *Impact of Bottleneck Queue Size on TCP Protocols and Its Measurement*, IEICE Trans. of Commun., vol. E89-B, no. 1, Jan 2006

[26] Yi Wang, Guohan Lu, Xing Li , *A Study of Internet Packet Reordering* Lecture Notes in Computer Science, vol. 3090, pp. 350-359, 2004

[27] S. Jaiswal, G. Iannaccone, C. Diot, J. Kurose, D. Towsley, *Measurement and Classification of Out-of-Sequence Packets in a Tier-1 IP Backbone* IEEE/ACM Trans. Netw. 15, 1 (Feb 2007), 54-66. doi: 10.1109/TNET.2006.890117

[28] M. Mathis, J. Heffner, *RFC 4821: Packetization Layer Path MTU Discovery*

[29] N. Hu , Li (Erran) Li, Z. Morley Mao, P. Steenkiste, J. Wang, Locating Internet Bottlenecks: Algorithms, Measurements, and Implications SIGCOMM Comput. Commun. Rev. 34, 4 (August 2004), 41-54, doi: 10.1145/1030194.1015474

[30] N. Hu, P. Steenkiste, Evaluation and Characterization of Available Bandwidth Probing Techniques IEEE Journal on Selected Areas in Communications, vol. 21, no. 6, August 2003

DOI: 10.15439/2015F259
ACSIS, Vol. 5

# Run-time UI Adaptation in the Context of the Device-Independent Architecture

Jacek Chmielewski
Poznan University of Economics
Department of Information Technology,
Al. Niepodległości 10,
61-875 Poznan, Poland
Email: chmielewski@kti.ue.poznan.pl

□

*Abstract*—**The increasing diversity of end-devices used by users to access their applications and systems strengthens the need for device-independent methods for implementing these applications. The Device-Independent Architecture (DIA) is one of the available approaches to this problem, but it does not directly address the issue of user interface (UI) device-independency. This issue can be addressed by real-time UI adaptation, but it is not clear whether the DIA architecture requires new UI adaptation methods or may use existing ones. This paper presents results of our analysis of this issue. Through theoretical model-based analysis of UI adaptation in various application architectures and through case studies of practical UI adaptation solutions we came up with a conclusion that the DIA-based systems may use existing real-time UI adaptation methods. Although, they have to be used with a different set of optimization criteria.**

## I. Introduction

THE development of software applications that use end-devices to communicate and interact with users becomes a complex and time-consuming issue. The increasing diversity of Internet-connected end-devices (especially mobile devices) forces application developers to implement multiple variants of each application. Each software platform (Windows, Android, iOS, etc.) and each device type (smartphone, tablet, laptop, watch, glasses, smart TV, etc.) has its own requirements and constraints, which makes it difficult to address all of them with a single uniform implementation. Device-independency of the application logic and data is hindered by different programming languages and disparate APIs provided by different software platforms. Device-independency of the application user interface (UI) is even harder to address because of the number and diversity of possible input and output user communication channels – starting with screen sizes and resolutions and ending with non-standard symbolic interfaces popular in Internet of Things solutions.

To cope with this problem we have proposed the Device-Independent Architecture (DIA) [1] which solves the logic and data device-independence issues. However the DIA does not directly address the UI device-independence aspect, which is supposed to be solved with proper UI design [2], [3] and UI adaptation [4], [5].

To make sure the DIA does not hinder the ability to use UI adaptation to provide UI device-independence in the reported research we have sought to answer the following question: **Does DIA-based software may use existing real-time UI adaptation methods?**

To be able to properly analyze the problem we have defined a model of the run-time UI adaptation and generation process. We have used this model to theoretically examine the run-time UI adaptation and generation process in various software architectures similar to the DIA. Additionally we have performed a series of case studies of real implementations of UI adaptation methods to check if these practical solutions confirm our theoretical conclusions.

Our main findings are the following. Through our research we have shown that DIA-based software may use existing real-time UI adaptation methods designed for client-server systems. Moreover, we have learned that the main limiting factor for DIA-based implementations of these UI adaptation methods is not the performance of an end-device, but network latency and throughput. Therefore, to provide properly optimized UIs for DIA-based solutions, existing real-time UI-adaptation methods have to be used with a different set of key metrics and guidelines.

The paper is composed of five sections. Section I is the introduction. Sections II and III provide background information on the topics of UI adaptation and Device-Independent Architecture. Section IV contains the main discussion and overview of case studies. Finally, the paper is briefly concluded in Section V.

## II. UI adaptation

UI adaptation activities can be split into two phases: design-time UI adaptation and run-time UI adaptation. These two UI adaptation phases focus on different aspects that may influence the UI adaptation process. The whole process, with its various aspects, is best described by the CAMELEON Reference Framework [6], which provides designers and developers with generic principles for structuring and

□ This work was supported by the Poznan University of Economics

978-83-60810-66-8/$25.00©2015, IEEE

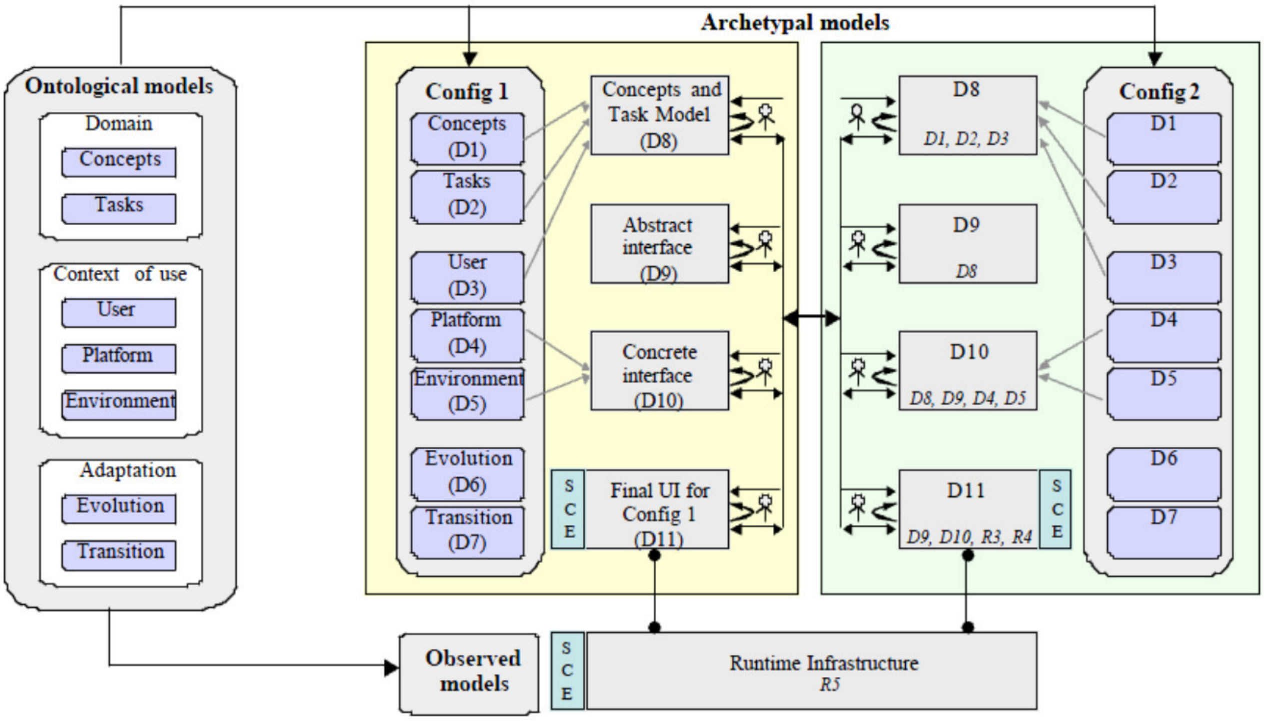


Fig. 1 CAMELEON Reference Framework

understanding a model-based UI development process. Model-based approaches [7], which rely on high-level specifications, provide the foundations for code generation and code abstraction. The framework fuses together different models that influence the overall UI adaptation. As shown in Figure 1, the framework covers the inference process from high-level abstract descriptions to run-time code, using a four-step reification process: from Concepts-and-Tasks Model (CTM), to Abstract User Interface (AUI), to Concrete User Interface (CUI), to Final User Interface (FUI). The CTM brings together concepts and tasks descriptions produced by designers for a particular interactive system and a particular target. The AUI is a universal description of the domain concepts and functions in a way that is independent of the UI implementation (in terms of UI widgets). At the CUI level the look and feel of a UI is defined, but the description is still device-independent. Finally, the FUI is expressed in a format suitable for a specific end-device and is tailored for this device. At each step the reification is influenced by the "context of use", defined as a set of parameters describing a user, a platform and the environment. Most of this process belongs to the design-time phase. The run-time phase includes the last reification from the CUI (device-independent) to the FUI (device-specific) and translations between FUI variants.

Both UI adaptation phases are different in nature. In our research on device-independent systems we do not address general UI design issues and we focus on the run-time UI adaptation phase, assuming that the design-time phase produces a device-independent UI description, which is used as a starting point for the run-time UI adaptation.

## III. Device-Independent Architecture

The Device-Independent Architecture (DIA) has been proposed to facilitate analysis and development of applications that can be made available to users via any capable device from the large, diverse and fast growing pool of Internet-enabled end-devices – i.e., devices that are used directly by users to interact with an application, but not sensors that passively record a state of an environment. As presented in [8], the idea of DIA originates from the Service-Oriented Architecture, where systems are decomposed into atomic services, and processes use such services without knowing much about their implementation. A similar approach can be used to decompose an end-device. Each end-device, be it a laptop or a smartphone, provides: resources, services, and user interaction channels. Resources encompass processing power, memory and storage. Services are providers of context information, such as location, temperature, light intensity, and data from other types of sensors. User interaction channels (both incoming and outgoing) are the means to communicate with a user and include: screen, vibration, keyboard, microphone, camera, etc. The key concept is to use external resources, instead of what is provided by an end-device, and to generalize the way services and user interaction channels are accessed. Therefore, in DIA, the separation of application from end-

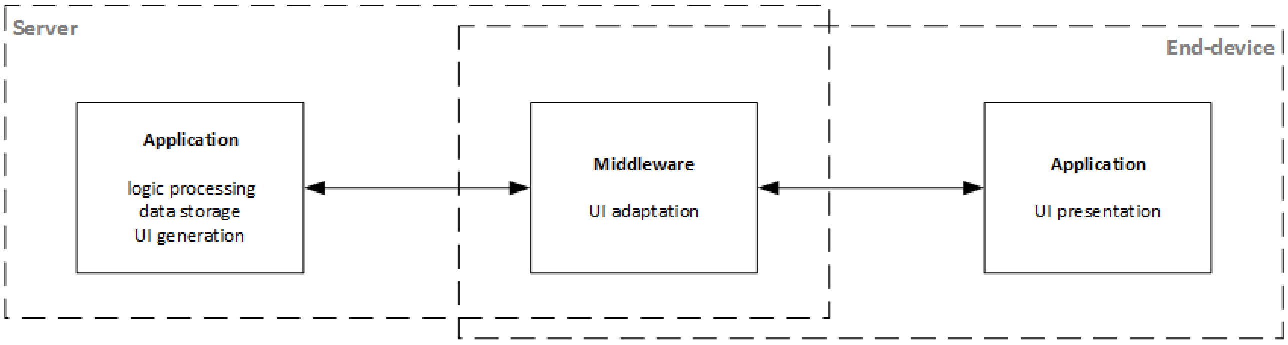


Fig. 2 Device-Independent Architecture diagram

devices, which enables the device independence, is achieved by:

- executing an application outside of end-devices,
- accessing sensor data provided by a device via a standardized API,
- using universal UI descriptions, and
- communicating with a user via a set of well-defined user interaction channels.

The execution of the application on external resources ensures that the application logic does not depend on the hardware or software platform of an end-device. The interesting consequence is that, in this architecture, end-devices could be deprived of their general purpose resources, as these resources are not needed. Services publish data in service-specific formats (e.g., location coordinates for a geolocation service, numerical data for a temperature sensor, and so on) independently of their implementation on a particular end-device. Therefore, it is feasible to build a middleware providing a device-independent API, such as the one proposed in Wolfram Language [9], to access such services. The usage of a universal UI description is a key requirement for making the UI of an application independent of parameters of user communication channels available on a given end-device (e.g. screen size and pixel density).

However, to enable a UI presentation tailored to parameters of a specific end-device, the generic UI description has to be properly adapted before reaching the user. That is why we have decided to research whether DIA-based software may use existing real-time UI adaptation methods.

## IV. Model and Analysis

Run-time UI adaptation is a process that transforms a high-level, device-independent UI description (often model-based) prepared at design-time into a final UI presentation. In ideal situations, the high-level UI description may be presented in different ways depending on the UI modality of available user communication channels. For example, presentation of the same UI could be done on screen (Graphical User Interface (GUI)) or via speakers (e.g. Voice User Interface (VUI)). In general, the execution of a run-time UI adaptation process requires three parameters: the UI description, the content and a context of use. The content is used to fill-in the UI. The context of use influences the UI adaptation process and allows tailoring the final UI to the user, her end-devices and situation (location, time, etc.).

### *A. Run-time UI adaptation model*

To be able to analyze the UI adaptation in different application architectures we have defined a simplified model of the UI adaptation and generation process. We call it the GARP model. The GARP model, presented in Figure 3, is composed of four main steps:

*Step 1: input gathering (G).* At the beginning of the process it is necessary to gather all input required for UI adaptation. The result of this step is a triplet of: UI

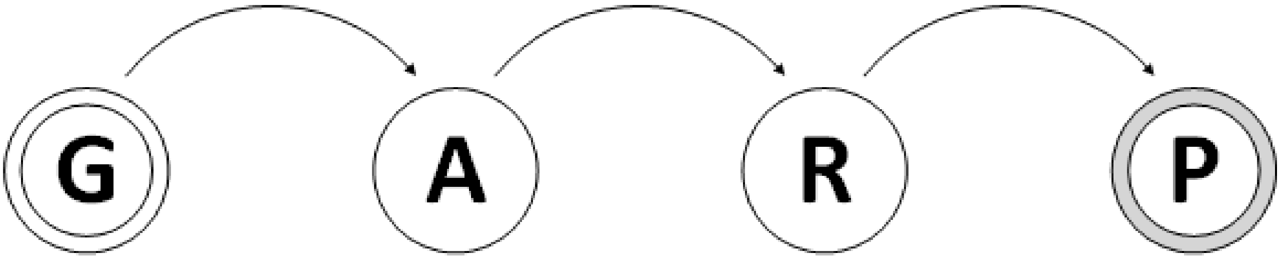


Fig. 3 Model of the UI adaptation and generation process

description, content and context.

*Step 2: adaptation (A).* In this step the content is used to fill-in the UI and the context is used to guide the transformation of the UI description into a final UI tailored for the user, her end-devices and situation. The result of this step is a device-specific UI description encoded with a specialized UI language such as HTML, QML, etc.

*Step 3: rendering (R).* The device-specific UI description provided by step 2 is interpreted here, in step 3 and the final UI presentation form is calculated. The final UI presentation form is data prepared for a specific user communication channel, e.g. pixels for screen or audio bits for speakers.

*Step 4: presentation (P).* The last step of the process is about presenting the UI to the user using a specific user communication channel of a specific end-device, e.g. showing images on screen or playing audio through speakers.

To make the analysis easier to follow, the model represents only the way towards a user and ignores the process of recording and interpreting user actions. Nevertheless, the path from a user can be modelled in a similar way, so our claims are valid for the whole user interaction loop.

For the analysis we have identified three classes of systems:

*Client-side adaptation systems* (CSA). Systems of this class include applications that are executed entirely on an end-device (a.k.a. local applications) and client-server applications with UI adaptation done on the client side.

*Server-side adaptation systems* (SSA). This class includes client-server applications with UI adaptation performed on the server side.

*DIA-based systems* (DIA), which include applications based on the Device-Independent Architecture.

Our goal is to see how the UI adaptation process differs among these classes of systems and how these differences influence the applicability of known UI adaptation methods. We acknowledge existence of in-between solutions. However, these three classes were selected to clearly show differences in the UI adaptation process.

### B. UI adaptation in CSA systems

The UI adaptation and generation process in CSA systems is done either entirely on an end-device (client side) or the G step is supported by the server side, which provides for example the content, UI description or user preferences. However, the fragment of the context gathering related to the end-device is local, so the G step can be seen as a task performed jointly by the server side and the client side. The way the G step is performed (locally or split between server and client sides) does not influence the actual UI adaptation, because from the point of view of the A step the results of the G step are always provided in the same way – locally.

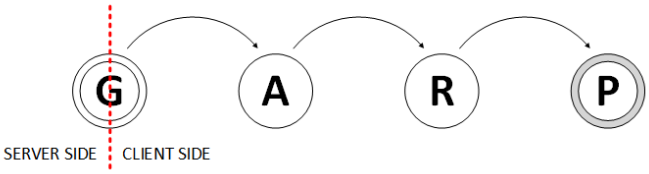


Fig. 4 GARP model in CSA systems

In CSA systems the A step may be implemented using existing UI adaptation methods designed for the use on an end-device. These methods are optimized for potentially limited processing capabilities of end-devices and are closely related to end-device characteristics and usage scenarios.

The local UI adaptation methods include solutions built-in into iOS and Android mobile operating systems and used by multiple mobile applications that run on various smartphones and tablets. On these mobile platforms the main issue is the diversity of screen sizes and pixel densities, so it is assumed that each application provides multiple variants of graphical assets (tailored to different screen densities) and some kind of a flexible layout that can be recalculated for any screen size. The drawback of these UI adaptation methods is that they are designed to cope with hardware parameters of a ‘standard device’ (in most cases a device with a touchscreen). Any UI adaptation that is supposed to take into account for example user preferences or non-standard devices, is not supported by the platform and has to be implemented manually.

The use of the server side for the G step usually does not change the fact that the adaptation implemented on the client side is somehow bound to the characteristic of an end-device. In our previous research [10] - [12] we have analyzed solutions that go beyond this local-only approach and use the server side to provide UI adaptation hints embedded in the high-level UI description provided by the G step, but even such extensions do not change the fact that the UI adaptation itself is device-specific, which makes it hard to reuse on other types of devices (devices with different hardware components, e.g. with two screens).

### C. UI adaptation in SSA systems

In SSA systems the two initial steps: G and A, are performed on the server side, and the two other steps: R and P, are performed on the client side. The server gathers all input data, runs the UI adaptation and sends the device-specific UI description to the client. The client then interprets the UI description and presents it to a user.

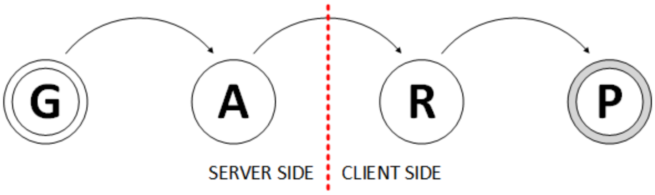


Fig. 5 GARP model in SSA systems

The SSA systems approach the issue of portability of the UI adaptation, shown for CSA systems, by implementing the A step on the server side. Such approach means that the UI adaptation is not bound by the performance of an end-device and can use external services to support the UI adaptation task (e.g. multimedia converters). Results of our previous research on UI adaptation in SSA systems [12] - [15] confirm that the A step in SSA systems may accommodate end-devices with disparate hardware configurations by using multiple or dynamic UI adaptation scenarios. However, the result of the A step is still interpreted on an end-device. Therefore is susceptible to differences in the final rendering and presentation on different end-devices. So full control of the resulting UI is not possible.

### *D. UI adaptation in DIA systems*

The Device-Independent Architecture is based on an assumption that the whole processing is done outside of an end-device (the client side) and the end-device receives a pre-rendered UI ready for presentation, without the need for any interpretation. So in the case of DIA systems all three initial steps of the GARP model are done on the server side and only the P step is performed on the client side. The data transferred between the server and the client is usually a stream (e.g. a video or audio stream) or a static UI state (e.g. an image of the UI to be presented on screen or audio file to be played through speakers) ready to be presented on an end-device.

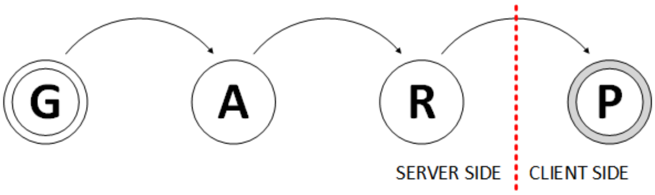


Fig. 6 GARP model in DIA systems

The DIA approach enables full control over the final UI presented to a user by implementing also the R step on the server side. UI adaptation methods used in DIA systems can be still the same as for SSA systems, but the fact that the end-device handles only the P step ensures that devices will not show a UI in a way that deviates from the designer intentions. The consequence of moving the R step to the server side is a different kind of data being transferred between the server side and the client side. In SSA systems, the client side receives a device-specific UI description encoded in a specialized UI language. In DIA systems, the server side has to send either a continuous stream of data tailored for specific user communication channels (e.g. video stream for a screen or an audio stream for speakers) or a static UI state composed of multiple files that are targeted at different user communication channels (e.g. image files to be shown on a screen or audio files to be played through speakers). The main difference here is the increased size of data that has to be transferred. More data to transfer could mean longer response times, but our previous research [16] showed that in analyzed scenarios DIA-based systems can still maintain proper response times to UI interactions initiated by a user, despite the increased size of transferred data.

## V. CONCLUSION

The Device-Independent Architecture can be treated as a special case of a client-server architecture, in which the client side is assumed to be an extremely thin client and in which all the processing is done on the server side. The DIA takes it even further and defines the client side as a set of user communication channels, which makes it possible to model multiple end-devices as a complex client device, but this distinction does not necessarily change the way the UI adaptation is performed. Therefore, DIA systems may use the same existing UI adaptation methods that were designed for SSA systems, or for client-server systems in general.

The main difference is related to the fact that in DIA systems the data transferred between the server side and the client side tends to be larger than in the case of SSA systems. Therefore, network usage optimization is crucial. Especially that the transmission delay will directly influence the UI responsiveness. Moreover, used communication protocols and formats of presentation data sent to an end-device have to be negotiated beforehand, to make sure that the end-device is able to receive and present it properly.

Summarizing, despite using a different implementation of the GARP model, the DIA systems may use existing real-time UI adaptation methods. The difference in the implementation of GARP model influences only the optimization of the UI adaptation and generation process. In CSA systems the key optimization aspect is end-device performance. In SSA systems the key optimization aspect is uniform interpretation of the device-specific UI description. While, in DIA systems the key optimization aspects are network-related. First, it is necessary to use data formats that minimize the amount of bits that have to be transmitted. Second, it is crucial to use the best possible data transfer protocols. The best are the ones with low overhead, low latency and support for QoS. Both points should be taken into account by the run-time UI adaptation task, because the nature of a UI (state-based or continuous) may influence the set of suitable transmission protocols.

We expect that different protocols will be best suited for different user interaction scenarios. Our next research goal in this area is be to identify user interaction patterns and UI design patterns, which could be used to define rules for selecting the best protocol and data formats for a given user interaction scenario.

## REFERENCES

[1] Chmielewski, J., Towards an Architecture for Future Internet Applications, in: The Future Internet, vol. Lecture Notes in Computer Science 7858 , Springer Berlin Heidelberg, 2013, pp. 214-219, ISBN 978-3-642-38081-5, DOI 10.1007/978-3-642-38082-2_18

[2] Meixner, G., Calvary, G., Coutaz, J., Introduction to Model-Based User Interfaces, W3C Working Group Note, December 2013, http://www.w3.org/2011/mbui/drafts/mbui-intro/

[3] Sottet, J. S., Calvary, G., Favre, J. M., & Coutaz, J., Megamodeling and metamodel-driven engineering for plastic user interfaces: MEGA-UI. In Human-Centered Software Engineering, pp. 173-200. Springer London. 2009, DOI 10.1007/978-1-84800-907-3_8

[4] Jaouadi, I., Ben Djemaa, R., & Ben Abdallah, H., Interactive Systems Adaptation Approaches: A survey. In ACHI 2014, The Seventh International Conference on Advances in Computer-Human Interactions, pp. 127-131. March 2014, ISSN: 2308-4138, ISBN: 978-1-61208-325-4

[5] Ye, J. H., & Herbert, J., Framework for user interface adaptation. In User-Centered Interaction Paradigms for Universal Access in the Information Society, pp. 167-174. Springer Berlin Heidelberg. 2004 DOI 10.1007/978-3-540-30111-0_14

[6] Calvary, G., Coutaz, J., Bouillon, L., Florins, M., Limbourg, Q., Marucci, L., Paternò, F., Santoro, C., Souchon, N., Thevenin, D., Vanderdonckt, J., The CAMELEON Reference Framework, Deliverable 1.1, CAMELEON Project, 2000

[7] Szekely, P., Retrospective and challenges for model-based interface development, in Design, Specification and Verification of Interactive Systems ’96, Eurographics 1996, pp. 1-27. Springer Vienna. 1996 DOI 10.1007/978-3-7091-7491-3_1

[8] Chmielewski, J., and K. Walczak, Application Architectures for Smart Multi-device Applications, in: Proceedings of the Workshop on Multi-device App Middleware 2012, Workshop on Multi-device App Middleware 2012, Montreal (Canada), December 3 – 7, 2012, ACM, New York, 2012, pp. 5:1 - 5:5, ISBN 978-1-4503-1617-0, DOI 10.1145/2405172.2405177

[9] Wolfram Language for Knowledge-Based Programming, https://www.wolfram.com/language/, 2015

[10] Chmielewski, J., K. Walczak, and W. Wiza, Mobile Interfaces for Building Control Surveyors, in: Software Services for e-World, IFIP Advances in Information and Communication Technology, Vol. 341, ed. Cellary, W., and E. Estevez, The 10th IFIP WG.6.11 Conference on e-Business, e-Services and e-Society I3E 2010, Buenos Aires, Argentina, November 3-5, 2010, Springer, 2010, pp. 29-39, ISBN 978-3-642-16282-4 DOI 10.1007/978-3-642-16283-1_7

[11] Rykowski, J., and J. Chmielewski, Automatyczna generacja zintegrowanego interfejsu człowiek-maszyna na potrzeby inteligentnego budynku , in: Inteligentne budynki - teoria i praktyka, ed. Mikulik, J. , Oficyna Wydawnicza Text, Kraków, 2010, pp. 166-188, ISBN 978-83-60560-54-9

[12] Chmielewski, J., K. Walczak, W. Wiza, and A. Wójtowicz, Adaptable User Interfaces for SOA Applications in the Construction Sector, in: SOA Infrastructure Tools - Concepts and Methods, ed. Ambroszkiewicz, S., J. Brzeziński, W. Cellary, A. Grzech, and K. Zieliński, Wydawnictwa Uniwersytetu Ekonomicznego w Poznaniu, Poznań, 2010, pp. 493-469, ISBN 978-83-7417-544-9

[13] Jansen, A., Bronmark, J., & Chmielewski, J. (2013). Method of adapting a user interface in industrial process monitoring and control applications. The Swedish Patent and Registration Office. SE 1300702-6

[14] Walczak, K., W. Wiza, D. Rumiński, J. Chmielewski, and A. Wójtowicz, Adaptable User Interfaces for Web 2.0 Educational Resources, in: IT Tools in Management and Education - Selected Problems, ed. Kiełtyka, L., Wydawnictwo Politechniki Częstochowskiej, Częstochowa, 2011, pp. 104-124, ISBN 978-83-7193-508-4

[15] Walczak, K., J. Chmielewski, W. Wiza, D. Rumiński, and G. Skibiński, Adaptable Mobile User Interfaces for e-Learning Repositories, IADIS International Conference on Mobile Learning, Avila (Spain), Marc 10-12, 2011, IADIS, 2011, pp. 52-60, ISBN 978-972-8939-45-8

[16] Chmielewski, J., Device-Independent Architecture for Ubiquitous Applications, in: Personal and Ubiquitous Computing, Volume 18, Issue 2, pp 481-488, Springer London, 2014, DOI 10.1007/s00779-013-0666-y

DOI: 10.15439/2015F157
ACSIS, Vol. 5

# The method for optimal server placement in the hypercube networks

Jan Chudzikiewicz
Military University of Technology,
ul. S. Kaliskiego 2, 00-908
Warszawa, Poland
Email: jchudzikiewicz@wat.edu.pl

Tomasz Malinowski
Military University of Technology,
ul. S. Kaliskiego 2, 00-908
Warszawa, Poland
Email: tmalin@ita.wat.edu.pl

Zbigniew Zieliński
Military University of Technology,
ul. S. Kaliskiego 2, 00-908
Warszawa, Poland
Email: zzielinski@wat.edu.pl

***Abstract*****—** In this paper, the problem of determining the most effective server placement in the hypercube network structure was considered. The algorithm consisting of two stages: first stage for the server placement and the second for generating the appropriate communication structure was described. The correctness of the algorithm has been verified through simulation tests, prepared and implemented in Riverbed Modeler environment. The results of these tests for exemplary structures were presented. Some properties of the server placement in the 4-dimensional hypercube network with soft degradation were investigated.

## I. Introduction

The computer networks with a regular structure as torus or hypercube ([1]–[4]) could be used in many kinds of specialized critical application (for instance military, aerospace or medical systems). An interconnection network with the hypercube logical structure is a well-known interconnection model for multiprocessor systems ([5]–[6]) and still hypercube networks are the field of interest of many theoretical studies concerning (among others) resource placement problem, which has been intensively studied in [7]–[12].

Specialized systems with critical application are usually used in real-time mode and required both very high reliability and high efficiency of data processing throughout all the network life cycle. In order to achieve high reliability of the system the network could be considered as soft degradable computer network [12]–[13]. In this kind of networks a processor identified as faulty is not repaired (or replaced) but access to it is blocked. New (degraded) network continues work after resources reassigning and under the condition that it meets special requirements. In turn, the efficiency of the system will depend heavily on the availability of resources (data bases, files or web data), which is determined by their placement in the network. So, for this kind of networks there is necessity for applying effective methods of resources placement. In the work [12] an analysis of the different schemas of resources placement in the 4-dimensional hypercube network with soft degradation was conducted.

Designing and exploitation of special networks in critical application is a comprehensive task that requires addressing a number of theoretical and practical problems. One of the problems is a skillful resources deployment in the network and modification of resources deployment after each phase of the network degradation. One of considered in the literature the resource placement problem is a combination the distance-$d$ and the $m$ adjacency problems, where a non-resource node must be a distance of at most $d$ from $m$ processors nodes [7]-[10], [12]. In [10] a perfect deployment has been introduced and analyzed which is defined as the minimum number of resources processors which are accessible by working processors. The definition - perfect deployment is a characteristic of the value of the generalized cost of information traffic in the network at a given load of tasks. In [12] the notion of $(m,d)$-perfect resources placement in the hypercube type structure $G$ has been extended to the such allocation of $k$ resources which minimizes the average distances between the working processors and resource processors in the structure $G$.

We investigate the case when a specialized computer system is based on the 4-dimensional hypercube skeleton network with communication nodes which could communicate between themselves via cable connections. The main task of the hypercube network is to provide efficient access to resources managed by the server (or cluster of servers) connected directly to one of the network nodes and semi-stationary clients communicating with the assigned network nodes via wireless links. The execution of applications by a client processor requires an access to server services and resources, also some results returned by the server must be submitted to other clients. We assume that all clients are responsible for performing the same or very similar tasks. Thus all clients will generate similar workload of the network. The problem which arises for the given network structure is to determine the most effective server placement in the network structure.

The main goal of this paper is to give an effective method of solving the server placement problem for the hypercube network along with its soft degradation process.

978-83-60810-66-8/$25.00©2015, IEEE

A generalized cost of a network traffic with a specified resources deployment and workload of a network is usually tested through experimental measurements or examined with the use of simulation methods. In the paper we apply a two phased approach. In the first stage we solve the problem of a server placement in the given network structure on the base of analytically determined attainability measure, which was proposed in [12]. It should be noticed that real cost of information traffic in a network for a given deployment of the server with resources depends on the nature of the tasks performed by clients in the network. In the second stage we have examined this problem with the use of simulation methods for the specified server deployment determined by the simple analytical method and given type of task load of the network.

We see our contributions as follows. Firstly, we have extended the approach proposed in [12] to the determining server placement in the hypercube network with soft degradation on the base of nodes attainability calculation. Secondly, we propose the algorithm of the communication structure assignation with the use of dendrite calculation. Next, we show the feasibility of this approach by applying obtained results to some possible structures of degraded 4-dimensional hypercube network and verifying effectiveness of server placement by simulation experiments.

The rest of the paper is organized as follows. In Section II, a basic definitions and properties were introduced. The calculation of radius and attainability for exemplary structures were presented. In Section III, the proposal of the algorithm determining server placement was presented. An illustration of the main algorithm steps for the exemplary structure was given. In Section IV, the results of simulation tests for verification the algorithm (implemented in Riverbed Modeler environment) were described. In Section V, some concluding remarks were presented.

## II. BASIC DEFINITIONS AND ASSUMPTIONS

**Definition 1.** The logical structure of processors network we call the structure of $n$-dimensional cube if is described by coherent ordinary graph $G = \langle E, U \rangle$ ($E$ – set of computer, $U$ – set of bidirectional data transmission links), which nodes can be described (without repetitions) by $n$-dimensional binary vectors (labels) in such a way that

$$[\delta(\varepsilon(e'), \varepsilon(e'')) = 1] \Leftrightarrow [(e', e'') \in U] \qquad (1)$$

where $\delta(\varepsilon(e'), \varepsilon(e''))$ is Hamming distance between the labels of nodes $e'$ and $e''$.

The Hamming distance between two binary vectors ε(e') and ε(e'') complies with the dependency:

$$\delta(\varepsilon(e'), \varepsilon(e'')) = \sum_{k\in\{1,\dots,n\}} (\varepsilon(e')_k \oplus \varepsilon(e'')_k)$$

where:

- $\varepsilon(e')_k$ – the $k$-th element of the binary vector ε(e'),
- $\oplus$ – modulo 2 sum.

We investigate the case when skeleton of the network has the logical structure of 4-dimensional hypercube (Fig. 1). A topology of the hypercube may be represented by an ordinary consistent graph whose nodes are described by 4-dimensional binary vectors such that the Hamming distance between vectors (labels) of the adjacent nodes equals one. If $|E| = 2^4$ and $|U| = 2|E|$, then such graph we called (non labeled) 4-dimensional cube and will be denote by $H^4$. Thus $H^4$ is a regular graph of degree of 4 i.e. such that the degree of a node $e \in E$ we determine as $\mu(e) = |E(e)|$, where $E(e)$ is a set of nodes adjacent to the node $e \in E$ and $\mu(e) = 4$ for each node $e$ of the graph $H^4$.

Let $d(e, e'|G)$ be the distance between nodes $e$ and $e'$ in a coherent graph G, that is the length of the shortest chain (in the graph $G$) connecting node $e$ with the node $e'$.

Let $r(e|G) = \max_{e'\in E(G)} d((e, e')|G)$ be the greatest distance from the node $e \in E(G)$ to another node of the set $E(G)$, and $r(G)$, and $D(G)$ (respectively) denote the radius and the diameter of a graph $G$ i.e. $r(G) = min\{r(e|G): e \in E(G)\}$ and $D(G) = max\{d(e', e''|G): \{e', e''\} \subset E(G)\}$.

**Property 1.** For the 4-dimensional cube $H^4$ the equation is complied

$$D(H^4) = r(H^4) = 4.$$

It is known that $D(G) \leq 2r(G)$.

If $r(e|G) = r(G)$ then the node $e$ is called the central node of the network $G$.

Denote by $E^{(d)}(e|G) = \{e' \in E(G): d(e, e'|G) = d\}$ for $d \in \{1, \dots, D(G)\}$, and by

$$\varsigma(e|G) = \left(\varsigma_1(e|G), \dots, \varsigma_{r(e|G)}(e|G)\right) \text{ for}$$

$$\varsigma_d(e|G) = \left|E^{(d)}(e|G)\right| \qquad (2)$$

**Definition 2.** Let $\varphi(e|G) = \sum_{e'\in E(G)} d(e, e'|G)\ \left(e \in E(G)\right)$ be attainability of the computer $e$ in the network $G$ and by $\Phi(G) = \sum_{e\in E(G)} \varphi(e|G)$ attainability of the network $G$.

Using (2) we have

$$\varphi(e|G) = \sum_{d=1}^{r(e|G)} d\varsigma_d(e|G) \qquad (3)$$

**Property 2.** $\Phi(H^4) = 512$ because $\forall_{e\in E(H^4)}: \left(r(e|H^4) = 4 \wedge \varsigma_d(e|H^4) = \binom{4}{d}\right)$. Using (3) we have $\forall_{e\in E(H^4)}: \varphi(e|H^4) = 32$ and $|E(H^4)| = 2^4$, then $\Phi(H^4) = |E(H^4)|\ \varphi(e|H^4)$ [12].

**Example 1.** Figure 1 presents all the seven possible cyclic structures upon the occurrence of $k = 7$ consecutive failures of processors of the network $H^4$ which are the subgraphs of $H^4$ [13].

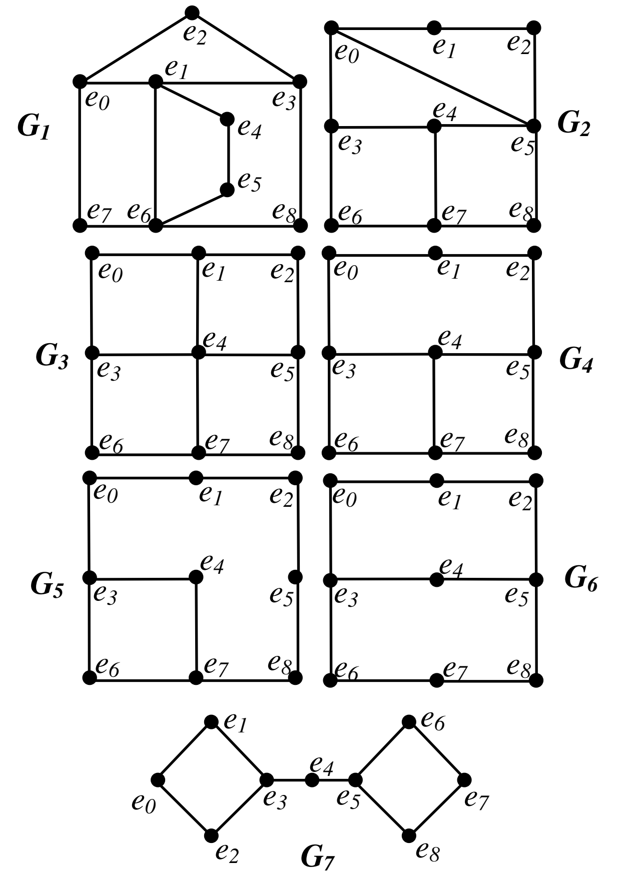


Fig. 1 Example of cyclic subgraphs of $H^4$ order 9 [13]

TABLE I.
THE $r(e|G), r(G)$ AND $D(G)$ FOR THE STRUCTURES PRESENTED ON THE FIGURE 1

| $r(e\|G_i)$ / $e \in E(G)$ | $r(e\|G_1)$ | $r(e\|G_2)$ | $r(e\|G_3)$ | $r(e\|G_4)$ | $r(e\|G_5)$ | $r(e\|G_6)$ | $r(e\|G_7)$ |
|---|---|---|---|---|---|---|---|
| $e_0$ | 3 | 3 | 4 | 4 | 4 | 4 | 6 |
| $e_1$ | 2 | 4 | 3 | 4 | 4 | 4 | 5 |
| $e_2$ | 4 | 4 | 4 | 4 | 4 | 4 | 5 |
| $e_3$ | 3 | 3 | 3 | 3 | 4 | 3 | 4 |
| $e_4$ | 3 | 3 | 2 | 3 | 4 | 3 | 3 |
| $e_5$ | 4 | 3 | 3 | 3 | 4 | 3 | 4 |
| $e_6$ | 3 | 4 | 4 | 4 | 4 | 4 | 5 |
| $e_7$ | 3 | 4 | 3 | 4 | 4 | 4 | 6 |
| $e_8$ | 3 | 3 | 4 | 4 | 4 | 4 | 5 |
| $r(G)$ | 2 | 3 | 2 | 3 | 4 | 3 | 3 |
| $D(G)$ | 4 | 4 | 4 | 4 | 4 | 4 | 6 |

It should be noticed, that for the given network structure $G$ on the base of the obtained measures $r(e|G)$ it would be rational to choose the server placement at the central node of the network or in the node with the minimum value $r(e|G)$. In some cases (let's consider the structures $G_2, G_4, G_5, G_6$) we are not able to choose the best server placement. Then we can have determined $\varsigma(e|G)$ using (2) and $\varphi(e|G)$ using (3) for these structures. The determined values of $\varsigma(e|G)$, $\varphi(e|G)$ and $\Phi(G)$ are presented in table II.

**Definition 3.** Let $T = \langle E, U^* \rangle$ be the dendrite i.e. such coherent acyclic partial graph of $G$ that:

$$\exists \langle e', e'' \rangle \in U \Rightarrow \langle e', e'' \rangle \in U^* \Leftrightarrow$$
$$\left[ \left( d(e_i, e') \neq d(e_i, e'') \right) \wedge d(e', e'') = 1 \right] \text{for } r(e_i) = \underset{e \in E(G)}{min} r(e).$$

The dendrite $T$ is a communication structure of $G$. The method for determined the dendrite $T$ is presented in section III.

TABLE II.
THE $\varsigma(e|G)$, $\varphi(e|G)$ AND$\Phi(G)$ FOR THE STRUCTURES $G_2, G_4, G_5, G_6$ PRESENTED ON THE FIGURE 1

| | $G_2$ | | | | | $G_4$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d(e,e'\|G)$ / $e \in E(G)$ | **1** | **2** | **3** | **4** | $\varphi(e\|G_2)$ | **1** | **2** | **3** | **4** | $\varphi(e\|G_4)$ |
| $e_0$ | 3 | 4 | 1 | 0 | 14 | 2 | 3 | 2 | 1 | 18 |
| $e_1$ | 2 | 2 | 3 | 1 | 19 | 2 | 2 | 3 | 1 | 19 |
| $e_2$ | 2 | 3 | 2 | 1 | 18 | 2 | 3 | 2 | 1 | 18 |
| $e_3$ | 3 | 3 | 2 | 0 | 15 | 3 | 3 | 2 | 0 | 15 |
| $e_4$ | 3 | 4 | 1 | 0 | 14 | 3 | 4 | 1 | 0 | 14 |
| $e_5$ | 4 | 3 | 1 | 0 | 13 | 3 | 3 | 2 | 0 | 15 |
| $e_6$ | 2 | 3 | 1 | 1 | 18 | 2 | 3 | 2 | 1 | 18 |
| $e_7$ | 3 | 2 | 2 | 1 | 17 | 3 | 2 | 2 | 1 | 17 |
| $e_8$ | 2 | 4 | 2 | 0 | 16 | 2 | 3 | 2 | 1 | 18 |
| | | $\Phi(G_2)$ | | | 144 | | $\Phi(G_4)$ | | | 152 |

| | $G_5$ | | | | | $G_6$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d(e,e'\|G)$ / $e \in E(G)$ | **1** | **2** | **3** | **4** | $\varphi(e\|G_5)$ | **1** | **2** | **3** | **4** | $\varphi(e\|G_6)$ |
| $e_0$ | 2 | 3 | 2 | 1 | 18 | 2 | 3 | 2 | 1 | 18 |
| $e_1$ | 2 | 2 | 3 | 1 | 19 | 2 | 2 | 3 | 1 | 19 |
| $e_2$ | 2 | 2 | 2 | 2 | 20 | 2 | 3 | 2 | 1 | 18 |
| $e_3$ | 3 | 2 | 2 | 1 | 17 | 3 | 2 | 3 | 0 | 15 |
| $e_4$ | 2 | 3 | 2 | 1 | 18 | 2 | 4 | 2 | 0 | 16 |
| $e_5$ | 2 | 2 | 3 | 1 | 19 | 3 | 3 | 2 | 0 | 15 |
| $e_6$ | 2 | 3 | 2 | 1 | 18 | 2 | 3 | 2 | 1 | 18 |
| $e_7$ | 3 | 2 | 2 | 1 | 17 | 2 | 2 | 3 | 1 | 19 |
| $e_8$ | 2 | 3 | 2 | 1 | 18 | 2 | 3 | 2 | 1 | 18 |
| | | $\Phi(G_5)$ | | | 164 | | $\Phi(G_6)$ | | | 156 |

## III. THE METHOD AND THE ALGORITHM FOR DETERMINING A SERVER PLACEMENT AND COMMUNICATION STRUCTURE

The method consists two stages. In the first stage for $G$, as the first node we choose a node that $r(e_i) = \underset{e \in E(G)}{min} r(e)$ or $\varphi(e_i|G) = \underset{e \in E(G)}{min} \varphi(e|G)$ – *server placement*. In the second stage for the chosen node we determine the dendrite $T$, which is a communication structure satisfying the condition $d_{max}(e_i|T) = r(e_i)$. Based on the presented method the algorithm for determining *the server placement* and the communication structure was developed**.**

*The algorithm for determining the server placement and communication structure.*

**Step 1.**

Determine $r(e|G)$ for $e \in E(G)$.

**Step 2.**

Choose a node $e_i \in E(G)$ such that $r(e_i) = \underset{e \in E(G)}{min} r(e)$. If $|\{e_i\}| > 1$ go to step 3 else go to step 5.

**Step 3.**
Determine $\varphi(e|G)$ for $e \in E(G)$.

**Step 4.**
Choose a node $e_i \in E(G)$ such that $\left(\varphi(e_i|G) = \min_{e\in E(G)} \varphi(e|G)\right) \wedge \left(\mu(e_i) = \max_{e\in E(G)} \mu(e)\right)$.
Selected node $e_i$ will be a central node of dendrite.

**Step 5.** *(second stage)*
Set $k = 1$ and $(E(T) = E^0(e_i|G) = \{e_i\})$.

**Step 6.**
$E(T) = E(T) + E^k(e_i|G)$.
$U^* = U^* \bigcup_{(e' \in E^{k-1}(e_i|G)) \wedge (e'' \in E^k(e_i|G))} \langle e', e'' \rangle$.

**Step 7.**
$k = k + 1$.
Check if $k > r(e_i)$.
YES
  Go to step 8.
NO
  Return to step 6.

**Step 8.**
The end of the algorithm.

Figure 2 is an illustration of the algorithm. In the first stage of the algorithm (steps 1-4) the central node $e_1$ for the structure $G_1$ (figure 1) was appointed. In the table III the results of determining radius (A) and attainability (B) of $G_1$ nodes are presented. Determination of the attainability of $G_1$ nodes is unnecessary, and should be used only if the steps 1-2 of the algorithm will not allow unambiguous to determine the central node for given structure $G$.

Dendrite $T$ determined in the second stage of the algorithm for the central node $e_1$, is one of the possible (but optimal ) communication structure.

TABLE III
THE RESULTS OF DETERMINING THE RADIUS (A), AND THE ATTAINABILITY (B) OF THE $G_1$ NODES

A

| $r(e\|G_1)$ / $e \in E(G_1)$ | $r(e\|G_1)$ |
|---|---|
| $e_0$ | 3 |
| $e_1$ | **2** |
| $e_2$ | 4 |
| $e_3$ | 3 |
| $e_4$ | 3 |
| $e_5$ | 4 |
| $e_6$ | 3 |
| $e_7$ | 3 |
| $e_8$ | 3 |

B

| $d(e,e'\|G_1)$ / $e \in E(G_1)$ | 1 | 2 | 3 | 4 | $\varphi(e,G_1)$ |
|---|---|---|---|---|---|
| $e_0$ | 3 | 3 | 2 | 0 | 15 |
| $e_1$ | 4 | 4 | 0 | 0 | 12 |
| $e_2$ | 2 | 3 | 2 | 1 | 18 |
| $e_3$ | 3 | 3 | 2 | 0 | 15 |
| $e_4$ | 2 | 3 | 3 | 0 | 17 |
| $e_5$ | 2 | 3 | 2 | 1 | 18 |
| $e_6$ | 4 | 3 | 1 | 0 | 13 |
| $e_7$ | 2 | 4 | 2 | 0 | 16 |
| $e_8$ | 2 | 4 | 2 | 0 | 16 |
| | | | | $\Phi(G_1)$ | 140 |

## IV. THE RESULTS OF SIMULATION STUDIES

Procedure for determining the best location of resources in the hypercube network specified in the section III has been verified through simulation tests. The aim of the test was to confirm the correctness of the theoretical considerations and arguments.

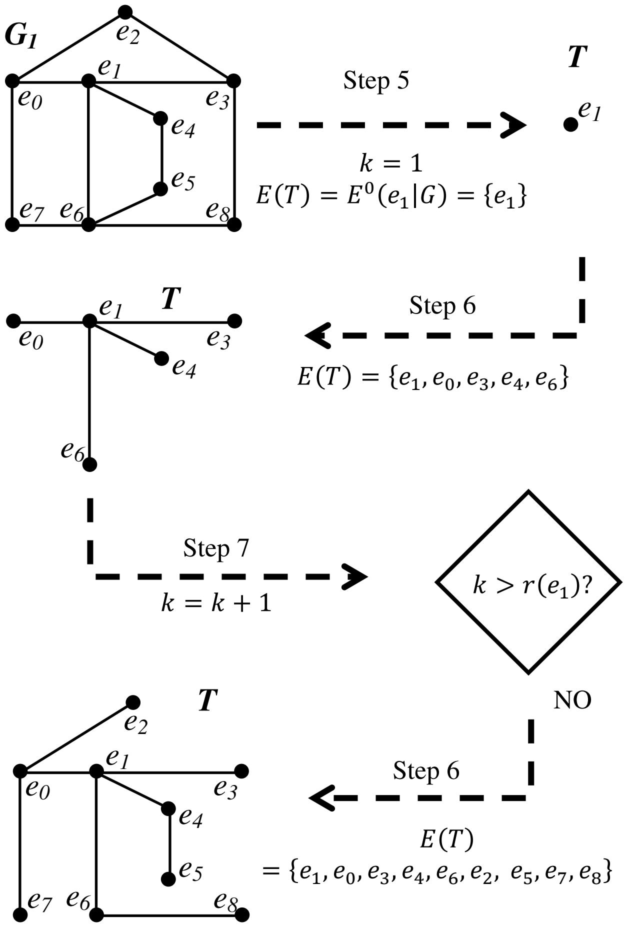


Fig. 2 An illustration of the algorithm steps

Simulation studies have been prepared and implemented in Riverbed Modeler environment. Subgraph $G_1$ (figure 1) has been the subject of research.

Nodes have been modeled as routers and LAN segments attached to them. The cases when the server (with different typical and popular network services, chosen arbitrarily by authors) is connected to the selected node within $T$ set of $G_1$ structure were examined. In the figure 3 different $T$ set of substructures (communication structures) correspond to different simulation scenarios is shown. The circle indicates node which the server is connected to.

For example, the network topology of the single scenario with $T$'s dendrite 1 was shown in the figure 4.

The name of the simulation scenario (*Dend_1* in figure 4) was associated with the node's number, which the server was connected to (number of the dendrite's central node).

The server was acting as a database server, ftp server, web server, and the node with which it was possible to communicate through VoIP (*Voice over IP*). Workstations within LAN segments (ten workstations in each segment) were functioning as the server's clients. All network services have used the standard application models, available at Riverbed Modeler ("High Load" ftp and database models, "Heavy Browsing" http model and "PCM Quality Speech" voice model) [15].

The communication structure (skeleton of the network) was modeled as a set of routers connected via 1,5 Mb/s links.

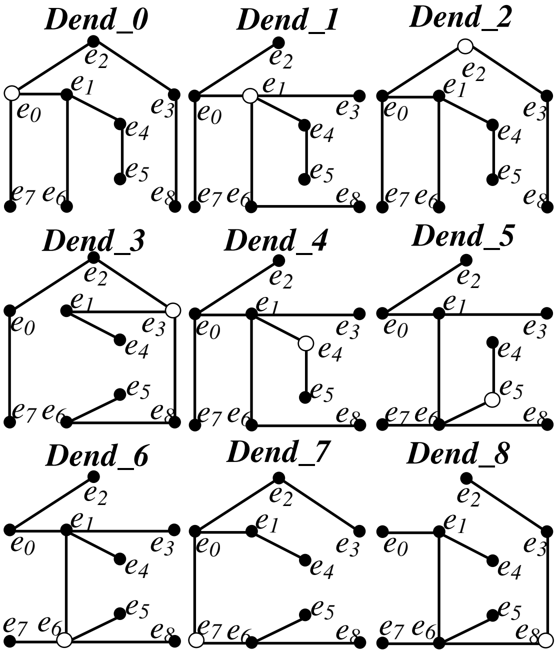


Fig. 3 The $T$ of dendrites of $G_1$

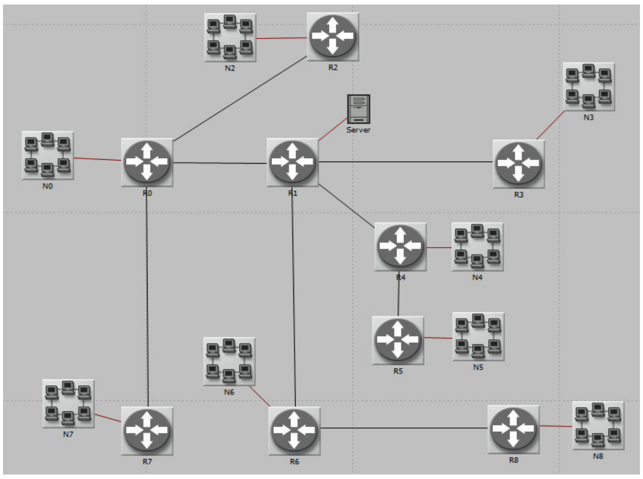


Fig. 4 Network topology of *Dend_1* scenario

Some interesting results, confirming the correctness of the procedure for determining the server placement and communication structure, are shown in the figures below. The dotted lines (e.g. Fig. 5) corresponds to the results obtained for the structure *Dend_1*, which is, according to the procedure, the most effective (the best) communication structure for subgraph $G_1$.

For each simulation scenarios (*Dend_0* to *Dend_8*) five characteristics were determined.

### A. End-to-end delay (EE_Del)

End-to-end delay is average delay in seconds for all LAN segments nodes communicating with the server through VoIP. The lowest value of *EE_Del* is desired (it is the best score).

Results obtained during the simulation are presented in the figure 5. The figure 6 shows the average values of *EE_Del*.

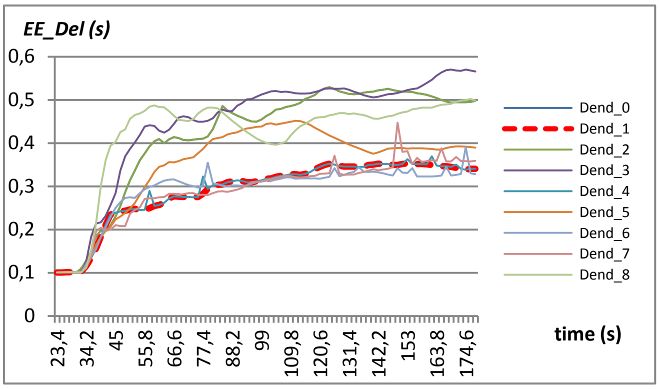


Fig. 5 *End-to-end delay* for VoIP transmission

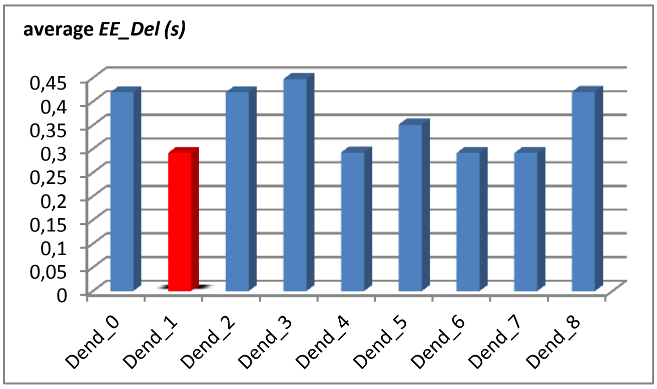


Fig. 6 Average *End-to-end delay* for *Dend_0* to *Dend_8* structures

### B. TCP Delay (TCP_Del)

TCP_del represents delay of TCP packets in seconds. This value is measured from the time an application data packet is sent from the source TCP layer to the time it is completely received by the TCP layer in the destination node. It is average delay in the complete network, for all connections. The lowest values are the best.

The results are presented in figures 7 and 8.

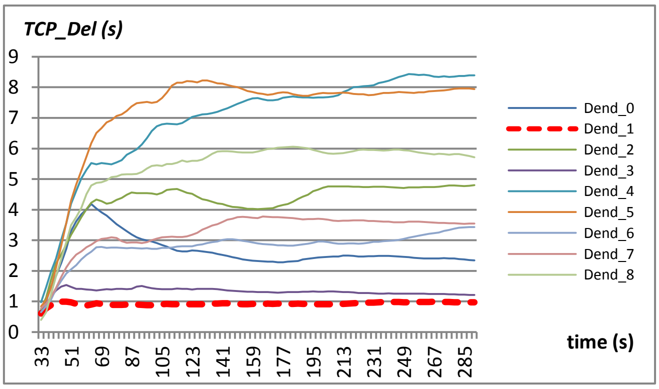


Fig. 7 *TCP delay* for TCP-based services

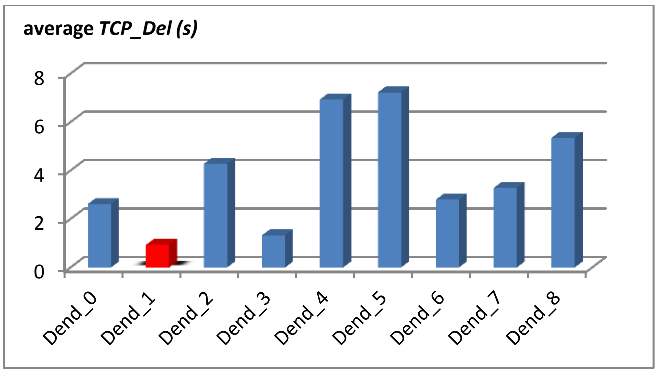


Fig. 8 Average *TCP delay* for TCP-based services

### C. Number of Hops (Nr_Hops)

*Nr_Hops* represents an average number of IP hops taken by data packets reaching at a destination node.

We expected the lowest value for *Dend_1* structure and results are presented in the figure 9.

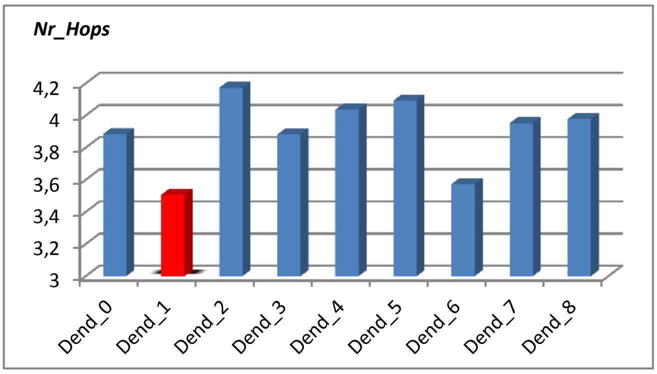


Fig. 9 Average *Number of Hops* for *Dend_0* to *Dend_8* structures

### D. Response Time (Res_Time)

Res_Time is average time elapsed between sending a request and receiving the response packet in seconds. It was measured for all the server's services (database, WWW and FTP).

The selected graph, *Response Time* for the database service was presented in the figure 10.

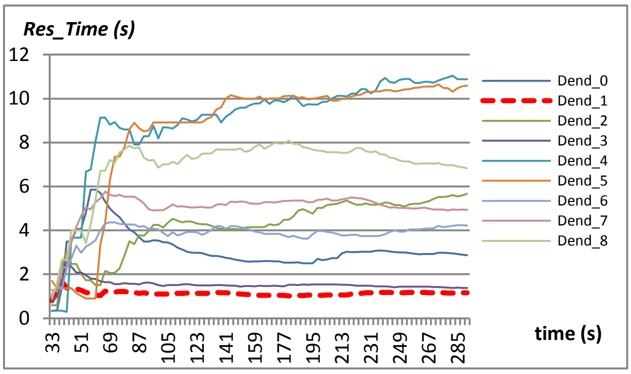


Fig. 10 *Response Time* for database service

Average values of the database server's response time for each simulation scenario are shown in the figure 11.

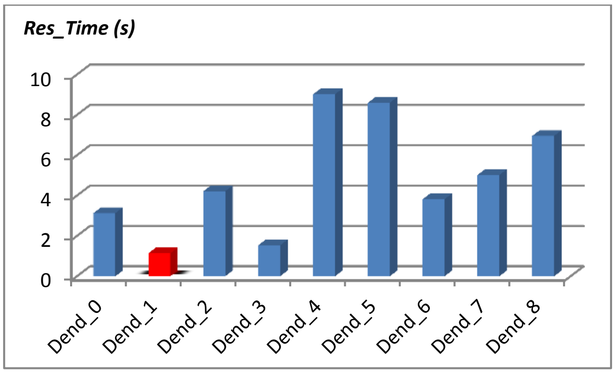


Fig. 11 Average *Response Time* for database service for all $G_1$'s dendrites

### E. Traffic Received (Traf_Rec)

*Traf_Rec* is an average number of bytes per second forwarded to server's application by the transport layer in the complete network. It was measured for all the server's service and treated as a transmission speed indicator, so we expected highest values for the best communication structure (*Dend_0* in the drawings of the selected service – Fig.12 and Fig.13).

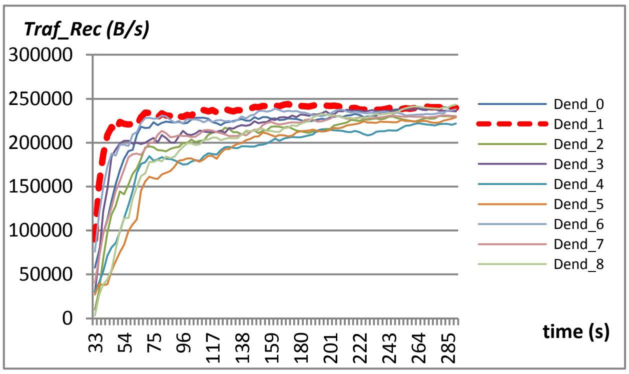


Fig. 12 *Traffic Received* for the FTP server

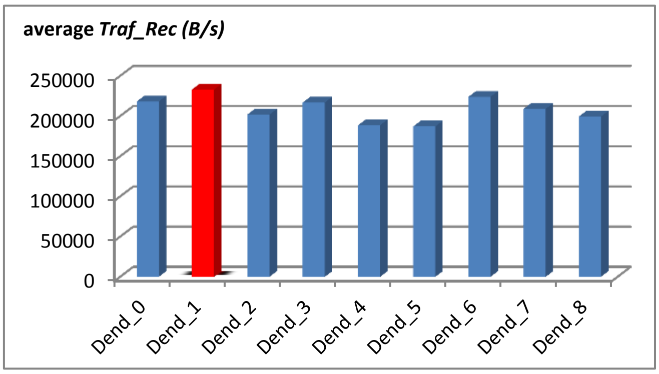


Fig. 13 Average FTP *Traffic Received* for all $G_1$'s dendrites

All the results are presented in Table III. It should be noted that the best result was reported in most measurements for *Dend_1* structure (winning factor - 78%).

TABLE IV
THE SIMULATION RESULTS FOR DENDRITES OF G 1 STRUCTURE

| | | | | HTTP | | DB | | FTP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Nr_Hops* | *TCP_Del (s)* | *EE_Del (s)* | ***Res_Time (s)*** | ***Traf_Rec (B/s)*** | ***Res_Time (s)*** | ***Traf_Rec (B/s)*** | ***Res_Time (s)*** | ***Traf_Rec (B/s)*** | **Winning factor (%)** |
| Dend_0 | 3.74 | 2.59245 | 0.419 | 1.113 | 70013 | 3.129 | 15163 | 8.22 | 170518 | |
| Dend_1 | 3.43 | 0.87554 | 0.291 | 0.641 | 136128 | 1.137 | 18214 | 5.37 | 188778 | **78** |
| Dend_2 | 4.10 | 4.043734 | 0.419 | 1.364 | 34099 | 4.210 | 14121 | 14.25 | 153116 | |
| Dend_3 | 3.75 | 1.307666 | 0.447 | 0.936 | 119991 | 1.532 | 18350 | 6.86 | 168044 | **11** |
| Dend_4 | 3.87 | 6.015415 | 0.292 | 1.959 | 25875 | 9.018 | 12745 | 19.35 | 140109 | |
| Dend_5 | 3.88 | 6.610183 | 0.351 | 1.968 | 28586 | 8.599 | 11137 | 19.67 | 134400 | |
| Dend_6 | 3.46 | 2.499316 | 0.291 | 0.873 | 66342 | 3.818 | 15677 | 10.45 | 179894 | |
| Dend_7 | 3.76 | 2.953067 | 0.291 | 1.397 | 43371 | 5.018 | 16540 | 11.25 | 161662 | |
| Dend_8 | 3.88 | 5.09135 | 0.420 | 1.417 | 36789 | 6.954 | 14086 | 16.71 | 143929 | |

## V. CONCLUSION

Correctness of developed algorithm and its usefulness for determining server placement and the optimal communication structure in the hypercube network with soft degradation on the base of nodes attainability calculation was confirmed by simulation tests. Although the simulation tests conducted mainly related to typical network services that results could be transferred to the network-critical applications. One of tested services which might be regarded as the critical application was capability of communicating through VoIP (Voice over IP). Obviously different system parameters aren't equally important and their importance depends on the application type. In the paper the influence of the network communication structure for obtained values of exemplary parameters was investigated.

Further work will address adaptation of the algorithm for various types of network structures. Work will focus on networks with dynamic structure reconfiguration.

## REFERENCES

[1] H. Hongwei1, S. Wei, X. Youzhi, Z. Hongke, "A Virtual Hypercube Routing Algorithm for Wireless Healthcare Networks", Chinese Journal of Electronics, Vol.19, No.1, Jan. 2010, pp. 138-144.

[2] Po-Jen Chuang, Bo-Yi Li, Tun-Hao Chao, "Hypercube-based Data Gathering in Wireless Sensor Networks", Journal Of Information Science And Engineering 23, 2007, pp. 1155-1170.

[3] A. Z. Zieliński, J. Chudzikiewicz, Arciuch, R. Kulesza, "Sieć procesorów o łagodnej degradacji i strukturze logicznej typu sześcianu 4-wymiarowego", in Metody wytwarzania i zastosowania systemów czasu rzeczywistego, L. Trybus, Ed., Warszawa: Wydawnictwo komunikacji i Łączności, 2011, pp. 219-232. (in Polish).

[4] A. Arciuch, "Reliability state of connections in a microprocessor network with binary hypercube structure", Electrical Review, R.86, No. 9/2010, pp. 154–156.

[5] T. Ishikawa, "Hypercube multiprocessors with bus connections for improving communication performance", IEEE Trans. Computers, Vol. 44, No. 11, 1995, pp. 1338–1344.

[6] A. B. Izadi, F. Özunger, "A Real-Time Fault_Tolerant Hypercube Multiprocessor", IEE Proceedings – Computers and Digital Techniques, Vol. 149, No. 5, 2002, pp. 197–202.

[7] B. F. AlBdaiwia, B. Bose, "On resource placements in 3D tori", Journal of Parallel Distributed Computer vol. 63, 2003, pp. 838–845.

[8] B. F. AlBdaiwia, B. Bose, "Quasi-perfect resource placements for two-dimensional toroidal networks", Journal of Parallel Distributed Computer vol. 65, 2005, pp. 815-831.

[9] M. M. Bae, B. Bose, "Resource Placement in Torus-Based Networks", IEEE Transactions on Computers, vol. 46, no. 10, October 1997, pp. 1083-1092.

[10] N. Imani, H. Sarbazi-Azad, A.Y. Zomaya, "Resource placement in Cartesian product of networks", Journal of Parallel Distribiuted Computer 70, 2010, pp. 481-495.

[11] P. Moinzadeh, H. Sarbazi-Azad, N. Yazdani, Resource Placement in Cube-Connected Cycles, The International Symposium on Parallel Architectures, Algorithms, and Networks, IEEE Computer Society, 2008, pp. 83-89.

[12] J. Chudzikiewicz, Z. Zieliński, "On some resources placement schemes in the 4-dimensional soft degradable hypercube processors network", Advances in Intelligent and Soft Computing. Proc. of the Ninth Int. Conf. on Dependability and Complex Systems DepCoS-RELCOMEX (W. Zamojski at al. Eds., Series Ed.: Kacprzyk Janusz), Springer 2014, pp. 133-143.

[13] Z. Zieliński, „Podstawy diagnostyki systemowej sieci procesorów o łagodnej degradacji i strukturze hiperszescianu", Wojskowa Akademia Techniczna, Warszawa, 2012, 182p. (in Polish).

[14] T. Malinowski, A. Arciuch, "The procedure for monitoring and maintaining a network of distributed resources", Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 947–954, ACSIS, Vol. 2, DOI: 10.15439/2014F159.

[15] A. S. Sethi, V.Y. Hnatyshin, "The Practical OPNET User Guide for Computer Network Simulation", Chapman and Hall/CRC, 2012.

DOI: 10.15439/2015F265
ACSIS, Vol. 5

# Molines – towards a responsive Web platform for flood forecasting and risk mitigation

João L. Gomes, Gonçalo Jesus,
João Rogeiro, Anabela Oliveira
Laboratório Nacional de
Engenharia Civil, Information
Technology in Water and
Environment Group, Av. do Brasil
101, 1700-066 Lisboa, Portugal
Email: {jlgomes, gjesus, jrogeiro,
aoliveira}@lnec.pt

Ricardo Tavares da Costa,
André B. Fortunato
Laboratório Nacional de
Engenharia Civil, Estuaries and
Coastal Zones Division, Av. do
Brasil 101, 1700-066 Lisboa,
Portugal
Email: {rcosta,
afortunato}@lnec.pt

*Abstract*— **This paper presents an innovative real-time information system for enhanced support to flood risk emergency in urban and nearby coastal areas, targeting multiple users with distinct access privileges. The platform addresses several user requirements such as 1) fast, online access to relevant georeferenced information from wireless sensors, high-resolution forecasts and comprehensive risk analysis; and, 2) the ability for a two-way interaction with civil protection agents in the field. The platform adapts automatically and transparently to any device with data connection. Given its specific purpose, both data protection and tailored-to-purpose products are accounted for through user specific access roles. This paper presents the platform's overall architecture and the technologies adopted for server-side, focusing on communication with the front-end and with the wireless sensor network, and the user interface development, using state-of-the-art frameworks for cross-platform standardized development. The advantages of the adopted solution are demonstrated for the Tagus estuary inundation information system.**

## I. Introduction and Motivation

NATURAL and hydraulic structures related floods are severe threats to life and property. The main goal of flood risk management in aquatic environments is to reduce human losses and the damages related to floods, and should be supported by adequate hazard monitoring and timely early warning of the events.

Some of the world's most densely populated cities are located in estuarine low-lying areas facing thus a high risk of inundation with a potential for significant economic costs and the loss of lives. These areas are highly vulnerable due to the growing human activity in their margins. Simultaneously, the hazards in these environments are severe due to the combined effects of oceanic, atmospheric and river forcings. Furthermore, they are increasing due to the effects of climate change, such as sea level rise, growing storminess and more extreme river flows. Floods in estuaries are associated to particular climatological conditions, namely very high tidal levels and large fresh-water discharges, or of high tides and storm surge conditions [1]. In addition to these progressive, slow phenomena, that are possible to predict a few days in advance, episodes of very intense and concentrated in time rainfall can lead to urban flooding in areas with insufficient drainage conditions and flash floods in small watershed tributaries to the estuary [2]. The effects of high water levels in estuaries can also be exacerbated by human interventions in the system, particularly in urban areas where drainage system behavior has to be considered.

Recently, the processes of prediction, detection, notification and population warning are becoming increasingly assured by automated systems, such as SAGE-B [3]. These information systems can be valuable assets for risk management, supporting all fundamental data related to flood events and the emergency elements needed for rescue in the predicted flooded areas. Unfortunately, most flood management systems still suffer from significant functional limitations, due to the difficulties in the access to monitoring data and unreliable, scattered, multiple sources of information, the use of inadequate flood forecasting due to inaccurate modeling tools, that either neglect relevant processes or are coarsely applied to the site at risk, and insufficient sharing of information across multiple emergency actors [4].

With the recent use of reliable automatic data acquisition systems and highly efficient and accurate numerical models, the most important constraints for the operational use of real-time information systems have been minimized, allowing for adequate forecasting of relevant events [5], [6]. The integration of these tools into interactive and flexible computational GIS-based platforms has paved the way to a change of paradigm in routine and emergency management of coastal resources and harbor operations [7], [8]. These platforms take advantage of novel technologies to provide on-line, intuitive and geographically-referenced access to real-time data and model predictions, and to produce on-demand services. However, much remains to be done on the interoperability between data providers and data consumers,

This work was funded by the Portuguese National Foundation for Science and Technology through project MOLINES (PTDC/AAG-MAA/2811/2012)

978-83-60810-66-8/$25.00©2015, IEEE

cross-platform and multiple users' flexibility and speed of access to data.

The project MOLINES (*Modelling floods in estuaries. From the hazard to the critical management*) aims at integrating existing and new wireless sensor networks, accurate model forecasts at both urban and estuarine scales and IT technology to create a Web platform that can contribute to a fast, coordinated mobilization of emergency agents and other managing entities for a timely response to inundation events in the Tagus estuary. The proposed platform is generic, interactive, facilitating the coordination between emergency agents and the individual contribution of civil protection agents in the field, and can be deployed elsewhere. It aims at contributing to a coordinated strategic planning and emergency response in urban and nearby estuarine regions, optimizing the alert to authorities, duly supported by real time monitoring and predictions of inundation.

This paper describes the platform, its architecture, and all innovative aspects related to the UI, product creation and choice of technologies. Besides this introduction, chapter 2 provides a background on IT technologies and platforms for real time information access, identifying the key aspects to be addressed. Chapter 3 presents the concept and implementation of the solution, focusing on requirements and technology choices. The application to the MOLINES case study is briefly presented in Chapter 4, and Chapter 5 closes the paper with some considerations for future work.

## II. Background

Technology is dramatically changing our ability to prepare for and respond to extreme events, facilitating the management of crisis incidents [9]. Information systems and technologies contribute to a better communication and action in complex systems, by helping in disaster response and in collecting information, analyzing it, sharing it, and timely disseminating it to the people at risk. In particular, timely information sharing amongst emergency actors is critical in emergency response operations [10]. Several research projects have been devoted to emergency and disasters management to create modelling and simulation techniques and tools for the emergency management. Relevant examples are the dynamic and adaptive models for operational risk management [11], [12].

Information technology is enhancing disaster management and communications through tools such as computer networks, virtual reality, remote sensing, GIS, and decision support systems. During the mitigation and preparation phases of an emergency the use of satellite communications and spatial analysis systems can be extremely valuable [13]. In recent years, many web-based emergency response systems have been developed and several studies shown the great complexities surrounding the design of this kind of systems [14]. Often, developments in other areas are overlooked and resources are spent looking for a solution that has been already implemented and proved in other environments. An example of an IT system to manage emergency situations is the Global Disaster Information Network (GDIN www.state.gov/www/issues/ relief/gdin.html). More elaborated examples with complex architectures, integrating geographic information systems, spatial databases and the internet are described in [15] and [16]. In [17] a WebGIS is presented addressing risk management related issues, providing authenticated users access to queryable information, depending on their authorization level. Hence, they achieve the goal of having a platform accessible anywhere with an internet connection, and multiple levels of access to different hierarchic roles. This is a similar approach to the one presented herein, except an existing tool has been adapted to the use-case, when compared with a tailored-made solution.

Building on these experiences and in the scope of several projects (INTERREG SPRES; FP7 Prepared), LNEC has been developing and applying a suite of Web platforms denoted as WIFF - Water Information Forecast Framework [7] to provide access to real time information to decision makers. These platforms were conceived for a single type of users and to provide full access to real time sensor data and model predictions, constituting at the same time a repository of past information, being available at each deployment site to the relevant end-users. For the SPRES platform, real time products were integrated with emergency planning information (hazard, vulnerability and risk maps as well as mitigation action sheets) to constitute a one-stop-shop for all data relevant to oil spill prevention and mitigation [7].

These platforms take advantage of novel technologies to provide on-line, intuitive and geographically-referenced access to real-time data and model predictions, and to produce on-demand services in support of routine management of coastal resources and harbor operations. Technology support include a) Drupal, a PHP-based Content Management System, to access model metadata, status and products, b) map server support (Geoserver) providing Web Map Services (WMS) to allow for geospatial placement of monitoring and forecast products, and model output query capabilities, and c) Flex, using the OpenScales library to handle geospatial information, for the WebGIS development.

However, the need for interaction between the multiple emergency actors and to have fast access to real time information (of both conventional data streams and on-the-fly in-the-field information during flood emergencies) from several users simultaneously, raises new requirements for these new platforms. Additionally, the new system should be cross-platform, i.e., to be built in a way that it is automatically and transparently adaptable to any device with a data connection, providing access to emergency information anywhere.

## III. Concept and implementation of the solution

The main goal for the platform described herein, is to provide a quick and responsive tool for flood risk forecast and assessment. End-users should be able to access information on the platform with no hassle and in any device from anywhere, providing that a data connection is available.

Moreover, since we are working directly with the civil protection agents as project partners, there is a major focus on developing on important issues for them. Specifically, our platform aims at fulfilling ease on usability and providing tools for quick decision taking by them, providing a product tailored for their needs. This tool is also being developed with modularity and reusability in mind, so that very little modifications to the platform code need to be made if the forecast product changes from "water levels" to other variables of interest. To achieve this, back and front-end are bind in such a way that the former provides access to the data while the front-end consumes (via a set of REST services) and displays the products without being content-aware, i.e., it shows data without considering data types. Since the front-end is not content-aware, a new instance of the platform, for visualizing and analyzing other types of data, can be created simply by changing the products made available by the back-end and performing basic adaptations to the front-end. Moreover, since the back-end serves data in a standard way (REST services), the front-end itself may also be substituted by another consuming service, be it a mobile application, a web-page or other data consuming service. This allows for interoperability between platforms, allowing other users to use these services straightforwardly, after being successfully authenticated and authorized.

This flow of information is illustrated in Fig. 1, where a separation of concepts, between back-end and front-end is clearly visible. The back-end consists of an instance of CakePHP, a MVC PHP framework, with a PostgreSQL storage database (with PostGIS extension), coupled with several instances of Geoserver and Perl and Python scripts.

The code developed with CakePHP handles user control and user accesses, based on access control lists and roles, and streamlines access to the database via REST requests from the front-end. Geoserver, an open source server for geospatial data sharing, manages the georeferenced imagery (both in raster and vector formats) and serves them using open standards, such as WMS. This also promotes interoperability by allowing that different systems exchange data with each-other using known standards. Geoserver uses data, both from the PostgreSQL/PostGIS database, but also from results produced by the flood prediction models, in the form of shapefiles, allowing for model results probing directly on the data served through the UI.

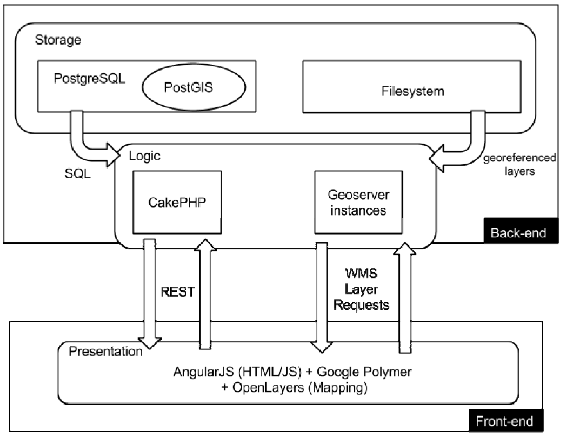


Fig. 1 Technological Architecture of the solution

Since a huge amount of requests are made to Geoserver, and it lags when cluttered with data, some maintenance must be performed on a recurrent basis. Instead of keeping all the data content in Geoserver' memory, existing content is moved to a secondary location on the filesystem daily through a script running as a cronjob. Reloading the products to Geoserver is performed on-demand by the user through a set of Perl and Python scripts, initiated by client request performed on the front-end. This strategy allows for a considerable performance increase in data access and gives the best priority of access to today's flood predictions.

The front-end consists of a single-page responsive web application which allows users to visualize all the products served by the back-end, in an intuitive interface available for multiple devices. With the goal of ubiquitousness in mind, less-intensive technologies were chosen: 1) HTML5 and CSS3, as the building base of all web-applications, 2) AngularJS, a javascript framework that offers dynamic templating and two-way databinding, 3) Google Polymer, a Google design specification implementation library, and 4) OpenLayers, a library for handling geospatial data and mapping tools on the client-side. Although some technologies, like AngularJS and Google Polymer, are still in a beta stage, they are supported and maintained by Google and have a huge community contributing for their development (AngularJS has over 6K commits on github, more than older and well-known javascript libraries such as JQuery with less than 6K commits). Using these technologies appears thus a good choice for future developments, since they shape the way we build the web (www.polymer-project.org) and fully fulfill the user requirements.

As referred before, the front-end maintains communication with the back-end via a set of REST services made available by the back-end. Responses to these request come in JSON, a lightweight data-interchange format easy to read/write by humans and to parse/create by machines.

The server-side information flow system has been planned and built to easily gather real-time data from the wireless sensor network, to parse the relevant information and store it in a persistent database system, to use the measurements to run forecasting models, and to provide the forecast results to the end-user. A detailed description of this flow of information can be seen in Fig. 2. From top to bottom, this system building blocks are 1) a wireless sensor network (WSN), comprised of several sensor nodes; 2) a data gathering server, which communicates with the WSN gathering and parsing real-time data; 3) prediction model instance and corresponding redundancy instance (to guarantee a fallback), that produce forecasting results with real-time data as one of the inputs; 4) several instances of Geoserver (again to guarantee a fallback, but also to guarantee data availability for a great amount of web-requests), that start by consuming forecast results in the form of visual imagery which later renders to the end-user in the form of WMS layers. All the instances of Geoserver are managed by a load-balancer that decides which instance should handle the client requests. The WSN consists of several real-time station nodes scattered in the domain of interest, which record water level data and transmit it to a central server. This central server is basically a set of scripts that trigger the transmission via either GPRS into an FTP server or Circuit Switched Data (CSD) directly into the file system. After the transmission is performed successfully, data is handled by the central-server to parse and store it, and to create input files for the prediction models. On their side, prediction models produce water level forecast results for the next 48 hours which are then consumed by the geographical web server. Geoserver accepts both raster images (geotiff) and vector files (shapefiles), but can also produce images from Postgis database data. These images are georeferenced and then presented in the user interface on a layer map.

## IV. Application to the Case Study

The flexibility and usefulness of the platform is illustrated here, applied to the MOLINES case study, mainly through examples of the functionalities available on the interface. For this project, the requirements were the following: 1) the platform would account for different user roles, providing differentiated access to dedicated products; 2) the platform should be able to host georeferenced products from the static risk analysis (hazard, vulnerability) and the dynamic real time forecast; 3) the platform should be agile, providing fast access to the alerts and their products; and 4) the platform should be prepared to incorporate and show in a georeferenced way the information uploaded by the civil protection agents in the field during emergencies.

This application user interface is composed of a top header with title, a sidebar for displaying the various links to the main functionalities offered and a detailed content area. The sidebar hides when opening the application in smaller devices (smartphones or tablets), and is accessible by a button on the top header, as seen in Fig. 3. This allows the main content to be shown in full screen, taking advantage of all available space on the device screen.

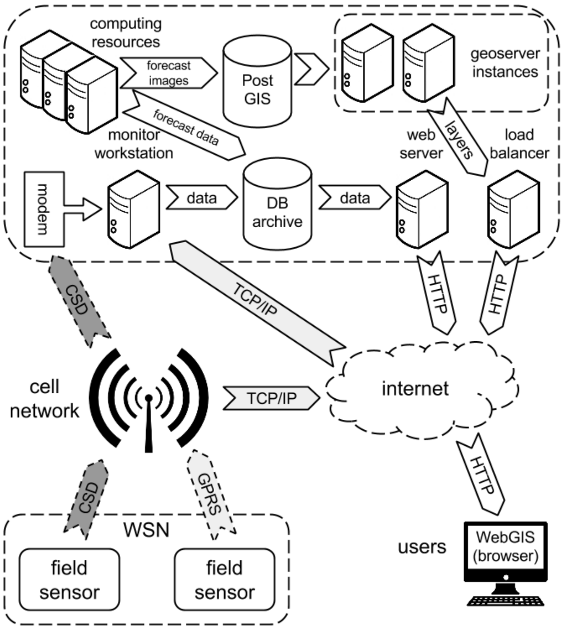


Fig. 2 Network information data flow

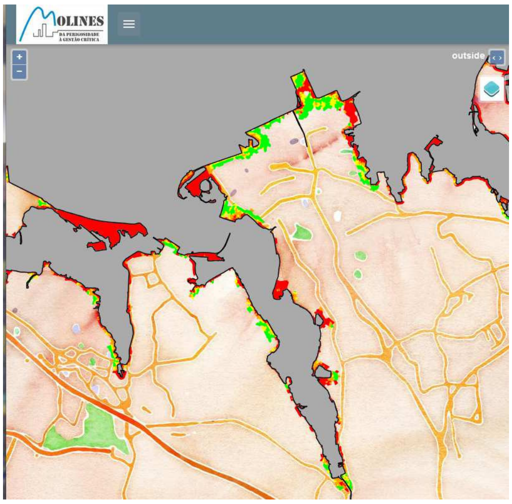
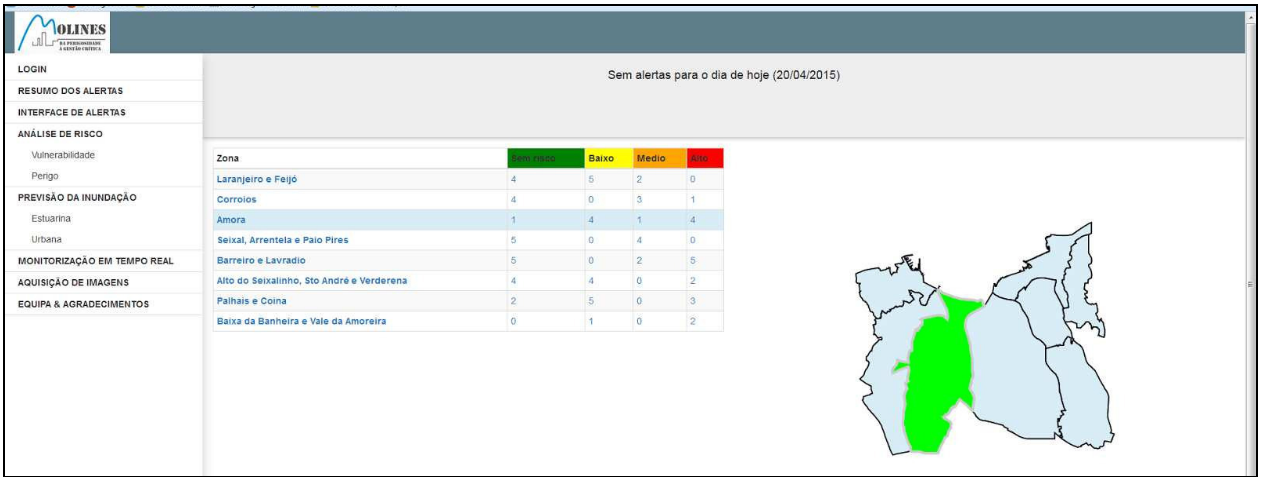


Fig. 3 Example of the interface adapted to a mobile device

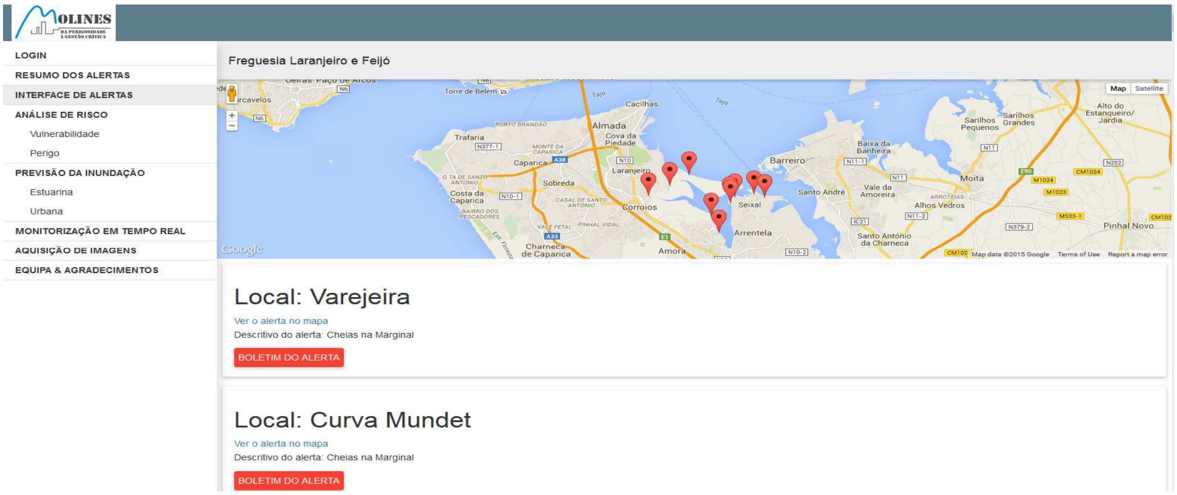
Since this is a platform for support of detailed risk management, it must provide the users with multiple ways to access all the risk events detected: a quick way, through both geographical representation of a specific area and a short list of areas, and a very detailed listing, supported by clear identification of the most vulnerable locations. This functionality can be seen in a) Fig. 4: a summary of study zones and risk alerts triggered for each zone as well as their risk type, b) Fig. 5: a detailed listing of locations at risk and the alert bulletins. Also, when hovering over the zones on the table in Fig. 4, the corresponding zone on the map gets highlighted for a better notion of where this zone is and provides a direct link to the detailed listing of Fig. 5.
Similarly, each zone name is a link for the more detailed description of the risk events occurring at that specific zone.



Sem alertas para o dia de hoje (20/04/2015)

| Zona | Sem risco | Baixo | Medio | Alto |
|---|---|---|---|---|
| Laranjeiro e Feijó | 4 | 5 | 2 | 0 |
| Corroios | 4 | 0 | 3 | 1 |
| Amora | 1 | 4 | 1 | 4 |
| Seixal, Arrentela e Paio Pires | 5 | 0 | 4 | 0 |
| Barreiro e Lavradio | 5 | 0 | 2 | 5 |
| Alto do Seixalinho, Sto André e Verderena | 4 | 4 | 0 | 2 |
| Palhais e Coina | 2 | 5 | 0 | 3 |
| Baixa da Banheira e Vale da Amoreira | 0 | 1 | 0 | 2 |

Fig. 4 Summary of risk events



Freguesia Laranjeiro e Feijó

Local: Varejeira

Ver o alerta no mapa

Descritivo do alerta: Cheias na Marginal

BOLETIM DO ALERTA

Local: Curva Mundet

Ver o alerta no mapa

Descritivo do alerta: Cheias na Marginal

BOLETIM DO ALERTA

Fig. 5 Detailed location of areas at risk

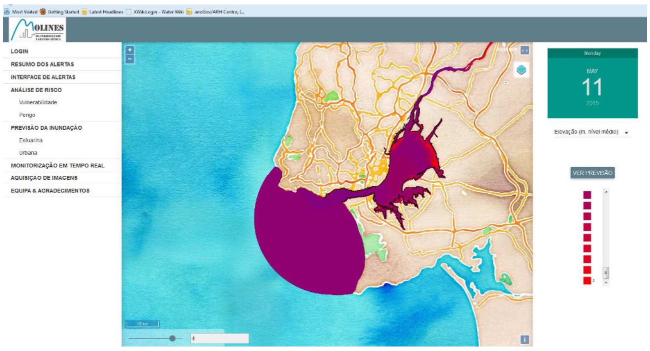


Fig. 6 Water level forecast

Other functionalities include a section to provide access to the real-time data gathered from WSN, which also allows an automatic comparison with model forecast results. This functionality allows the end-users to access data being measured at a point of interest but also to validate model predictions with real-time data. Indeed, this comparison is fundamental for the emergency agents and other decision-makers to provide them a measure of reliability on the model predictions and confidence on the actions to be promoted in the field.

## V. Discussion and Future Work

Herein, an interactive, flexible and multiple user roles Web platform is presented, which takes advantage of novel technologies to provide fast access to all relevant online, intuitive and geographically referenced real-time data and model predictions for urban and estuarine floods.

Future work on the platform is planned to further improve its performance. For instance, one of the strategies includes using GeoJSON, an open standard format for encoding geographical information features using Javascript Object Notation, on the client-side to render georeferenced data layers on top of maps, instead of depending on Geoserver to serve those layers. This would put the workload on the client-side instead of the server-side, which would produce faster results overall.

At the same time, the overall goal of these IT platforms is to have them integrated in the everyday's workflow of end-users. Their operation requires considerable computational efforts that may not available in many decision-makers IT infrastructures. As LNEC's computational resources are finite and often related with on-going scientific projects, a solution should be looked up to provide the best solution for end-users. In order to allow for stakeholders to run their own prediction model instances and operate these IT platforms, future work also includes the creation of prediction model deployments on the cloud.

For the MOLINES application, the future work will be concentrated on the integration, in the interface, of the uploaded data provided by the agents in the field. Challenges are the automatic check on the reliability of this information and the way to integrate them in the interface in a simple and easy to probe manner. Other add-ons include also the issuing of the alert based on the model predictions.

## VI. Acknowledgment

The authors thank Paula Freire and Ana Rilo, from LNEC, Alexandre Tavares and Pedro Santos, from CES, and Luis Sá, from the Portuguese Civil Protection, for helping with the definition of the platform requirements. They also thank the SCHISM community for making their model code available at http://ccrm.vims.edu/schism/.

## References

[1] Townend, I. and Pethick, J. (2002) Estuarine flooding and managed retreat. Society (2002), Volume: 360, Issue: 1796, pp. 1477-1495. DOI: 10.1098/rsta.2002.1011

[2] Ugarelli, R., Leitão, J.P., Almeida, M.C., Bruaset, S. (2011) Overview of climate change effects which may impact the urban water cycle (PREPARED 2011.011 report). PREPARED: enabling change project.

[3] Jesus, G., Oliveira, A., Santos, M.A. and Palha-Fernandes, J. (2010) – Development of a Dam-break Flood Emergency System, Proceedings of the 7th International ISCRAM Conference, pp. 5.

[4] Pradhan, A.R., D.F. Laefer and W.J. Rasdorf (2007). Infrastructure Management Information system framework requirements for disasters, Journal of Computing in Civil Engineering, March-April, pp. 90-101. DOI: 10.1061/(ASCE)0887-3801(2007)21:2(90)

[5] Carracedo, P., et al. (2006) Improvement of pollutant drift forecast system applied to the Prestige oil spills in Galicia Coast (NW of Spain): Development of an operational system. Marine Pollution Bulletin 53(5-7) pp. 350-360. DOI:10.1016/j.marpolbul.2005.11.014

[6] Rodrigues M., et al. (2013) Application of an estuarine and coastal nowcast-forecast information system to the Tagus estuary. Proceedings of the 6th SCACR – International Short Course/Conference on Applied Coastal Research (Lisboa, Portugal), pp.10.

[7] Oliveira, A., et al. (2014) An interactive WebGIS observatory platform for enhanced support of coastal management. Journal of Coastal Research, Special Issue No. 66, , ISSN 0749-0208, pp. 507-512

[8] Deng Z. Q. Namwamba, F., Zhang, Z.H. (2014) Development of decision support system for managing and using recreational beaches, Journal of Hydroinformatics, Volume: 16, Issue: 2, pp. 447-457. DOI:10.2166/hydro.2013.185

[9] Rinaldi, S., Peerenboom, J., and Kelly, T. (2001) Complexities in Identifying, Understanding, and Analyzing Critical Infrastructure Interdependencies, IEEE Control Systems Magazine, pp. 11-25. DOI: 10.1109/37.969131

[10] Corbacioglu, S. and Kapucu, N. (2006) Organizational Learning and Self-adaptation in Dynamic Disaster Environments, Disasters, 30, 2, pp. 212-233. DOI: 10.1111/j.0361-3666.2006.00316.x

[11] Beroggi, G.E.G. and Wallace, W.A. (1994) Operational Risk Management: A New Paradigm for Decision Making. IEEE Transactions on Systems, Man, and Cybernetics, 24, pp. 1450-1457. DOI: 10.1109/21.310528

[12] Beroggi, G.E.G. and Wallace, W.A. (2000) Multi-Expert Operational Risk Management, IEEE Transactions on Systems, Man and Cybernetics Part C 30, pp. 32-44. DOI: 10.1109/5326.827452

[13] Marincioni, F. (2007) Information Technologies and The Sharing of Disaster Knowledge: The Critical Role of Professional Culture, Disasters, 31, 4, pp. 459–476. DOI: 10.1111/j.1467-7717.2007.01019.x

[14] Kyng, M., Nielsen, E. T. and Kristensen, M. (2006) Challenges in designing interactive systems for emergency response. In Proceedings of the 6th ACM Conference on Designing interactive Systems ACM Press, pp. 301-310. DOI: 10.1145/1142405.1142450

[15] Herold, S., M. Sawada, B. Wellar. (2005) Integrating geographic information systems, spatial databases and the internet: a framework for disaster management. Proceedings of the 98 th Annual Canadian Institute of Geomatics Conference, pp. 13-15.

[16] Fahland, D., Gläber, T.M., Quilitz, B., Weibleder, S. & Leser, U., HUODINI – Flexible Information Integration for Disaster Management, Proceedings ISCRAM2007, 2007, pp. 255-138.

[17] Kulkarni, A. T., et al. (2014) "A web GIS based integrated flood assessment modeling tool for coastal urban watersheds." Computers & Geosciences 64: pp. 7-14. DOI: 10.1016/j.cageo.2013.11.002

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1177–1181

DOI: 10.15439/2015F228
ACSIS, Vol. 5

# Two approaches to dynamic power management in energy-aware computer networks - methodological considerations

Andrzej Karbowski
NASK, Research and Academic Computer Network
ul. Wąwozowa 18
02-796 Warszawa, Poland
and
Institute of Control and Computation Engineering
Warsaw University of Technology
ul. Nowowiejska 15/19
00-665 Warszawa, Poland
E-mail: A.Karbowski@elka.pw.edu.pl

Przemysław Jaskóła
NASK, Research and Academic Computer Network
ul. Wąwozowa 18
02-796 Warszawa, Poland
E-mail: pjaskola@nask.pl

***Abstract*—The paper compares two formulations of dynamic power management in energy-aware computer networks. In the first approach the only criterion is energy consumption, in the second there is an additional one - the quality of service. It is shown, that the second approach is appropriate when the routing problem with fixed demands is inadmissible. Fortunately, by some optimization modeling transformations it still allows for using the same standard mixed integer solvers as the first approach.**

## I. Introduction

METHODS for increasing energy efficiency of computer networks gained much attention last years. The reason is, that we are witnessing a rise of energy costs, customer increase, more on-demand services using cloud architectures, mobile Internet, a diffusion of broadband access and a growing number of services offered by Internet service providers. The growth of the energy consumption by network infrastructure may be well illustrated by the overall energy requirements of European Internet operators: in 2005 they needed 14 TWh, in 2010 - 21 TWh, and the forecast for 2020 is 36 TWh [1].

At the same time the capacity surplus becomes a standard in almost all networks. Consequently, so-called green network technologies are quickly becoming a high-priority issue for the Internet [1], [2].

In the European Union an additional motivation is 2020 Energy Strategy, which, among other goals, assumes achieving by 2020 a 20% improvement in energy efficiency [3].

Efforts to reduce power consumption in telecommunication networks follow in two mutually related directions – design of a more efficient equipment and development of energy-aware network control strategies and protocols. Initial efforts were aimed at assessment of energy characteristics of network equipment and building elementary models [4]. Upon this knowledge some local, i.e., concerning single device, strategies were built – see e.g., [5].

However, it is possible to save even more energy by employing network-wide solutions.

The paper [6] presented a model of energy-aware router, an architecture of a control framework and various formulations of a network-wide energy saving optimization problem. They start from the exact mixed integer linear programming (MILP) formulation, which is aimed at solving the problem of a minimum energy routing. The objective is the minimization of the total power utilized by network components while ensuring end-to-end Quality of Service (QoS). The basic link-node formulation (*LNb*) is a network management problem with binary decision variables describing full routing in a network and corresponding energy state assignments to all routers, line cards and communication ports. A more advanced version may be proposed, in which some parts of the network can be shifted to low energy mode as a result of the optimization algorithms, where both paths and flow rates are decision variables. It will exploit the fact, that Internet traffic used to be elastic in a large part, which means, that a quality of service is little aggravated by small deviations from assumed flow rate.

In this paper first a model with fixed flow rates inspired by the paper [6] will be shortly presented and assessed. Then, an improved version of it with flexible rates - in the authors' opinion much more practical - will be proposed and discussed.

## II. A model with given demands

A hierarchical network model proposed in [6] considers every single communication port $p \in \{1,\ldots,P\}$ of every line card of a router $r \in \{1,\ldots,R\}$. In our paper, for simplification, we do not consider individual cards of the router, because they do not bring anything into the model except additional summations.

Directed links connecting pairs of ports are denoted by $l \in \{1,\ldots,L\}$; any network component can operate in

978-83-60810-66-8/$25.00©2015, IEEE

$k \in \{1,\ldots,K\}$ energy states, but two ports connected by a link are in the same state. A demand $d \in \{1,\ldots,D\}$ is characterized by its source $s_d$ and the destination $t_d$ node (router) and the volume $V_d$.

The topology of the physical network is described by four matrices of binary indicators: $g_{pr}, a_{lp}, b_{lp}$, which indicate, whether, respectively: port $p$ belongs to the router $r$, link $l$ is incoming to the port $p$ and link $l$ is outgoing from the port $p$. If $l$ is a link outgoing from the port $p$, the link $\tilde{l}$ denotes its partner link in the edge going in the opposite direction, that is:

$$b_{lp} = 1 \Longleftrightarrow a_{\tilde{l}p} = 1 \tag{1}$$

We assume that the numbers $l$, $\tilde{l}$ have different parities (e.g., they are consecutive numbers). The decision variables are two vectors of binary indicators $x_p, z_r$ - whether the port $p$ or router $r$ is used for data transmission and two incidence matrices with elements: $y_{lk}$ - whether the link $l$ is in the state $k$ and $u_{dl}$ - whether the demand $d$ uses the link $l$.

The full optimization problem is as follows:

$$\min_{\substack{x_p, y_{lk}, z_r, u_{dl} \\ p\in\overline{1,P}, l\in\overline{1,L}, \\ k\in\overline{1,K}, r\in\overline{1,R}, \\ d\in\overline{1,D}}} \Bigg[ F_{LNb} = \sum_{l=1,3,5,\ldots}^{L-1} \sum_{k=1}^{K} \xi_{lk} y_{lk} + \\ + \sum_{p=1}^{P} W_p x_p + \sum_{r=1}^{R} T_r z_r \Bigg] \tag{2}$$

subject to the constraints:

$$\forall_{\substack{d=1,\ldots,D, \\ p=1,\ldots,P}} \quad \sum_{l=1}^{L} a_{lp} u_{dl} \le x_p \tag{3}$$

$$\forall_{\substack{d=1,\ldots,D, \\ p=1,\ldots,P}} \quad \sum_{l=1}^{L} b_{lp} u_{dl} \le x_p \tag{4}$$

$$\forall_{\substack{r=1,\ldots,R, \\ p=1,\ldots,P}} \quad g_{pr} x_p \le z_r \tag{5}$$

$$\forall_{l=1,\ldots,L} \quad \sum_{k=1}^{K} y_{lk} \le 1 \tag{6}$$

$$\forall_{\substack{d=1,\ldots,D, \\ r=1,\ldots,R}} \sum_{p=1}^{P}\sum_{l=1}^{L} g_{pr} a_{lp} u_{dl} - \sum_{p=1}^{P}\sum_{l=1}^{L} g_{pr} b_{lp} u_{dl} = \\ = \begin{cases} -1 & r = s_d \\ 1 & r = t_d \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$\sum_{d=1}^{D} V_d u_{dl} \le \sum_{k=1}^{K} M_{lk} y_{lk}, \ l = 1,3,\ldots,L-1 \tag{8}$$

$$\sum_{d=1}^{D} V_d u_{dl} \le \sum_{k=1}^{K} M_{\tilde{l}k} y_{\tilde{l}k}, \ l = 2,4,\ldots,L \tag{9}$$

$$x_p, z_r \in \{0,1\} \ p \in \overline{1,P}, r \in \overline{1,R}, \tag{10}$$

$$y_{lk}, u_{dl} \in \{0,1\} \ l \in \overline{1,L}, k \in \overline{1,K}, d \in \overline{1,D} \tag{11}$$

where $M_{lk}$ and $\xi_{lk}$ are, respectively, the capacity and the power consumption of the edge comprising links $l, \tilde{l}$ in the state $k$, and $W_p$ and $T_r$ are power cost coefficients of the port $p$ and the router $r$.

As in the paper [6] constraints (3)-(5) determine the number of ports and routers that are used for data transmission. The conditions (6) assure, that each edge can be in one energy-aware state. The constraints (7) are formulated according to 1st Kirchhoff's law applied to source, destination and transit nodes, and finally, the constraints (8),(9) assure, that the flow will not exceed the capacity of a given edge.

Any MILP algorithm can be used to solve (2)-(11) problem. Popular solvers such as CPLEX and Gurobi may be applied.

## III. Shortcomings of the *LNb* model and possibilities to overcome them

In the *LNb* model (2)-(11) the issue of determining the flow demand matrix, which is crucial for implementing the control system, is not addressed. The simplest approach is to measure the actual flow rates carried by the network (averaged over a selected timescale), and adopt it as optimization constraints. Then, this model can help to answer the question whether it is possible and worthwhile to temporarily reduce network capacity and maybe also reroute some paths in order to save some energy without sacrificing the QoS, that is maintaining the current values of the flow rates. It should work especially in the case, when the network is underutilized.

Shortcomings of the method may manifest themselves however in more complex scenarios, when the controlled network, or at least a part of it, works on its capacity limit. The measured traffic is no longer a good estimate of the demanded bandwidth, because flows are already truncated by the network. The demand matrix must be provided in some other way, e.g., by predictive model or educated guess.

It is then crucial, that demands, coupled through link capacity constraints, do not exceed the bandwidth offered by the network in any spot, which can be hard to assure in nontrivial network topologies. Otherwise, a feasible solution will not exist.

In such cases, a modified approach can be proposed, in which flow rates are represented by variables rather than constants. It exploits the fact, that Internet traffic used to be elastic in a large part, which means, that the QoS is only a little aggravated by small deviations from the assumed flow rate. The combined routing and rate control problem has to be solved, which leads to the solution feasible in terms of the formulated model, even when the traffic demand is greater than the capacity offered by the network.

Moreover, in some cases a minor reduction of flow rates, which is accepted by the comprehensive model taking into account the elasticity of a demand, may allow to accommodate the traffic in a smaller number of links, thus allowing for further great reduction of power consumption.

## IV. A GENERALIZATION – TWO-CRITERIA ROUTING PROBLEM

The most important modification of the *LNb* model consists in relaxing flow rates - from now on they are variables denoted by $v_d$. In a consequence, the objective function $F_{LNb}$ (2) has to be augmented with a QoS related criterion $Q_d$, which represents a penalty for not achieving the assumed flow rate $V_d$ by the flow $d$. $Q_d(v_d)$ is a convex and continuous function, decreasing on interval $[0, V_d]$. It is reaching minimum (zero) at $V_d$, the point in which user expectations are fully satisfied. The convexity of $Q_d(v_d)$ is associated with the conviction, that small deviations from the nominal throughput $\Delta = V_d - v_d$ are neglected by network users, while large deviations are noticed and should be avoided. Moreover, since $Q_d(v_d)$ is monotonically decreasing, it assures that the slope of the curve becomes steeper, as the rate $v_d$ approaches zero.

A two criteria - i.e., reflecting energy costs and QoS - mixed integer network problem of simultaneous optimal bandwidth allocation and routing may be formulated in the following way:

$$\min_{\substack{x_p, y_{lk}, z_r, u_{dl}, v_d, \\ p\in\overline{1,P}, l\in\overline{1,L}, k\in\overline{1,K} \\ r\in\overline{1,R}, d\in\overline{1,D}}} \Bigg\{ F_{2C} = \alpha F_{LNb} + (1-\alpha) \sum_{d=1}^{D} Q_d(v_d) =$$

$$= \alpha \left[ \sum_{l=1,3,5,\ldots}^{L-1} \sum_{k=1}^{K} \xi_{lk} y_{lk} + \sum_{p=1}^{P} W_p x_p + \sum_{r=1}^{R} T_r z_r \right] +$$

$$+(1-\alpha) \sum_{d=1}^{D} Q_d(v_d) \Bigg\} \tag{12}$$

subject to constraints (3)-(7), (10)-(11) from *LNb* and

$$\sum_{d=1}^{D} v_d u_{dl} \leq \sum_{k=1}^{K} M_{lk} y_{lk},\ l = 1, 3, \ldots, L-1 \tag{13}$$

$$\sum_{d=1}^{D} v_d u_{dl} \leq \sum_{k=1}^{K} M_{\tilde{l}k} y_{\tilde{l}k},\ l = 2, 4, \ldots, L \tag{14}$$

$$0 \leq v_d \leq V_d,\ d \in \overline{1,D} \tag{15}$$

The parameter $\alpha \in [0, 1]$ is a scalarizing weight coefficient, which can be altered to emphasize any of the objectives.

Such a problem in the particular case when $\alpha = 0$, i.e., without the energy components, was first addressed by Jaskóła and Malinowski [7] and independently by Wang et al. [8]. A decomposed algorithm to solve it was proposed by Karbowski [9].

A two-criteria optimal routing and bandwidth allocation problem, taking into account the energy component, for a completely different network and cost model than that of Section II was presented in [10].

In general, the formulation (12)-(15), (3)-(7), (10)-(11) has some drawbacks: it defines a mixed-integer nonlinear programming problem with nonconvex, bilinear link capacity constraints (13),(14). At present the leading solvers - e.g., CPLEX, Gurobi - can solve efficiently convex quadratic mixed-integer quadratically constrained problems MIQCP, with positive semidefinite matrices of constraints quadratic forms, which is not the case of (13),(14) constraints. The general nonlinear, mixed-integer, nonconvex solvers are very slow.

Fortunately, the problem (12)-(15), (3)-(7), (10)-(11) can be quite easily transformed to the form accepted by fast mixed-integer solvers, what we describe below. The reformulation of nonlinear network problems to the input format of standard MILP or MIQP solvers seems to be a very promissing approach nowadays [11].

## V. ELIMINATION OF THE NONLINEARITY FROM CONSTRAINTS

From the QoS components of the objective function $F_{2C}$ (12) it is usually expected, that they assure so-called proportional-fairness of the allocations of the bandwidth, when the network is subject to a congestion [12]. Quadratic functions may be used to achieve it [13] (unfortunately, linear - not), so the objective function $F_{2C}$ can be quadratic and convex.

The only problem that still remains to solve is nonconvex nonlinearity of the constraints (13), (14). It can be eliminated by a transformation proposed in [14].

It consists in the introduction of auxiliary variables $w_{dl} = v_d u_{dl}$, $d \in \overline{1,D}$, $l \in \overline{1,L}$ (denoting the part of a traffic rate in the link $l$ assigned to the flow $d$) and the substitution of these inequalities with subsequent set of linear inequalities:

$$\forall_{l=1,3,\ldots,L-1} \quad \sum_{d=1}^{D} w_{dl} \leq \sum_{k=1}^{K} M_{lk} y_{lk} \tag{16}$$

$$\forall_{l=2,4,\ldots,L} \quad \sum_{d=1}^{D} w_{dl} \leq \sum_{k=1}^{K} M_{\tilde{l}k} y_{\tilde{l}k} \tag{17}$$

$$\forall_{\substack{d=1,\ldots,D, \\ l=1,\ldots,L}} \quad w_{dl} \leq V_d u_{dl} \tag{18}$$

$$\forall_{\substack{d=1,\ldots,D, \\ l=1,\ldots,L}} \quad w_{dl} \leq v_d \tag{19}$$

$$\forall_{\substack{d=1,\ldots,D, \\ l=1,\ldots,L}} \quad w_{dl} \geq v_d - V_d(1 - u_{dl}) \tag{20}$$

$$\forall_{\substack{d=1,\ldots,D, \\ l=1,\ldots,L}} \quad w_{dl} \geq 0 \tag{21}$$

## VI. THE FINAL FORMULATION OF THE PROBLEM

Summing up, the final formulation of our two criteria energy-aware integrated routing and flow control problem is as follows:

$$\min_{\substack{x_p, y_{lk}, z_r, u_{dl}, v_d, w_{dl} \\ p\in\overline{1,P}, l\in\overline{1,L}, k\in\overline{1,K} \\ r\in\overline{1,R}, d\in\overline{1,D}}} \Bigg\{ F_{2C} = \alpha F_{LNb} + (1-\alpha) \sum_{d=1}^{D} Q_d(v_d) =$$

$$= \alpha \left[ \sum_{l=1,3,5,\ldots}^{L-1} \sum_{k=1}^{K} \xi_{lk} y_{lk} + \sum_{p=1}^{P} W_p x_p + \sum_{r=1}^{R} T_r z_r \right] +$$

$$+(1-\alpha) \sum_{d=1}^{D} Q_d(v_d) \Bigg\} \qquad (22)$$

subject to the constraints:

$$\forall_{\substack{d=1,\ldots,D,\\ p=1,\ldots,P}} \quad \sum_{l=1}^{L} a_{lp} u_{dl} \leq x_p \qquad (23)$$

$$\forall_{\substack{d=1,\ldots,D,\\ p=1,\ldots,P}} \quad \sum_{l=1}^{L} b_{lp} u_{dl} \leq x_p \qquad (24)$$

$$\forall_{\substack{r=1,\ldots,R,\\ p=1,\ldots,P}} \quad g_{pr} x_p \leq z_r, \qquad (25)$$

$$\forall_{l=1,\ldots,L} \quad \sum_{k=1}^{K} y_{lk} \leq 1 \qquad (26)$$

$$\forall_{\substack{d=1,\ldots,D,\\ r=1,\ldots,R}} \sum_{p=1}^{P} g_{pr} \sum_{l=1}^{L} \left(a_{lp} - b_{lp}\right) u_{dl} = \begin{cases} -1 & r = s_d \\ 1 & r = t_d \\ 0 & \text{otherwise} \end{cases} \qquad (27)$$

$$\sum_{d=1}^{D} w_{dl} \leq \sum_{k=1}^{K} M_{lk} y_{lk}, \; l = 1, 3, \ldots, L-1 \qquad (28)$$

$$\sum_{d=1}^{D} w_{dl} \leq \sum_{k=1}^{K} M_{\tilde{l}k} y_{\tilde{l}k}, \; l = 2, 4, \ldots, L \qquad (29)$$

$$\forall_{\substack{d=1,\ldots,D,\\ l=1,\ldots,L}} \quad w_{dl} \leq V_d u_{dl} \qquad (30)$$

$$\forall_{d=1,\ldots,D,} \quad w_{dl} \leq v_d \qquad (31)$$

$$\forall_{\substack{d=1,\ldots,D,\\ l=1,\ldots,L}} \quad w_{dl} \geq v_d - V_d(1 - u_{dl}) \qquad (32)$$

$$\forall_{\substack{d=1,\ldots,D,\\ l=1,\ldots,L}} \quad w_{dl} \geq 0 \qquad (33)$$

$$\forall_{d=1,\ldots,D} \quad 0 \leq v_d \leq V_d \qquad (34)$$

$$x_p, z_r \in \{0,1\} \;\; p \in \overline{1,P}, r \in \overline{1,R}, \qquad (35)$$

$$y_{lk}, u_{dl} \in \{0,1\} \;\; l \in \overline{1,L}, k \in \overline{1,K}, d \in \overline{1,D} \qquad (36)$$

When QoS components $Q_d(v_d)$ are quadratic and convex, the obtained mixed-integer quadratic problem can be solved by effective MILP/MIQP solvers, such as CPLEX, Gurobi.

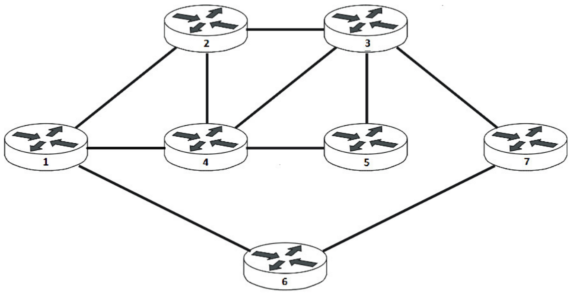


Fig. 1. Test network

## VII. Numerical evaluation

The problem (22)- (36) was formulated, implemented and solved with the help of the CPLEX solver for the test network presented in Fig. 1, with: $R = 7$, $L = 20$, $P = 20$. Every edge between two routers had two unidirectional links. We performed the experiments for three demands ($D = 3$) : $1 \rightarrow 7, 1 \rightarrow 7, 5 \rightarrow 6$. It was possible for each link to operate in five energy-aware states ($K = 5$). As their (penalty for not achieving) QoS functions we took [7]:

$$Q_d(v_d) = \frac{1}{2} V_d^2 - v_d \cdot (V_d - v_d/2) \qquad (37)$$

The throughput of a given link $l \in \overline{1,L}$ and the power consumption in energy-aware state $k \in \overline{1,K}$ were as follows: ($M_{l1} = 20$, $\xi_{l1} = 15$), ($M_{l2} = 40$, $\xi_{l2} = 30$), ($M_{l3} = 60$, $\xi_{l3} = 40$), ($M_{l4} = 80$, $\xi_{l4} = 60$), ($M_{l5} = 100$, $\xi_{l5} = 75$). As the power cost coefficients we took for ports $W_1 = W_2 = \ldots = W_{20} = 30$ and for routers $T_1 = T_2 = \ldots = T_7 = 300$. The maximum demand volumes were: $V_1 = 100, V_2 = 50, V_3 = 150$. The scalarizing coefficient was taken $\alpha = \frac{1}{2}$.

From calculations we got the following optimal paths: $d = 1,2 : 1 \rightarrow 6 \rightarrow 7$; $d = 3 : 5 \rightarrow 3 \rightarrow 7 \rightarrow 6$ and the objective function $F_{2C} = 1957.499221$. It means, that only routers $1, 3, 5, 6, 7$ were used and the links and ports connecting them, that is, the rest of the network was in the sleeping mode. The obtained optimal transmission rates were: $\hat{v}_1 = 75, \hat{v}_2 = 25, \hat{v}_3 = 100$. All the used links worked at the highest energy levels.

## VIII. Conclusions

We modified the dynamic power management of energy-aware computer networks model presented in [6] to capture these situations, when the users' demands are so high, that there is no admissible solution of the problem *LNb* (2)-(11). We suggest to use then a two-criteria model with flow rates as additional decision variables. When the QoS functions are quadratic and convex, it is possible to reformulate the problem in such a way, that the same standard solvers, e.g., CPLEX or Gurobi, can be used to find the solution. The resulting mixed-integer programming problem has more variables, but the new are only real, not binary, what should not influence too much the time of calculations.

The performed numerical test confirmed the appropriateness of the formulation.

## REFERENCES

[1] D.G. Recupero, “Toward a Green Internet", *Science*, vol. 339, 2013, pp. 1533–1534, http://dx.doi.org/10.1126/science.1235623

[2] R. Bolla, R. Bruschi, F. Cucchietti, and F. Davoli, “Setting the Course for a Green Internet", *Science*, vol. 342, 2013, pp. 1316, http://dx.doi.org/10.1126/science.342.6164.1316-a

[3] European Commission “Energy 2020 A Strategy For Competitive, Sustainable And Secure Energy", *Communication From The Commission To The European Parliament, The Council, The European Economic and Social Committee and The Committee Of The Regions*, COM(2010) 639, Brussels, 2010, http://dx.doi.org/10.2833/78930

[4] R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, “Energy efficiency in the Future Internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures", *IEEE Communications Surveys & Tutorials*, vol. 13, 2011, pp. 223–244, http://dx.doi.org/10.1109/SURV.2011.071410.00073

[5] S. Nedevschi, L. Popa, G. Iannacone, D. Wetherall, S. Ratnasamy, “Reducing network energy consumption via sleeping and rate adaptation", in *Proc. of 5th USENIX Symposium on Networked Systems Design and Implementation*, 2008, pp. 323–336.

[6] E. Niewiadomska-Szynkiewicz, A. Sikora, P. Arabas, M. Kamola, K. Malinowski, P. Jaskóła, and M. Marks, “Network-wide power management in computer networks", in *Proc. 22nd International Teletraffic Congress Specialist Seminar on Energy Efficient and Green Networking* (Best Paper Award), 2013, http://dx.doi.org/10.1109/SSEEGN.2013.6705398

[7] P. Jaskóła, and K. Malinowski, “Two methods of optimal bandwidth allocation in TCP/IP networks with QoS differentiation", in *Proc. Summer Simulation Multiconference (SPECTS'04)*, 2004, pp. 373–378.

[8] J. Wang, L. Li, S. H. Low, J. C. Doyle, “Cross-layer optimization in TCP/IP networks", *IEEE/ACM Transactions on Networking*, vol. 13, 2005, pp. 582–595, http://dx.doi.org/10.1109/TNET.2005.850219

[9] A. Karbowski, “Integrated routing and network flow control embracing two layers of TCP/IP networks - methodological issues", *Journal of Telecommunications and Information Technology*, 2012, pp. 51–54, http://www.itl.waw.pl/czasopisma/JTIT/2012/2/51.pdf

[10] P. Jaskóła, P. Arabas and and A. Karbowski, “Combined Calculation of Optimal Routing and Bandwidth Allocation in Energy Aware Networks", in *Proceedings of the 26th International Teletraffic Congress (ITC)*, 2014, http://dx.doi.org/10.1109/ITC.2014.6932962

[11] M. Tvorogova, “Efficient models for special types of non-linear maximum flow problems", in *Proc. Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2013, pp. 409–416.

[12] F. Kelly, “Charging and rate control for elastic traffic", *European Transactions on Telecommunications*, 1997, vol. 8, pp. 33–37, http://dx.doi.org/10.1002/ett.4460080106

[13] C. Touati, E. Altman, J. Galtier, “On fairness in bandwidth allocation", *Tech. Rep. 4269, Unité de recherche INRIA, Sophia Antipolis, France*, 2001.

[14] J. Bisschop, *AIMMS Optimization Modeling*, Paragon Decision Technology B.V., 2007.

DOI: 10.15439/2015F6
ACSIS, Vol. 5

# Analysis of notification methods with respect to mobile system characteristics

Piotr Nawrocki [*], Mikołaj Jakubowski [†] and Tomasz Godzik [‡]
[*]AGH University of Science and Technology,
al. A. Mickiewicza 30, 30-059 Krakow, Poland
e-mail:piotr.nawrocki@agh.edu.pl
[†]e-mail:mkl.jakubowski@gmail.com
[‡]e-mail:tomek.godzik@gmail.com

***Abstract*—Recently, there has been an increasing need for secure, efficient and simple notification methods for mobile systems. Such systems are meant to provide users with precise tools best suited for work or leisure environments and a lot of effort has been put into creating a multitude of mobile applications. However, not much research has been put at the same time into determining which of the available protocols are best suited for individual tasks. Here a number of basic notification methods are presented and tests are performed for the most promising ones. An attempt is made to determine which methods have the best throughput, latency, security and other characteristics. A comprehensive comparison is provided, which can be used to select the right method for a specific project. Finally, conclusions are provided and the results of all the tests conducted are discussed.**

## I. Introduction

THE PURPOSE of this paper is the analysis and tests of several selected notification methods for mobile platforms. The reason for this research is the need to determine the best way of sending simple as well as more advanced messages about the events involved in the operation of grid systems or telemetric networks. This makes it possible to use the optimal approach in numerous projects that need to inform users about their current status. This aspect is currently of utmost importance for the industry as such notification methods enable developers to engage users much more fully and keep them in constant contact with their leisure and work interests. These considerations have guided us throughout our research and affected all our decisions on the selection and ways of testing of the methods in question.

Several protocols and methods were considered based on their purposes and current industry standards. The main candidates were: CoAP (Constraint Application Protocol), XMPP (Extensible Messaging and Presence Protocol) and XMPP over SOAP (Simple Object Access Protocol), MQTT (Message Queuing Telemetry Transport), MQTT-SN (Message Queuing Telemetry Transport for Sensor Networks), AMQP (Advanced Message Queuing Protocol), Cloud notification systems (Google Cloud Messaging, Urban Airship), SMS (Short Message Service) and Restful HTTP (Hypertext Transfer Protocol).

Of course, these are not all the protocols that could be used for mobile notifications, but these listed appear to hold the most promise and therefore the purpose is to discern their usefulness in the best way possible.

In addition to the protocols and methods above, we investigated other solutions, such as the Apple push notification or Line application which, for various reasons, were not considered further. The Apple push notification technology is a good solution, but it is proprietary, i.e. limited to Apple devices and that is why we decided to test more universal solutions first. There are also solutions (applications) that use their own protocols. A good example is Line application, which uses a proprietary protocol. We considered testing this solution; however, there are significant difficulties with accessing the documentation for this protocol.

## II. Related Work

Mobile systems are a relatively new field of study and picking a specific topic such as comparing available notification methods does not return many related work results. Some of the protocols have been covered in separate articles and while these took the sending of notifications into account, tests were not always conducted in mobile systems.

The one available article [1] that compared notification methods only covered cloud systems [2] and applications. It discussed the following methods: C2DM (Google Cloud to Device Messaging—the predecessor to GCM), Xtify, XMPP and Urban Airship. As during our research that article was relatively new, one might think that the information contained there would still be relevant, but it turned out to be already out of date. Google has meanwhile redesigned and rebranded its notification system and Xtify was purchased by IBM. Only Urban Airship is still available on the market in the same configuration as previously. The article is more a comparison of available commercial products than a real world testing suite. As expected, the conclusion was that the fastest protocol of the four tested was XMPP, but it had a characteristic slightly different from the others.

Another article [3] only tested the MQTT protocol. The author believed that it was the best possible choice and only aimed to describe its main features and capabilities. Just a single simple test and its averaged results together with the amount of data transferred and power consumption over a period of time were provided. In the conclusion, the author

978-83-60810-66-8/$25.00©2015, IEEE

described the MQTT protocol as being both lightweight and perfect for mobile platforms.

In [4], the authors investigated XMPP in the field of collaborative applications. Its main purpose was to assess the usefulness of XMPP in exchanging location data between mobile clients and web servers. No testing was conducted, but a thorough description of XMPP and the Android platform was provided while also taking into account the ways of integrating them. The article described XMPP as a general purpose messaging protocol that is easily extensible.

An important aspects in the context of notification methods are SLA parameters [5] and the power consumption of battery-powered devices. In [6], the authors discussed the problem of sending notification data using GPRS connectivity from remote telemetry stations [7]. They proposed the concept of adaptive message aggregation which extends the MQTT-SN protocol, adjusting its behaviour to the GPRS (General Packet Radio Service) connectivity profile in order to decrease the energy consumption related to data transmission.

## III. Notification methods

The following section generally describes and analyses the possible notification methods for mobile devices mentioned in the introduction. As a result of this analysis, it was decided to select some of them in order to perform the thorough tests described later in this paper.

### A. SMS

It is possible to use the Short Message Service as a notification mechanism. An application would have to intercept the SMS messages received by an Android phone and analyse them to check whether they contain notifications from the system. One could just use simple text messages without a dedicated client application, but this would severely limit the functionality available to users.

This approach has several major issues. First, the cost of sending multiple messages to numerous clients could be immense. Secondly, it is not guaranteed that the message will be delivered on time or (sometimes) even on the same day. What is more, all text messages have a maximum undelivered period (which cannot exceed 7 days), and this means that some notifications would not be delivered at all.

### B. Google Cloud Messaging

In order to simplify the development of applications and to extend phone battery life, Google has created a simple built-in notification system for the Android platform, which only maintains a single connection at any time.

This approach has some obvious drawbacks. Firstly, the number of messages sent concurrently is limited to four per application and there is no guarantee that the message will be delivered, especially while the service is shared. Secondly, there is no specified maximum delay, which is not acceptable for most modern systems. Moreover, in posts like [8] it is claimed that the method is not all that well documented and it is not easy to make an application work reliably with Google Cloud Messaging (GCM). Another problem with GCM is that some people do not trust Google not to abuse its capabilities, citing privacy or security concerns. One must also keep in mind that GCM can be used by some malware applications as described in [9].

### C. Restful HTTP

Another possible solution would be to use a RESTful HTTP service based on a pull queue model [10]. Such an implementation would have to pull notifications from the server at certain intervals or when the user turns on the application. Currently creating such a service is a very simple process and does not require additional knowledge from most developers, which is the main advantage of this approach.

However, using this method is very inefficient as it is not clear at what intervals requests should be made. Using too long an interval between requests may result in multiple notifications being sent all at once, making the older messages meaningless. Conversely, if the interval were too short, it would use too much device resources. Moreover, much of the workload is shifted to the mobile device and the amount of data sent between server and client is sometimes doubled.

Some ideas for REST notification systems are discussed in [11], however using a pure REST approach is highly discouraged. Using the AMQP/REST mixed approach seems much more plausible.

### D. XMPP

XMPP is basically an open technology for real-time communication, using XML (Extensible Markup Language) as the base format for exchanging information. It was designed to be easily extensible and one of its main uses are publish-subscribe systems. It is the most mature protocol among all the solutions selected as it was already in use in 1999. Throughout its history it was used by companies such as Google in the Google Talk communicator, by Microsoft in Skype or by Facebook in WhatsApp Messenger.

The idea behind XMPP is similar to that of e-mail, with a distributed server network in which each and every server can create its own service. The XMPP standard enables message encryption and XML support allows for the use of such technologies as SOAP or EDI (Electronic Data Interchange).

A standard that is tightly coupled with XMPP is SOAP over XMPP, which can be tested using the same means, as sending a SOAP message is basically sending some content over XMPP, which provides effective and reliable messaging—both asynchronous and synchronous.

XMPP is a general purpose protocol that is easily extensible. It was only designed to meet mobile platform requirements and was not expected to outperform any other protocols. However, its flexibility makes it a choice worth considering. In [4], a few add-ons are mentioned like group chats or streaming services with a possibility to transfer files.

### E. SOAP over HTTP

SOAP is a lightweight protocol for message exchanges that is independent from the system platform programming language. Its specification does not define a specific transport layer protocol, but most implementations use HTTP. It is important to mention, however, that HTTP is of no use for asynchronous messaging and because of that SMTP is often used instead. The protocol makes it possible to send many short messages.

In the discussion on the use of SOAP in notification systems, the following solutions should be considered: polling, both endpoints having their SOA interfaces, using WS-notification and using the message queueing solution encapsulated in HTTP.

All the solutions above have been analysed and none of them are easy to adapt to the needs of mobile notification systems. The first solution requires the client to make requests at certain points in time, which generates a lot of unnecessary traffic and is quite resource-heavy on small devices. The second idea is better, but would not work for most mobile devices as not all requests would pass from the server to the device since such HTTP requests are often blocked. A good solution is to use WS-notification, but the problem with making requests from the server is still present. What is more, it is not a standard supported by all web servers. The final solution uses queueing, but it involves a lot of unnecessary technology, especially given that there are ready-made queueing mechanisms that do not have to be encapsulated in HTTP requests.

### *F. MQTT/MQTT-SN*

MQTT is a publish-subscribe lightweight messaging protocol based on TCP/IP. It was designed to be open, simple, lightweight and easy to implement, since it was intended to be used in constrained environments with limitations such as: expensive, low bandwidth, unreliable network, limited processor or memory resources.

The entire protocol is based upon a central message broker, which distributes messages published on a topic to all the interested consumers. The "MQ" part of the name comes from "Message Queueing", however this protocol does not support queuing by default. It has three types of quality of service for message delivery, which are "At most once", "At least once" and "Exactly once". It also has a mechanism that can be used to inform interested parties about an abnormal disconnection using the "Testament" and "Last Will" features.

What is interesting is the fact that MQTT has already been used in numerous applications. The first implementation of GCM(C2D)[1] used exactly this protocol. DeltaRail's latest version of their IECC (Integrated Electronic Control Centre) also uses MQTT for communications within their signalling system, which is covered in [12].

This standard was created by IBM and because of that fact the IBM MQTT client Java library was used for testing and the Mosquitto open source message broker for distributing messages. Mosquitto's simple construction allowed us to create a bash script sending a set number of messages. The Android client connects to the broker using the IBM library and is fed the messages sent by the script.

MQTT-SN is a variation of MQTT designed to be used in sensor networks. In particular, it is supposed to be lightweight and easily implementable on small devices (e.g. in non-TCP/IP[2] networks).

### *G. CoAP*

CoAP[3] is a specialised web transfer protocol for use with constrained nodes and networks based on UDP (User Datagram Protocol). Its main task is to allow for communication between small devices such as sensors, switches, etc. It was designed on the basis of HTTP in order to simplify its architecture and allow for multicast. It also provides a simple mapping between CoAP and HTTP, which can be used to create RESTful services. The messages are sent in a binary format and their size is limited by the maximum size of a datagram. Messages can be sent with acknowledgements or without them depending on the designer's needs. Although it is a relatively new standard, it already has some additional features proposed like "Observable", which makes it possible to notify all the clients subscribed about changes to the resource.

### *H. AMQP*

AMQP is an open standard application layer protocol for message-oriented middleware that uses a binary format to send its messages. It was designed to solve the problem of interoperability between heterogeneous systems and message brokers. It was first used in 2006 by JP Morgan. It offers both point-to-point and publish-subscribe messaging types.

The most important advantage of AMQP is the fact that it is independent of programming languages and platforms unlike most messaging standards, for example JMS (Java Message Service) [13]. Moreover, it offers several types of quality of service in terms of delivery guarantees; these types are at-most-once, at-least-once or exactly-once guarantees. It also allows to encrypt messages, which is important especially in the case of valuable scientific data. Currently it is a widely used standard and has a large number of implementing libraries like Apache Qpid, RabbitMQ [14] or StormMQ.

## IV. Tests

There are currently three main mobile operating systems (Android, iOS and Windows Phone) available on the market and numerous devices that support them. As it would be neither possible nor sensible to test each and every one of them, only one testing platform and device was chosen.

Google's Android system was selected as the mobile platform for testing purposes because of the considerable availability and open nature of the solution. All major protocols and methods selected have working implementations for this system.

[1] Android Developer Central - GCM Advanced Topic - http://developer.android.com/google/gcm/adv.html

[2] Transmission Control Protocol/Internet Protocol

[3] CoAP RFC 7252 - http://tools.ietf.org/html/rfc7252

As a mobile device the Nexus 5 (LG D821) was used with Android version 4.4.3 using the standard Dalvik engine. At this point Android Runtime was already available but it seemed not yet ready for serious testing. All data (from a mobile device) was transmitted using an HSPA technology (operator: T-Mobile).

In order to conduct all tests, a server platform was also needed, which consisted of an Asus laptop with the Intel i5-3320M processor and 8GB of RAM with Ubuntu 12.04LTS and Oracle Java 1.7.0_60 installed. All data (from a server platform) was transmitted over the Wi-Fi network using the 802.11g standard (54 Mbps).

For time-related test cases ClockSync[4] was used to synchronise with time servers on the mobile device. All applications launched their message connectors in separate threads. Services were not used so all memory usage diagrams show the combined values of the connector and activity screen.

All useful notification methods should meet most of the specifications listed below:

- the financial cost should be low—mostly for open source and university projects;
- it should be possible to transmit more than just simple text—to enable interaction between the application and the main system;
- energy and memory usage should be minimal—the solution is to be used on mobile devices;
- message contents should be secure—confidential information could be transmitted;
- minimal message loss—important data could be transmitted;
- minimal delay—fast interaction is sometimes needed.

As a result of analysing the notification methods available (see Section III) and taking the above assumptions into account, it was decided to test the following protocols: XMPP, SOAP, MQTT, CoAP and AMQP.

In order to conduct testing for each protocol, the following solutions were used:

- XMPP—in order to prepare the XMPP test, the Smack library was used to implement the mobile client. It has been ported to Android in a version called Asmack. ejabber was used as a message broker. The second client, which was sending messages to the mobile client, was implemented in Python using the SleekXMPP library. The mobile solution was plain and simple with its task limited to keeping an open connection to the broker.
- SOAP—an attempt was made to test a basic polling mechanism using a simple Python SOAP server[5] and a basic Android client[6]. It involved making a number of requests for stress testing and a single request to measure single message performance. After obtaining initial results this method was discarded as it was more than 10 times slower than any other and used a lot of resources for polling, which is unacceptable for mobile devices. It was decided to concentrate efforts on other solutions specifically intended for such devices. The results collected are shown together with the other protocols tested, but are not included in graph comparisons (in Figure I) as they were much worse than for any other test conducted.
- MQTT/MQTT-SN—a broker that works with the MQTT/MQTT-SN protocol is the RSMB (Really Small Message Broker) from IBM and while it is quite easy to find, locating an appropriate client library (especially for MQTT-SN) is much more difficult. The one that is available for MQTT-SN[7] is written in the C programming language so it was of no use whatsoever for Android devices. The project also appeared to have been abandoned (no recent contributions). A further search led to a library written in Python[8], which was then used to implement a client that would be run using the QPython interpreter. After a certain amount of research and testing a stage was reached where messages were delivered from the broker to the device but never reliably. Attempts to change QoS settings failed as it appeared that the client library implementation was not complete yet. For these reasons MQTT-SN was excluded from testing and only MQTT was tested. It appears that the protocol is not mature enough yet to be used on a larger scale and that it has no reliable or finished implementations.
- CoAP—there is a limited number of implementations. The main Java libraries are jCoAP and nCoAP. jCoAP is not up to date with the RFC (Requests for Comments) 7252 and therefore nCoAP was used which additionally implements the “Observable” feature. To test the protocol, a simple server was created with a time service and a mobile client that was sending GET requests to the server. At first it was intended to use the “Observable” feature, but during stress testing it turned out that messages cannot be sent too often using this implementation because of errors. As a consequence, although “Observable” can be quite useful, especially in mobile systems, it was decided to have each notification sent as an answer to a separate GET request.
- AMQP—RabbitMQ was selected, which is one of the best documented and popular libraries. For testing purposes a simple Python script was developed that can send a set number of simple messages containing timestamps and an Android client application. The client connects to the RabbitMQ message broker and then the Python script is used to send messages.

To assess the efficiency and usefulness of the notification methods selected, several test cases were created and run for each protocol. It is not claimed that it is a complete test suite,

[4] ClockSync - http://amip.tools-for.net/wiki/android/clocksync

[5] Python simple and lightweight SOAP Library (a.k.a. soap2py) - https://code.google.com/p/pysimplesoap/wiki/SoapServer

[6] A simple SOAP client for Android - https://code.google.com/p/droidsoapclient

[7] MQTT-SN client in C - https://github.com/njh/mqtt-sn-tools

[8] MQTT-SN client in Python - http://git.eclipse.org/c/mosquitto/org.eclipse.mosquitto.rsmb.git/tree/rsmb/src/MQTTSClient/Python

but rather preliminary testing. More work remains to be done in this field.

### A. Time per message

Each of the four applications designed to check how much time it takes to process a single message while stress testing was used with different numbers of concurrent messages sent. Set sizes of 10, 50, 100, 200, 300, 400 and 500 messages were chosen. Each message contained its timestamp in order to enable the calculation of exact delay. Figures 1 (nCoAP), 2 (MQTT), 3 (RabbitMQ), 4 (XMPP), 5 (SOAP over HTTP) show how much time it took to process a single message for different stress test set sizes.

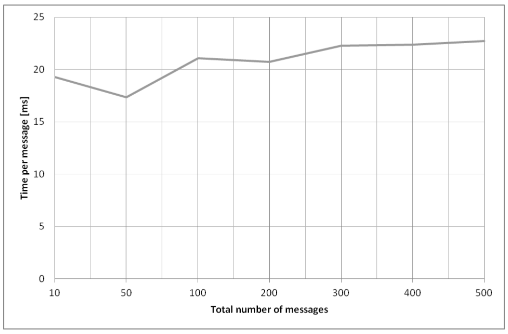


Fig. 1. nCoAP

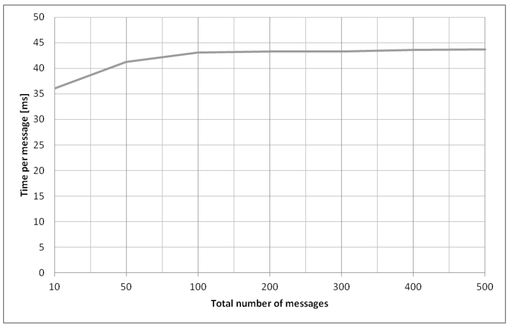


Fig. 2. MQTT

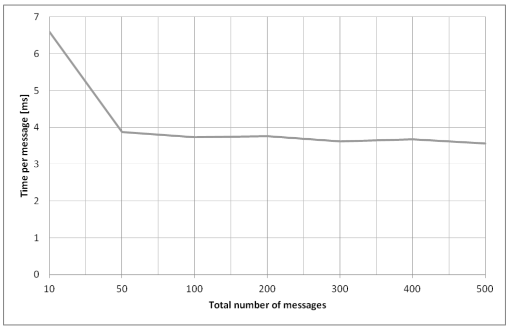


Fig. 3. RabbitMQ

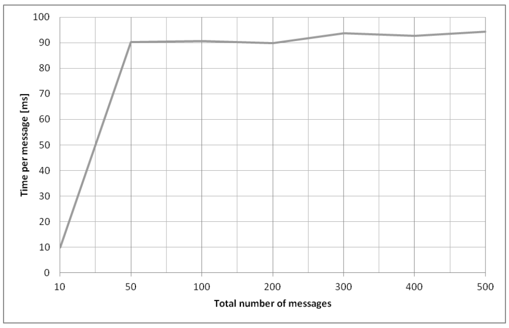


Fig. 4. XMPP

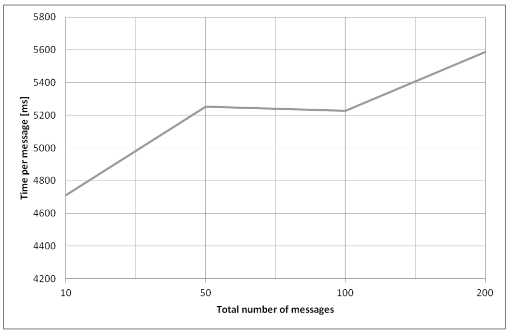


Fig. 5. SOAP over HTTP

Based on the graphs generated it can be stated that RabbitMQ is the fastest in terms of performance and its performance actually improves as more messages are sent concurrently. The nCoAP was also quite effective, but it should be kept in mind that each message was sent in response to a GET request, so it could be faster yet. Both MQTT and XMPP exhibit quite long message sending times. However SOAP over HTTP being definitely the slowest solution. In this test, RabbitMQ was the clear winner.

### B. Resource usage

The second most important criterion after performance was resource usage. It is crucial to use as little device resources as possible on a mobile platform in order to consume less power and allow for greater efficiency. In this section, peak memory usage (shown in Table I - "RAM (Random Access Memory) usage peak") and CPU (Central Processing Unit) power consumption (by using PowerTutor tool [15]) were measured as presented in Figures 6 (nCoAP), 7 (MQTT), 8 (RabbitMQ), 9 (XMPP), 10 (SOAP over HTTP), while sending 1000 messages concurrently to be processed by each of the mobile clients developed.

It is clearly visible that almost all protocols used similar amounts of memory, with the only outlier being MQTT with 10 MB less RAM usage than others. A much larger difference can be seen in power consumption levels. These seem to be strongly correlated with each individual protocol's processing time. nCoAP and RabbitMQ consumed the least power. About two times more power was consumed by XMPP and SOAP

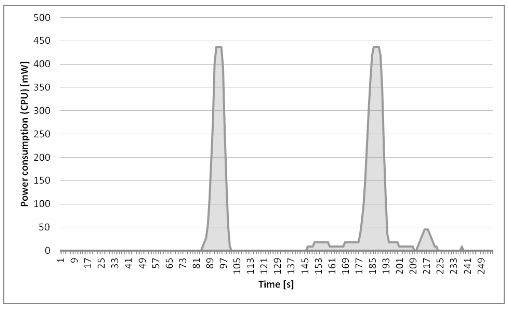


Fig. 6. nCoAP (2 consecutive runs shown)

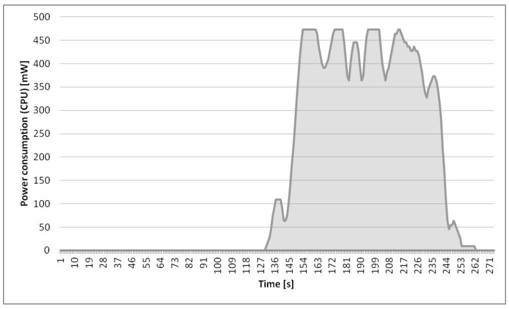


Fig. 7. MQTT

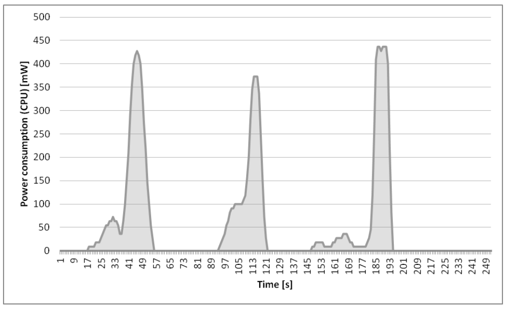


Fig. 8. RabbitMQ (3 consecutive runs shown)

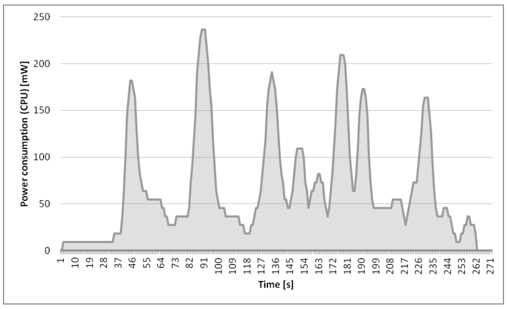


Fig. 9. XMPP

over HTTP. The worst result was achieved by the MQTT protocol.

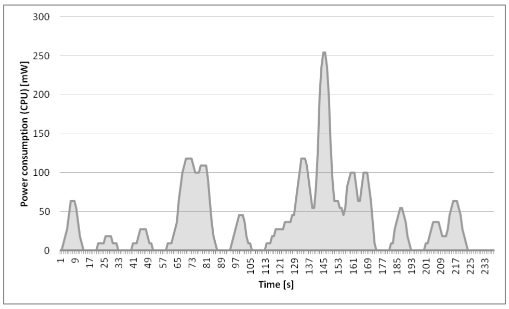


Fig. 10. SOAP over HTTP

### C. Reliability

Message sending reliability was also tested as it is one of the crucial issues to be tackled in mobile notification applications. Especially important is the issue of what happens to messages in a queue when the connection to the client is lost. This was simulated by reconnecting to a Wi-Fi network while sending a set of 1000 messages. All protocols were tested using default settings. It turned out that only nCoAP managed to deliver all messages, while the other protocols lost some or most of the messages sent. Developers must take care to use the correct settings for each protocol as QoS is not usually switched on by default.

### D. Ordering

This test case was meant to show whether protocols deliver messages in the same order in which they were sent. Similarly to the first test, a set of messages was sent containing timestamps and the comparison of arrival times of successive messages made it possible to determine whether they were correctly ordered. Only nCoAP changed the order of messages, which is most probably caused by using UDP. All other protocols delivered messages in the correct order even during high load.

### E. Average delay

The final test case was used to calculate the average delay when sending a single isolated message using each of the five protocols (Table I - “Average delay”). It turns out that nCoAP is the fastest when it comes to sending individual messages and SOAP over HTTP is the slowest one among all the protocols tested.

## V. Conclusions

The results as shown in Figure 11 clearly demonstrate that in terms of the maximum number of messages delivered per second RabbitMQ is the leader; however, when it comes to minimal delay, nCoAP tends to be able to deliver single messages much quicker. This means that if large numbers of notifications are to be sent, RabbitMQ could be used, while in a sparse notification system CoAP should perhaps be recommended. In the power consumption test the best results were also achieved by RabbitMQ and nCoAP.

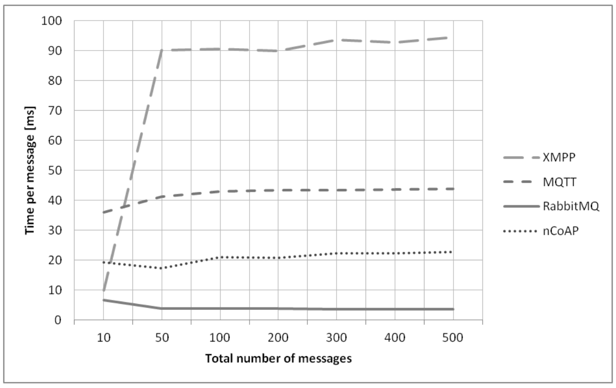


Fig. 11. Comparison of protocols

It should also be noted that not all protocols are able to easily pass through firewalls and NATs like XMPP, which is the most mature of all the protocols tested. When it comes to RAM usage, MQTT turned out to require the least megabytes, which can be of great importance on mobile devices. A summary of test results is shown in Table I.

In conclusion, the most promising solution seems to be RabbitMQ but none of the protocols proposed outperformed all the others in all test cases. This test suite only demonstrates the performance of some implementations currently available and the results might change for future releases or different platforms. Before any protocol is selected, it is important to specify the needs of the project in question and then compare the protocols to determine which one best suits these needs.

TABLE I
OVERVIEW OF PROTOCOL PROPERTIES

| | nCoAP | RabbitMQ | MQTT | XMPP | SOAP |
|---|---|---|---|---|---|
| RAM usage peak [MB] | 47 | 44 | 35 | 46 | 38 |
| Average delay [ms] | 91 | 185.5 | 339.6 | 192.3 | 972.2 |
| Ordered | no | yes | yes | yes | yes |
| Lost messages | no | yes | yes | yes | yes |
| Content | binary | binary | binary | text | text |
| Based on | UDP | TCP | TCP | TCP | TCP |
| SSL support | no | yes | yes | soon | yes |

## ACKNOWLEDGMENT

The research presented in this paper was partially supported by the National Centre for Research and Development (NCBiR) under Grant No. PBS1/B9/18/2013 and by the Polish Ministry of Science and Higher Education under AGH University of Science and Technology Grant 11.11.230.124 (statutory project).

## REFERENCES

[1] J. Hansen, T.-M. Grønli, and G. Ghinea, "Towards cloud to device push messaging on android: Technologies, possibilities and challenges," *International Journal of Communications, Network and System Sciences*, vol. 5, no. 12, pp. 839–849, 2012. doi: 10.4236/ijcns.2012.512089

[2] P. Nawrocki and M. Soboń, "Public cloud computing for software as a service platforms," *Computer Science*, vol. 15, no. 1, 2014. doi: 10.7494/csci.2014.15.1.89. [Online]. Available: http://journals.agh.edu.pl/csci/article/view/519

[3] K. Tang, Y. Wang, H. Liu, Y. Sheng, X. Wang, and Z. Wei, "Design and implementation of push notification system based on the MQTT protocol," in *2013 International Conference on Information Science and Computer Applications (ISCA 2013)*. Atlantis Press, 2013. doi: 10.2991/isca-13.2013.20

[4] D. Schuster, I. Koren, T. Springer, D. Hering, B. Söllner, M. Endler, and A. Schill, *Creating Applications for Real-Time Collaboration with XMPP and Android on Mobile Devices*. IGI Global: Handbook of Research on Mobile Software Engineering: Design, Implementation and Emergent Applications, 2012.

[5] J. Kosinski, P. Nawrocki, D. Radziszowski, K. Zielinski, S. Zielinski, G. Przybylski, and P. Wnek, "SLA monitoring and management framework for telecommunication services," in *Networking and Services, 2008. ICNS 2008. Fourth International Conference on*, March 2008. doi: 10.1109/ICNS.2008.31 pp. 170–175.

[6] T. Szydlo, P. Nawrocki, R. Brzoza-Woch, and K. Zielinski, "Power aware MOM for telemetry-oriented applications using GPRS-enabled embedded devices—levee monitoring use case," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 2. IEEE, 2014. doi: 10.15439/2014F252 pp. pages 1059–1064. [Online]. Available: http://dx.doi.org/10.15439/2014F252

[7] R. Brzoza-Woch, M. Konieczny, B. Kwolek, P. Nawrocki, T. Szydło, and K. Zieliński, "Holistic approach to urgent computing for flood decision support," *Procedia Computer Science*, vol. 51, no. 0, pp. 2387 – 2396, 2015. doi: 10.1016/j.procs.2015.05.414 International Conference On Computational Science, ICCS 2015 Computational Science at the Gates of Nature. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915012223

[8] R. Oldenburg, "Keeping google cloud messaging for android working reliably [technical post]," http://blog.pushbullet.com/2014/02/12/keeping-google-cloud-messaging-for-android-working-reliably-techincal-post, February 2014.

[9] C. Duckett, "Android malware utilising google cloud messaging service," http://www.zdnet.com/android-malware-utilising-google-cloud-messaging-service-7000019427/, August 2013.

[10] G. Ghinamo, F. Vadala, C. Corbi, P. Bettassa, F. Risso, and R. Sisto, "Vehicle navigation service based on real-time traffic information: A RESTful netAPI solution with long polling notification," in *Ubiquitous Positioning, Indoor Navigation, and Location Based Service (UPINLBS), 2012*, Oct 2012. doi: 10.1109/UPINLBS.2012.6409749 pp. 1–8.

[11] K. Wylie, "REST requires asynchronous notification," http://kirkwylie.blogspot.com/2008/12/rest-requires-asynchronous-notification.html, December 2008.

[12] D. Wood and D. Robson, "Message broker technology for flexible signalling control," in *Proc. ASPECT 2012 Conference*, 2012.

[13] M. Richards, "Understanding the difference between AMQP and JMS," *NFJS Magazine*, May 2011.

[14] M. Rostański, K. Grochla, and A. Seman, "Evaluation of highly available and fault-tolerant middleware clustered architectures using RabbitMQ," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 2. IEEE, 2014. doi: 10.15439/2014F48 pp. pages 879–884. [Online]. Available: http://dx.doi.org/10.15439/2014F48

[15] L. Zhang, B. Tiwana, Z. Qian, Z. Wang, R. P. Dick, Z. M. Mao, and L. Yang, "Accurate online power estimation and automatic battery behavior based power model generation for smartphones," in *Proceedings of the Eighth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, ser. CODES/ISSS '10. New York, NY, USA: ACM, 2010. doi: 10.1145/1878961.1878982. ISBN 978-1-60558-905-3 pp. 105–114. [Online]. Available: http://doi.acm.org/10.1145/1878961.1878982

DOI: 10.15439/2015F128
ACSIS, Vol. 5

# Open Data collection using mobile phones based on CKAN platform

Katarzyna Oświecińska (1,2)
(1) Orange Labs
CBR
ul. Obrzeżna 7
02-691 Warsaw, Poland
(2) Polsko-Japońska Akademia Technik Komputerowych
ul. Koszykowa 86
02-008Warsaw, Poland
Email: k.oswiecinska@gmail.com

Jarosław Legierski (1,3)
(1) Orange Labs
CBR
ul. Obrzeżna 7
02-691 Warsaw, Poland
Email: jaroslaw.legierski@orange.com

(3) Warsaw University of Technology, Faculty of Mathematics and Information Science,
ul. Koszykowa 75, 00-662 Warsaw, Poland

***Abstract* — This paper presents concept and prototype of universal easy-to-use and self-configurable mobile application dedicated for open data sets collection and sending them to the central database storage based on CKAN platform. This article describes the concept of open data collection in crowdsourcing model using mobile phones and also an architecture of end user application (CKAN client) dedicated for Android operating system.**

## I. Introduction

The concept of Open Data has been known in science and business environments for many years. The main definition of such kind of information describes that “certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control” [1], [2].

In the world there are many installed platforms, exposed open data sets produced by: cities, government institutions, scientific, and private companies. Large sets of this information being offered under open licenses such as mentioned in [3] are exposed in up to bottom model. Main part of such data comes from government institutions collecting and processing a large number of information about: their citizens, financial data or statistical information. Another idea regarding open data collection is concentrated on up to button approach. This model is based on idea of crowdsourcing and allows citizens to collect data.

Because of growing popularity of mobile devices (e.g. smart phones or smart watches) in recent years this kind of devices becomes the most convenient tool for collecting and aggregating open data sets. Therefore, dedicated, easy-to-use mobile application use mobile phone functions such as GPS receiver and photo camera. In many cases it becomes the best tool for the data acquisition using crowdsourcing.

## II. Existing Solutions

On the Internet we can find large set of mobile applications dedicated to open data collecting. For example Yanosik [4] – the most popular, based on crowdsourcing model, mobile application in Poland. This service allows drivers to report information about dangerous situations on the roads: accidents, construction zones, police patrols locations etc. Other application users, basing on the data collected and exposed in this system, can adapt their behavior to local conditions and thus improve the safety on the roads. Another exemplary system based on crowdsourcing concept is mobile application 19115 [5] developed by City of Warsaw. This service (mobile application is only one component of the 19115 system) is dedicated to report non-emergency incidents in Warsaw. Residents using this service can report any information about the city (road damage, problems with garbage, problems which needs the intervention of the municipal guard, etc.). It could be noticed that applications presented above are dedicated only to specific use cases (road traffic reporting or processing interventions in City of Warsaw) and it is not possible to use them to collect any kind of open data. These solutions are dedicated to specific usage, not configurable and closed from systems integration point of view.

In literature we can find some open solutions dedicated to open data collection. Another type of mobile applications are those, mostly concentrated on processing and transferring open data from end user in universal form in which the scope and type of collected open data is configurable by the system administrators. The very good example is Open Data Kit [6]- set of tools which manages mobile data collection. Open Data Kit allows to build a data collection in survey form, collect the data

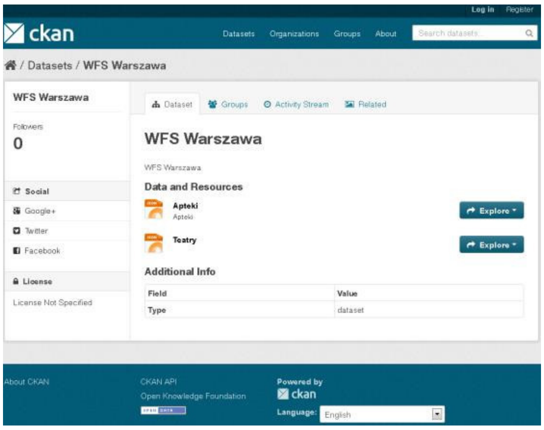
978-83-60810-66-8/$25.00©2015, IEEE

using mobile device and send data to the server. Tool supports also aggregation of data on the server side and visualize them.

In literature, there can be found also applications dedicated for CKAN platform [7] - open source framework de'facto standard in open data exposition. First one is Ukansearch - mobile web application to improve access and availability to open data. Unfortunately actually Ukansearch allows only to search using tags datasets on the www.data.ug portal [8] and doesn't support open data collection process. The second solution is sample CKAN Android client [9]. This tool was developed about 3 years ago and still exists only in beta version.

## III. CKAN

CKAN (Comprehensive Knowledge Archive Network) – is open source data portal and open data exposition platform. This project was started in March 2006 [7] and is maintained by The Open Knowledge Foundation [10]. After 9 years of development CKAN is world-leading open data platform and there are more than 60 instances [11] of the system installed worldwide such as: data.gov - a portal of the United States government [12] British data.gov.uk [13], or the publicdata.eu European Union open data portal [14].



/ Datasets / WFS Warszawa

WFS Warszawa

Followers
0

License

License Not Specified

WFS Warszawa

WFS Warszawa

Data and Resources

Apteki
Apteki

Teatry

Additional Info

| Field | Value |
|---|---|
| Type | dataset |

Fig. 1. Example data set in CKAN platform

CKAN platform allows for storage and management of open data repository, publication, search and visualize of data sets. System also allows developers to download and upload data sets via rich JSON based on API (Application Programming Interface). CKAN is open source software distributed under open license Affero GNU GPL v3.0.

## IV. IDEA OF CROWDENABLER

Because currently there is no universal mobile application dedicated to open data collection using CKAN portal the main goal was to build the concept, development and to test this type of application.

Crowd Enabler is an application dedicated to the collection of open data. With it, users of smartphones with the Android operating system can complement the individual datasets hosted on the CKAN platform. Native application function allows to add individual records to CKAN using geolocation information based on implemented in the phone GPS receiver. The main feature of the Crowd Enabler application allows to add objects containing text data, image files and the geographical coordinates by the end-user of the mobile phone. A user with the application, may, at any place (with enabled internet access and GPS location), take a photo of any object, add descriptions, and add its geolocation.

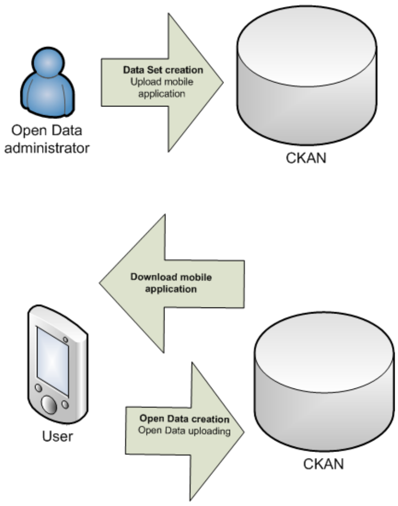


Fig. 2. Idea of Crowd Enabler

From the CKAN administrator and mobile application end user point of view the application usage looks as follows:

1) Data Set creation and upload Crowd Enabler.
In this step system administrator creates in CKAN an empty data set, prepares and uploads it as an CKAN resource. Then installs version of Crowd Enabler mobile application. The url to the Crowd Enabler is distributed (by e-mail or web page) to the end application users.
2) Download Crowd Enabler mobile application.
In this step application users responsible for open data collection download and install mobile application on their phones. It should be noticed that because of not publishing Crowd Enabler in Google Play application shop the option install from "Unknown sources" in mobile devices during this step must be turned on.
3) Data collection.
In this step application user can collect and store

data sets using mobile application. Application after start retrieves all information from CKAN platform using RDF metadata extension and reconfigures (changes names of the buttons and hints for editable fields). Application process open data in three following steps:

a) collected geographic location based on GPS receiver,

b) collected object description (in text form),

c) collected photo,

d) sending data to the repository (CKAN).

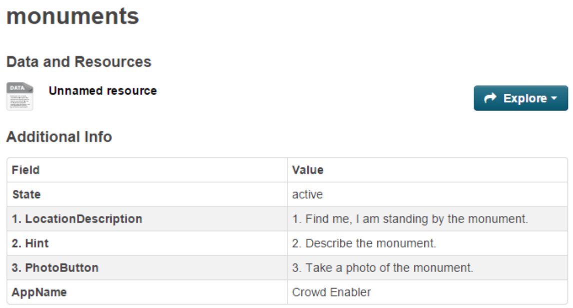
## V. RDF ONTOLOGY USAGE

Resource Description Framework (RDF) [15] is a standard for describing web resources. RDF using some of the extension such as Data Catalog Vocabulary (DCAT) can be used for description of any resources on the Internet and used for building semantic Web concept. Semantic internet is the vision of Tim Berners-Lee and in data layer it's architecture is based on Linked Data principles. These principles are as follows [16], [17]:

- Usage URIs as names for things.
- Usage HTTP URIs so that people can look up those names.
- When someone looks up a URI, provides useful information, using the standards (RDF, SPARQL).
- Exposition links to other URIs, so that they can discover more things.

In presented in this paper application nonstandard RDF usage was proposed and implemented – this additional elements describing web resources were used for exposition of configuration data for mobile application.

In Crowd Enabler RDF extensions were used to define some application parameters. (e.g. Text box description etc.). Application uses RDF extension in Data Catalog Vocabulary (DCAT) standard implemented in CKAN platform.



monuments

Data and Resources

Unnamed resource

Additional Info

| Field | Value |
|---|---|
| State | active |
| 1. LocationDescription | 1. Find me, I am standing by the monument. |
| 2. Hint | 2. Describe the monument. |
| 3. PhotoButton | 3. Take a photo of the monument. |
| AppName | Crowd Enabler |

Fig. 3. DCAT RDF extension in CKAN platform

In CKAN in format key-value there were 4 parameters defined and described in the table below:

Table 1. DCAT RDF parameters defined in CKAN dedicated for Crowd Enabler

|  | Parameter | description |
|---|---|---|
| 1 | LocationDescription | Task name / name of collected object |
| 2 | Hint | Collected object detailed description |
| 3 | PhotoButton | Task name (photo) |
| 4 | AppNAme | Name of application |

Based on RDF XML extension in CKAN parameters presented above are accessible for developers during access to the data set via API using additional header in HTTP request **Accept: application/rdf+xml.**

```
<rdf:RDF     xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#"         xmlns:dcat="http://www.w3.org/ns/dcat#"
xmlns:dct="http://purl.org/dc/terms/">
  <dcat:Dataset
rdf:about="http://demo.ckan.org/dataset/monuments">
    <owl:sameAs     rdf:resource="urn:uuid:3aecf676-
e1cd-4eda-9e35-6805d50bafae"></owl:sameAs>
    <dct:description></dct:description>
    <foaf:homepage
rdf:resource="http://demo.ckan.org/dataset/monuments
"></foaf:homepage>
    <rdfs:label>monuments</rdfs:label>
    <dct:identifier>monuments</dct:identifier>
    <dct:title>monuments</dct:title>
        <dcat:distribution>
            <dcat:Distribution>
                <dcat:accessURL
rdf:resource="http://demo.ckan.org/dataset/3aecf676-
e1cd-4eda-9e35-6805d50bafae/resource/53ed83a3-8d89-
4c81-b7a9-
eaa55553ea5b/download/table.txt"></dcat:accessURL>
            </dcat:Distribution>
        </dcat:distribution>
        <dcat:distribution>
            <dcat:Distribution>
                <dcat:accessURL
rdf:resource="http://demo.ckan.org/dataset/3aecf676-
e1cd-4eda-9e35-6805d50bafae/resource/5f3441dd-8473-
4c44-af20-
de71af3c625c/download/photo.jpeg"></dcat:accessURL>
               <dct:title>photo</dct:title>
            </dcat:Distribution>
        </dcat:distribution>
        <dct:relation>
          <rdf:Description>
            <rdfs:label>1.
LocationDescription</rdfs:label>
            <rdf:value>1. Find me, I am standing by
the monument.</rdf:value>
          </rdf:Description>
        </dct:relation>
        <dct:relation>
          <rdf:Description>
            <rdfs:label>2. Hint</rdfs:label>
            <rdf:value>2.       Describe       the
monument.</rdf:value>
          </rdf:Description>
        </dct:relation>
        <dct:relation>
          <rdf:Description>
            <rdfs:label>3. PhotoButton</rdfs:label>
            <rdf:value>3.  Take  a  photo  of  the
monument.</rdf:value>
          </rdf:Description>
        </dct:relation>
        <dct:relation>
          <rdf:Description>
            <rdfs:label>AppName</rdfs:label>
            <rdf:value>Crowd Enabler</rdf:value>
          </rdf:Description>
        </dct:relation>
  </dcat:Dataset>
</rdf:RDF>
```

Fig. 4. DCAT RDF extension in CKAN XML dataset description

## VI. SYSTEM ARCHITECTURE

This section of this publication contains the architecture and functionality of Crowd Enabler application. This mobile application is dedicated for Android operating system and was developed using Eclipse and Android Development Tools (ADT)

### *A) Programing environment*

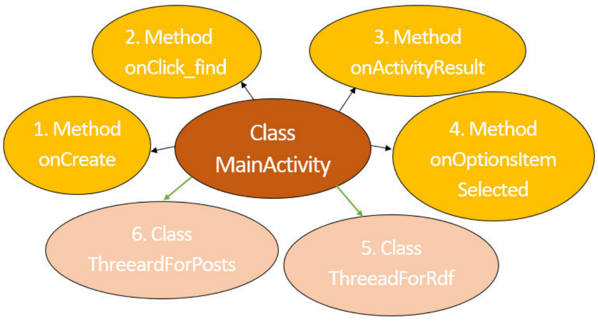

2. Method onClick_find
3. Method onActivityResult
1. Method onCreate
Class MainActivity
4. Method onOptionsItem Selected
6. Class ThreeardForPosts
5. Class ThreeadForRdf

Fig. 5. Application class structure

The main methods and classes implemented in the application are as follows:

1. Method onCreate handles methods which are connected to buttons and starting ThreeadForRdf and ThreeadForPosts.
2. Method onClick_find sets GPS coordinates.
3. Method onActivityResult makes sure that our photo exists and is displayes in the app.
4. Method onOptionsItem Selected handles pressing Settings.
5. The main Method getRdfXml in Class ThreeadForRdf, gets useful informations from rdf about dataset. Using this, sets descriptions of buttons and editText.
6. Method makeRequestImage in Class ThreardForPosts sends image to dataset. Method makeRequest sends full data (in JSON) of record to dataset.

-

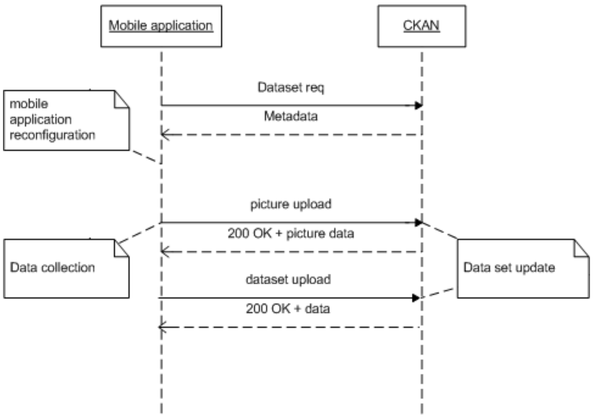


Fig. 6. Application data flow

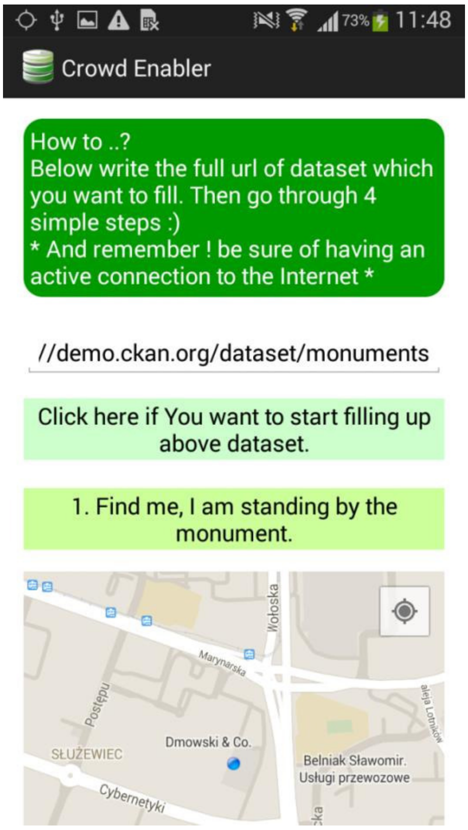
Data flow between Crowd Enabler and CKAN platform is presented on Fig.6

1) Dataset request – application after launch based on defined http resource connects with CKAN Platform, downloads metadata and RDF extensions in key – value form used for Crowd Enabler configuration.
2) Crowd Enabler self-configures and shows GUI.
3) User takes photo and uploads it on the CKAN server as open data resource.
4) User performs an additional description of data record (text ,GPS coordinates), uploads and sets it to the CKAN.

### *B) Application GUI*

End application user is a person who uses Crowd Enabler on it's mobile phone – and collects open data records. After start, application connects to CKAN and catches from RDF scheme of metadata to change names of buttons and hints defined in the application GUI.

When user wants to collect data using geolocation, must stand with his smartphone nearby an object which he wants to upload (e.g. memorial or monument).



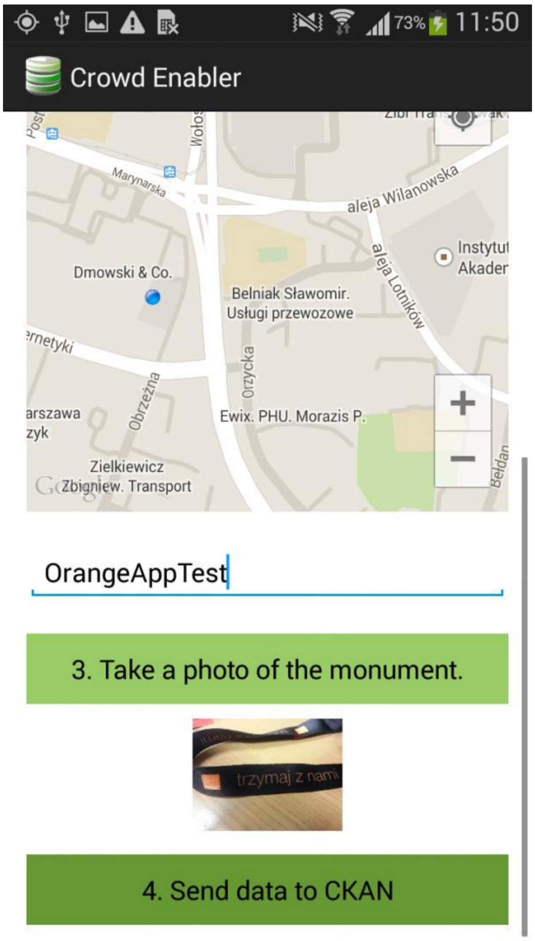
How to ..?
Below write the full url of dataset which you want to fill. Then go through 4 simple steps :)
* And remember ! be sure of having an active connection to the Internet *

//demo.ckan.org/dataset/monuments

Click here if You want to start filling up above dataset.

1. Find me, I am standing by the monument.

Fig. 7. Application GUI



OrangeAppTest

Fig. 8. Application GUI

After taking photo and adding additional description end application user sends data to a server by pressing Send Button.

## VII. MEASURENMENT

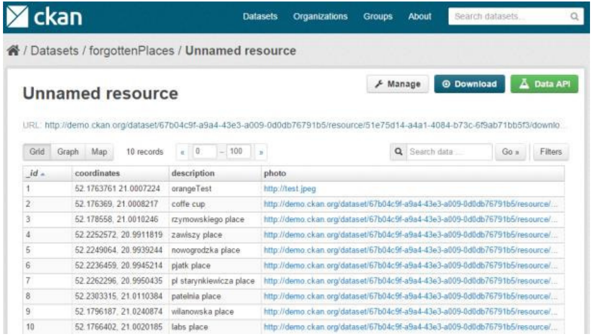
Using Crowd Enabler application several data sets were collected

1) monuments - historic buildings and monuments in Warsaw

2) forgottenPlaces - some interesting places in the city

3) Trees - the test group of trees

The Figures bellow shows the example data sets with forgottenPlaces and monuments in Warsaw



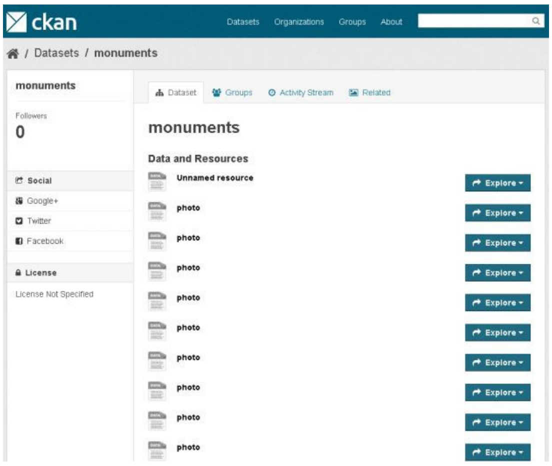
/ Datasets / forgottenPlaces / Unnamed resource

Unnamed resource

URL: http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/51e75d14-a4a1-4084-b73c-6f9ab71bb5f3/downlo...

| _id | coordinates | description | photo |
|---|---|---|---|
| 1 | 52.1763761 21.0007224 | orangeTest | http://test.jpeg |
| 2 | 52.176369, 21.0008217 | coffe cup | http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/... |
| 3 | 52.178558, 21.0010246 | rzymowskiego place | http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/... |
| 4 | 52.2252572, 20.9911819 | zawiszy place | http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/... |
| 5 | 52.2249064, 20.9939244 | nowogrodzka place | http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/... |
| 6 | 52.2236459, 20.9945214 | pjatk place | http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/... |
| 7 | 52.2262296, 20.9950435 | pl starynkiewicza place | http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/... |
| 8 | 52.2303315, 21.0110384 | patelnia place | http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/... |
| 9 | 52.1796187, 21.0240874 | wilanowska place | http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/... |
| 10 | 52.1766402, 21.0020185 | labs place | http://demo.ckan.org/dataset/67b04c9f-a9a4-43e3-a009-0d0db76791b5/resource/... |

Fig. 9. Data set with forgottenPlaces in Warsaw



/ Datasets / monuments

monuments

Followers
0

License Not Specified

monuments

Data and Resources

Unnamed resource
photo
photo
photo
photo
photo
photo
photo
photo
photo

Fig. 10. Example dataset with monuments in Warsaw

It needs to be stressed that for application test only small data sets in the amount of several records each were collected.

## VIII. SUMMARY AND FUTURE WORK

In the future, authors intend to make some modifications in the Crowd Enabler application

1) Improvement of application installation process - application can be exposed in Google Play application shop and installed after connection and download without "install from unknown source" Android option usage

2) Usage of another geolocation sources instead of GPS (e.g. location based on Telco operator infrastructure, Wifi infrastructure etc.).

The main goal of this project was to build a mobile client for CKAN open data platform.

The universality of developed application based on self-configuration capabilities as well as proposed and tested nonstandard RDF meta data usage needs to be emphasized. It should be mentioned that Crown Enabler application has been created in Open Middleware Model using Open Application Programming Interfaces allowing in ease way to create innovative applications and services such as: [17], [19], [20], [21].

The potential of usage of the presented application is very wide. End application users can store large sets of data, that can be described by location, photos and extended text description.

Prototype of Crowd Enabler application was made as part of the Open Middleware 2.0 Community by Orange Labs program [22]

## References

[1] Auer, S. R.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. (2007). "DBpedia: A Nucleus for a Web of Open Data". "The Semantic Web". Lecture Notes in Computer Science 4825. p. 722.

[2] Portal http://en.wikipedia.org/wiki/Open_data [27.04.2015]

[3] Portal http://opendatacommons.org/licenses/ [27.04.2015]

[4] Portal http://yanosik.pl/ [29.04.2015]

[5] Portal https://warszawa19115.pl/ [29.04.2015]

[6] Open Data Kit (ODK) portal https://opendatakit.org/about/deployments/ [23.10.2014]

[7] CKAN Portal http://ckan.org/ [24.10.2014]

[8] https://github.com/davidebukali/Ukansearch

[9] https://github.com/47deg/labs-opendata-adopta-playa-android

[10] The Open Knowledge Foundation Portal https://okfn.org/ [24.10.2014]

[11] Portal http://datacatalogs.org/catalog?q=platform [13.06.2014]

[12] Portal https://www.data.gov/ [23.10.2014]

[13] Portal data.gov.uk [23.10.2014]

[14] Portal http://publicdata.eu [23.10.2014]

[15] Klyne, G., Carroll, J. J.: Resource description framework (RDF): concepts and abstract syntax. Technical report W3C, 2 (2004)

[16] http://www.w3.org/DesignIssues/LinkedData.html

[17] Sören Auer Volha Bryl Sebastian Tramp (Eds.) Linked Open Data – Creating Knowledge Out of Interlinked Data Results of the LOD2 Project Lecture Notes in Computer Science Springer 2014

[18] Trusiewicz, P.; Witan, M.; Kuzia, M., "Mobile Payment System - Telco 2.0 application dedicated for payments," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.859,864, 8-11 Sept. 2013

[19] Wawrzyniak, P.; Korbel, P.; Borowska-Terka, A., "Student information delivery platform using telecommunications open middleware APIs," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.871,874, 8-11 Sept. 2013

[20] Korbel, P.; Skulimowski, P.; Wasilewski, P.; Wawrzyniak, P., "Mobile applications aiding the visually impaired in travelling with public transport," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp. 825,828, 8-11 Sept. 2013

[21] P. Wawrzyniak, Ł. Wronkowski, D. Kuniszewski, A. Cackowski, P. Czapliński i K. Szymański "Send It Safe – A Novel Application for Secure Key Exchange Using Telecommunications Open Middleware APIs," w Frontiers in Network Applications, Network Systems and Web Services (SoFAST-WS'14) w Federated Conference on Computer Science and Information Systems FedCSIS 2014, Warszawa 2014

[22] Open Middleware 2.0 Community portal – http://www.openmiddleware.pl [20.05.2013]

# 4th International Conference on Wireless Sensor Networks

A FEW years ago, the applications of WSN were rather an interesting example than a powerful technology. Nowadays, this technology attracts still more and more scientific audience. Theoretical works from the past, where WSN principles were investigated, grew into attention-grabbing applications practically integrated by this time in a real life. It could be said, that countless application fields, from military to healthcare, are already covered by WSN. Together with this technology expansion, still new and new tasks and interesting problems are arising. Simultaneously, such application actions stimulate the progress of WSN theory that at the same time unlocks new application possibilities. The typical examples are developments within the "Internet-of-Things" field as well as advancements in eHealth domain with WBAN IEEE 802.15.6 standard progress.

Wireless sensor networks, as the spatially distributed networks consisted of a number of relatively simple, low-cost, low-power components interconnected mutually, provide quite wide application portfolio for different branches of economy. As the main examples could be mentioned military, industry, transport, agriculture and healthcare. However, in the near future, even stronger expansion of WSN application assortment is expected. In order to make this expansion possible, it is necessary to continually work on the solving of typical questions/problems related to the WSN development, e.g. standardization of communication protocols; the lack of energy-efficient power sources; the development of new ultra-low-power microelectronic components; etc.

An integration of WSN within the public data networks as well as within the domains where confidential and private data are processed (e.g. E-Health) brings along problems related to the ethical and legal questions too. Therefore, the terms as social safety or ethical safety should not be neglected.

The problematic of WSN is one of actual activities getting to the fore in the European Research Area since the issue of sensor networks was covered through "IoT" in FP7 program and strong continual extension is planned to be included also in Horizon 2020 program, especially in sections such Smart Transport; Health; Climate Action covered under Societal Challenges Pillar.

It is therefore essential to create an experience-sharing platform for scientific researchers and experts from research institutes, SMEs and companies who work in WSN domain where they can exchange some relevant skills and experiences as well as discuss upcoming trends and new ideas from this field. Moreover, the conference should also serve a function of a kind of networking platform facilitating interconnectivity between participants in case of a future collaboration.

## Topics

Original contributions, not currently under review to another journal or conference, are solicited in relevant areas including, but not limited to, the following:

*Development of sensor nodes and networks*

- Sensor Circuits and Sensor devices – HW
- Applications and Programming of Sensor Network – SW
- Architectures, Protocols and Algorithms of Sensor Network
- Modeling and Simulation of WSN behavior
- Operating systems

*Problems dealt in the process of WSN development*

- Distributed data processing
- Communication/Standardization of communication protocols
- Time synchronization of sensor network components
- Distribution and auto-localization of sensor network components
- WSN life-time/energy requirements/energy harvesting
- Reliability, Services, QoS and Fault Tolerance in Sensor Networks
- Security and Monitoring of Sensor Networks
- Legal and ethical aspects related to the integration of sensor networks

*Applications of WSN*

- Military
- Health-care
- Environment monitoring
- Transportation & Infrastructure
- Precision agriculture
- Industry application
- Security systems and Surveillance
- Home automation
- Entertainment – integration of WSN into the social networks
- Other interesting applications

## Event Chairs

**Hodoň, Michal,** University of Žilina, Slovakia
**Kapitulík, Ján,** University of Žilina, Slovakia
**Micek, Juraj,** University of Žilina, Slovakia
**Ševcik, Peter,** University of Žilina, Slovakia

## Program Committee

**Al-Anbuky, Adnan,** Auckland University of Technology, New Zealand

**Baranov, Alexander,** Russian State University of Aviation Technology, Russia

**Brida, Peter,** University of Zilina, Slovakia

**Dadarlat, Vasile-Teodor,** Univiversita Tehnica Cluj-Napoca, Romania

**Diviš, Zdenek,** VŠB-TU Ostrava, Czech Republic

**Elmahdy, Hesham N.,** Cairo University, Egypt

**Fortino, Giancarlo,** Università della Calabria

**Fouchal, Hacene,** University of Reims Champagne-Ardenne, France

**Furtak, Janusz,** Military University of Technology, Faculty of Cybernetics, Poland, Poland

**Giusti, Alessandro,** CyRIC - Cyprus Research and Innovation Center, Cyprus

**Grzenda, Maciej,** Orange Labs Poland and Warsaw University of Technology, Poland

**Gu, Yu,** National Institute of Informatics, Japan

**Hudik, Martin,** University of Zilina

**Husár, Peter,** Technische Universität Ilmenau, Germany

**Jin, Jiong,** Swinburne University of Technology, Australia

**Jurecka, Matus,** University of Žilina, Slovakia

**Kafetzoglou, Stella,** National Technical University of Athens, Greece

**Karastoyanov, Dimitar,** Bulgarian Academy of Sciences, Bulgaria

**Karpiš, Ondrej,** University of Žilina, Slovakia

**Kochláň, Michal,** University of Žilina, Slovakia

**Laqua, Daniel,** Technische Universität Ilmenau, Germany

**Milanová, Jana,** University of Žilina, Slovakia

**Monov, Vladimir V.,** Bulgarian Academy of Sciences, Bulgaria

**Ohashi, Masayoshi,** Advanced Telecomunications Research Institute International / Fukuoka University, Japan

**Papaj, Jan,** Technical university of Košice, Slovakia

**Ramadan, Rabie,** Cairo University, Egypt

**Scholz, Bernhard,** The University of Sydney, Australia

**Shaaban, Eman,** Ain-Shams university, Egypt

**Shu, Lei,** Guangdong University of Petrochemical Technology, China

**Smirnov, Alexander,** Linux-WSN, Linux Based Wireless Sensor Networks, Russia

**Staub, Thomas,** Data Fusion Research Center (DFRC) AG, Switzerland

**Stojmenovic, Ivan,** University of Ottawa, Canada

**Teslyuk, Vasyl,** Lviv Polytechnic National University, Ukraine

**Wang, Zhonglei,** Karlsruhe Intitute of Technology, Germany

**Xiao, Yang,** The University of Alabama, United States

DOI: 10.15439/2015F146
ACSIS, Vol. 5

# Real-Time Schedule for Mobile Robotics and WSN Aplications

Michal Chovanec *
University of Žilina
Faculty of Management Science and Informatics,
Univerzitná 8215/1 Žilina 010 26,
michal.chovanec@fri.uniza.sk

Peter Šarafín
University of Žilina
Faculty of Management Science and Informatics,
Univerzitná 8215/1 Žilina 010 26,
peter.sarafin@fri.uniza.sk

***Abstract*—This paper presents real-time scheduler in operating system running on ARM Cortex (M0, M3, M4) usable in small mobile robotics with kernel response around 1ms. Thanks to the strong modularity, advanced sleep modes and event driven programming ability, it can be used for WSN applications too.**

***Index Terms*—operating system, ARM Cortex M, mobile robotics, WSN node, low-power, real-time**

## I. Introduction

REAL-TIME scheduler provides added value for embedded software development in the form of strong modularity, reusable code and rapid development [1]. Many embedded applications work without operating system - usually single purpose tasks or interrupt driven tasks. For more complex applications, operating system can provide better results when some common problems occure [2] [3] :

- Multiple sensors (or any inputs) reading
- Multiple control loops with different sampling time
- Communication (routing, resending)
- Power management
- System modularity and extension posibilities
- GUI running on background of the main process

From these, we can consider following operating system requirements:

- Multiple parallel threads (often with priority scheduling)
- Real-time processing ability
- Code size acceptable for microcontroller abilities
- Sleep modes support
- Multiplatform compilation ability
- Modular architecture

## II. Sytem architecture

The operating system runs on a single chip microcontroller. Supported cores are ARM Cortex M0, M0+, M3, M4 and M4F. For testing, TI TivaC TM4C123G [5] (Fig. 1) with cortex M4F core has been used. This MCU is running on 80MHz and it disposes 256K flash memory and 32K SRAM. Other devices, such as STM32L053, STM32F103, STM32F407, LPC812, MKL02Z32 and MKL25Z4 have also been tested.

All parts are compiled using GNU GCC (using C99 standard) into single binary file, which can be loaded into flash memory [4]. Recent source files can be downloaded from [6].

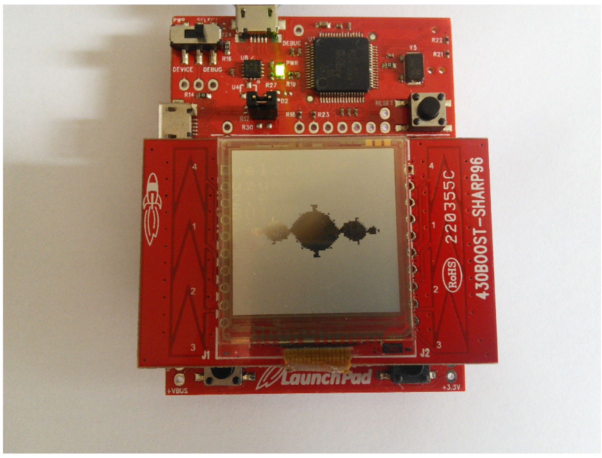


Fig. 1. TI TivaC launchpad testing board

Presented operating system consists of these parts:

- User application
- User libraries
- Kernel
- OS libraries
- Device low level libraries

All parts can work independently on each other. Only necessary part is device low level libraries, represented as HAL (hardware abstraction layer). Operating system is written with microkernel architecture, where kernel only creates and schedules threads. Other functions are implemented as optional libraries.

In the following text, we briefly describe OS structure. The priority scheduling algorithm compared with common round robin is described in more details.

### A. Booting process

After microcontroller reset, HAL is initialized first, especially clock configuration, GPIO initialization, UART timers and ADC setup. All parts are initialized only if they are linked in the binary. In other case, initialization of missing parts is skipped. Absolute minimum requirement for OS running is main clock initialization. For common problems, UART and timers are necessary.

978-83-60810-66-8/$25.00©2015, IEEE

Operating system libraries such as STDIO, software timers, messages subsystem and mutexes are initialized next. After this, kernel is initialized and user main thread is started.

### B. User application and libraries

Users main loop is discussed in this section. The main function is called $void\ main\_thread()$. This main thread can create other threads, by calling $create\_thread$ function. Following code shows how another thread can be created. When $void\ main\_thread()$ is running, user can initialize it's own libraries, usually sensors, displays or communication module. Boot up and four running threads screenshot is displayed in the Fig. 2.

Listing 1. Thread creating

```
thread_stack_t thread_01_stack
[THREAD_STACK_SIZE];

void thread_01(){
}

void main_thread(){
    create_thread(
        thread_01,
        thread_01_stack,
        sizeof(thread_01_stack),
        PRIORITY_MAX);

    while (1)
    {
    }
}
```

## III. Scheduling algorithm

Main part of OS is the microkernel core. Preemptive multitasking with two options, round robin scheduling or time decrease priority scheduling is implemented. To compare different schedule algorithms (especially real-time processing), we first need to define error function. Consider set of threads as

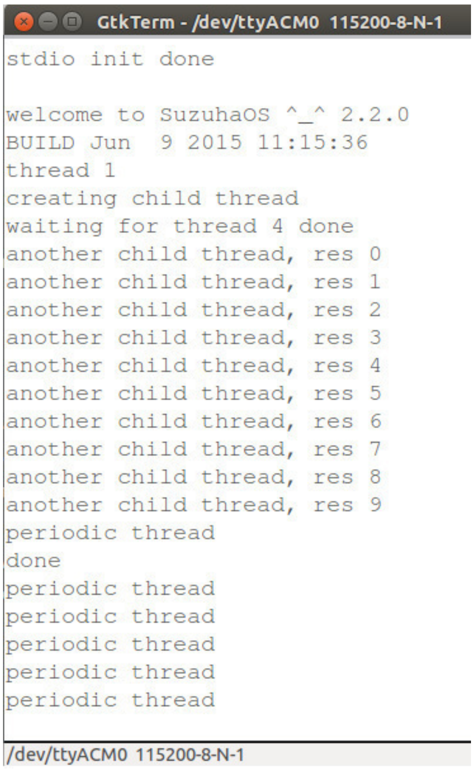
$$t_i \in T(p, k, s, d, c), \tag{1}$$

where $p$ represents thread priority (lower number - higher priority), $k$ is the counter of thread priority current value, $s$ stands for thread state (running, waiting, created), $d$ states thread deadline time (set by user, usually in ms), $c$ is thread running code (represented as Turing machine).

Let us define thread execution time function as $g(t_i)$ and error function as

$$e = \sum_{i=1}^{Tc} |d_i - g(t_i)|, \tag{2}$$

where $Tc$ is threads count. This function represents the error, which corresponds with the difference between required deadline time and measured time of running thread. Using priorities we can define error as



```
stdio init done

welcome to SuzuhaOS ^_^ 2.2.0
BUILD Jun  9 2015 11:15:36
thread 1
creating child thread
waiting for thread 4 done
another child thread, res 0
another child thread, res 1
another child thread, res 2
another child thread, res 3
another child thread, res 4
another child thread, res 5
another child thread, res 6
another child thread, res 7
another child thread, res 8
another child thread, res 9
periodic thread
done
periodic thread
periodic thread
periodic thread
periodic thread
periodic thread
```

GtkTerm - /dev/ttyACM0 115200-8-N-1
/dev/ttyACM0 115200-8-N-1

Fig. 2. OS terminal screenshot

$$e = \sum_{i=1}^{Tc} |d_i - g(t_i)| \frac{1}{p_i}, \tag{3}$$

where lower $p_i$ means higher priority.

Consider that the faster execution of the thread is not an issue. This fact means that CPU spends remaining time waiting (executing other threads or sleeping). We can write this as (4) and (5).

$$e_i = \begin{cases} d_i - g(t_i) & if\ d_i < g(t_i) \\ 0 & else \end{cases} \tag{4}$$

$$e = \sum_{i=1}^{Tc} |e(t_i)| \frac{1}{p_i} \tag{5}$$

Threads with higher priority (smaller $p_i$) have bigger influence on the total error. To implement priorities, we define following structure for each thread:

Listing 2. thread structure

```
struct sThread{
    u16 cnt, icnt;
    u32 flag;
    u32 *sp;
};
```

where $cnt$ and $icnt$ are counters used for priority scheduling, corresponding with $p$ and $k$ respectively in (1). When thread is created, $p$ and $k$ are set to $priority$ value and remain constant (variability during execution is also possible, but not tested yet). Each nonzero $k$ is decremented after each timer interrupt. Thread with smaller $k$ is chosen for the next execution and its $k$ is loaded back to $p$. Realization in C code is presented on following code listings.

Listing 3. priority scheduler

```
u32 i, min_i = 0;

/*find thread with minimum cnt*/
for (i = 0; i < THREADS_MAX_COUNT; i++)
{
    if (__thread__[i].cnt <
      __thread__[min_i].cnt)
        min_i=i;

    /*decrement counters*/
    if (__thread__[i].cnt != 0)
        __thread__[i].cnt--;
}

__thread__[min_i].cnt =
        __thread__[min_i].icnt;
__current_thread__ = min_i;
```

For full function, other common functions like thread creating, waiting or setting into waiting state are implemented.

Listing 4. kernel functions

```

void sched_off();
void sched_on();


void yield();

u32 get_thread_id();

void kernel_init();
void kernel_start();

u32 create_thread(
    void (*thread_ptr)(),
    thread_stack_t *s_ptr,
    u32 stack_size,
    u16 priority);

void kernel_panic();

void set_wait_state();
void wake_up_threads();
void wake_up_threads_int();

void join(u32 thread_id);
```

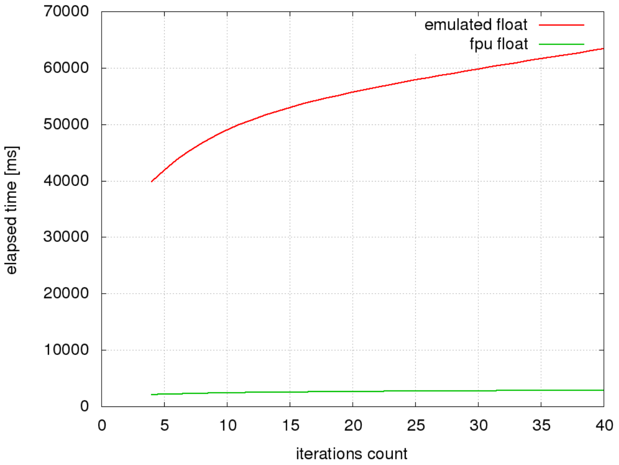


Fig. 3. FPU and CPU calculation times

## IV. Experimental results

For testing OS, few experiments were performed. First one was aimed for the basic multitasking test and comparison of performance using hardware and software emulated float performance. There were four running threads in this experiment. One thread was calculating Julia set fractal and results were displayed on LCD. Total number of calculating points was set to 96x96. Quantity of algorithm iterations was changing from interval $\langle 4, 40\rangle$. Performance result is represented in the Fig. 3. In this test, basic functionality has been tested, especially preemptive multitasking, timers and terminal interface.

Next testing was focused on the real-time processing ability. There were two main and six child threads (only first three are shown in figures). Each thread was waiting specified time (required waiting time) and this time was measured. Difference between required and measured time was used to compare round robin and priority scheduler scheduling algorithms. We

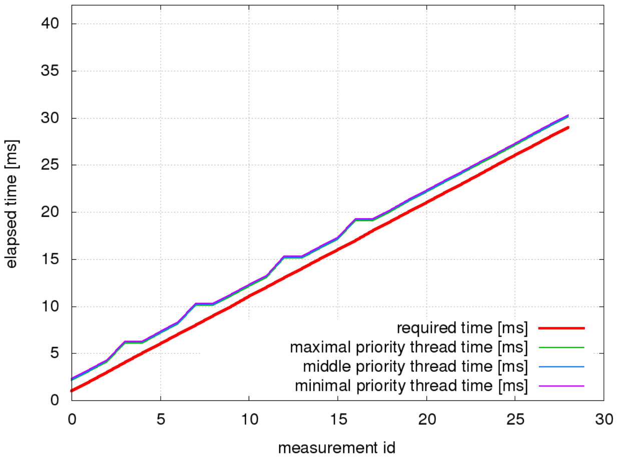


Fig. 4. Round robin real-time test

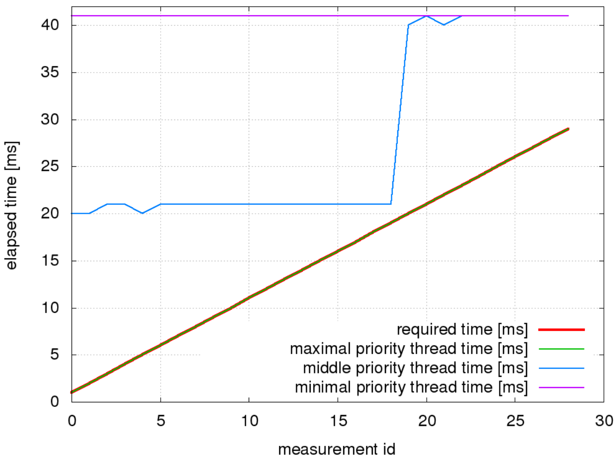


Fig. 5. Priority scheduler real-time test

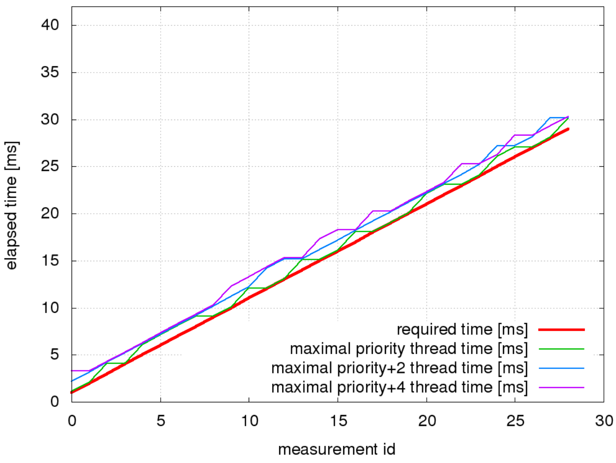


Fig. 7. Priority scheduler real-time test with similar priorities

can see round robin results in the Fig. 4 (all threads have same results). It can be seen, that required value is bellow measured lines and the time difference is around 1ms. Little peaks are consequence of timer resolution, which is 1ms. Situation with priority scheduler can be seen in the Fig. 5.

Threads with the maximum priority perfectly meet the conditions. Threads with lower priorities were executing for much longer time. Following priority values $p_i$ have been used:

- PRIORITY_MAX = 8
- PRIORITY_MID = 128
- PRIORITY_MIN = 255

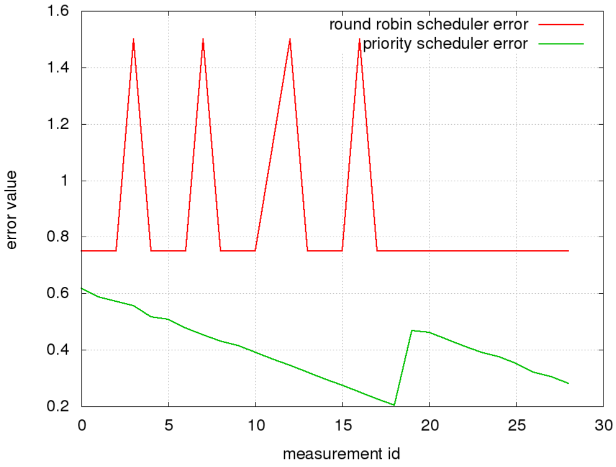


Fig. 6. Schedulers error comparison

We can use (5) to compute total error from meassured times. Result is shown in the Fig. 6. From priority scheduling algorithm, it can be seen that it converges into round robin when priorities are equal. This experiment was accomplished and results are presented in the Fig. 7.

## V. Conclusion

In this paper, priority scheduler has been explained and OS was briefly introduced. From experimental results, we can see that priority scheduler has better results, when considering error function definition (5). Of course, if we consider maximal deadline time without looking for priorities, round robin has better results. For applications, where same priority is necessary, round robin (or priority scheduler with same priorities represented in the Fig. 7) provides better solution. For applications, where it is needed to prioritize some processes, priority scheduler is of course better choice.

## References

[1] Whill Hentzen, The Software Developer's Guide, 3rd Edition, ISBN: 1-930919-00-X, 2002

[2] John A. Stankovic, Anthony D. Wood, Tian He, Realistic Applications for Wireless Sensor Networks, http://www.ent.mrt.ac.lk/dialog/documents/ERU-2-wsn.ppt

[3] Nuwan Gajaweera, Wireless Sensor Networks, http://www.ent.mrt.ac.lk/dialog/documents/ERU-2-wsn.ppt

[4] LM4 flahsing tool, https://github.com/utzig/lm4tools/tree/master/lm4flash

[5] Texas instruments TivaC Launchpad, http://www.ti.com/tool/ek-tm4c123gxl

[6] Suzuha OS sources https://github.com/michalnand/suzuha_os

[7] Ishwari Singh Rajput and Deepa Gupta : A Priority based Round Robin CPU Scheduling Algorithm for Real Time Systems, ISSN: 2319 1058, 2012

DOI: 10.15439/2015F169
ACSIS, Vol. 5

# Trust Security Mechanism for Marine Wireless Sensor Networks

Walid Elgenaidi, Thomas Newe
University of Limerick,
Optical Fibre Sensors Research Centre
Department of Electronic and Computer Engineering
Limerick, Ireland
Email: {walid.elgenaidi, thomas.newe }@ ul.ie }

*Abstract*—**To provide a strong security service in Wireless Sensor Networks (WSNs), cryptographic mechanisms are required. Generally these security mechanisms demand intensive use of limited resource, such as memory, and energy to provide a defense against attacks. Monitoring the behavior of nodes and detecting risks according to these behaviors, and then taking decisions based on these measurements generally requires the use of a trusted Key Management scheme. In this paper we compare two existing security key management schemes that were designed for use in mobile ad hoc networks: "An overlay approach to data security in ad-hoc networks" authored by Jorg Liebeherr, Guangyu Dong, and "A hierarchical key management scheme for secure group communications in mobile ad hoc networks" authored by Nen-Chung Wang, Shian-Zhang Fang. Then a Hybrid Security Key Management Mechanism designed for use in the marine environment is proposed. This scheme focuses on reducing the memory storage of keys, using a leader node that is responsible for both the node joining and the node revoke processes. This security mechanism is implementing in real time on the Waspmote sensor platform.**

## I. Introduction

Design of smart security solutions for wireless sensor networks for specific fields such as marine environments is a big challenge. Since smart security protocols must be designed to have efficient and flexible key distribution systems to prevent attacks while conserving energy [1], [5]. This work aims to provide a secure technique using both Symmetric and Asymmetric key algorithms. Our designed scheme seeks protection against attacks by providing the standard security services such as confidentiality and authentication in addition to addressing the re-keying process between adjacent nodes, as well as reducing the number of stored keys. Therefore the sensor nodes are configured in a point-to-point topology which is suitable for marine coastal WSN systems.

The system proposed in [2] uses public/private keys, pre-message keys and requires that offline signed certificates are stored in each node. Authentication between nodes is performing without coordination with other nodes; this makes trust revocation difficult for 'bad' nodes. Each node maintains a single symmetric key that it shares with its current neighbours in the network topology. Furthermore, a public key is required in the authentication of new neighbours, and every node requires its neighbour's public key for use with the RSA public key algorithm.

The system proposed in [3] offers key management for secure group communications using a two-layer structure. The selected cluster head constructs and transmits a group key to all nodes. This scheme uses symmetric keys for subgroup keys and communication keys. The Diffie–Hellman "DH" key exchange scheme is utilized to achieve secure key transmission between subgroups. These schemes are discussed in section II below.

Section III presents a comparison between the neighbourhood key and hierarchical key management schemes for secure group communications. Section IV proposes a new hybrid key management scheme and section V concludes.

## II. Trust Key Management Techniques for Wireless Sensor Networks

There are essential practices for developing a good trust management system for WSNs and for the management of the necessary cryptographic keys [3]. In [2] Jorg Liebeherr, Guangyu Dong presented a key management and encryption scheme, called the neighbourhood key method, that ensures integrity and confidentiality of application data in overlay networks. The neighbourhood key method avoids network wide re-keying operations and payload data re-encrypting at each hop.

### A. Updating and Exchanging Neighbourhood Keys

The solutions presented in this scheme [2] are orthogonal to the problem of secure routing, which seeks protection against attacks to routing protocols. Each node has its own certificate and this certificate has been signed by a trusted third party using X.509 Version 3. These certificates, which include secret keys are exchanged between neighbours to use in encrypting or signing messages of

The authors would like to acknowledge the support of the SFI funded MaREI (Marine Renewable Energy Ireland) Centre and Ministry of Higher Education and Scientific Research - Libya

978-83-60810-66-8/$25.00©2015, IEEE

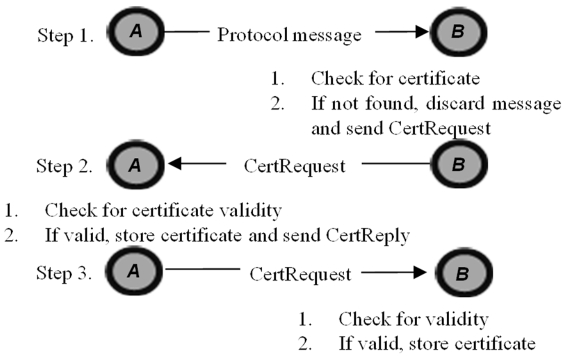

Step 1.
A
Protocol message
B
1. Check for certificate
2. If not found, discard message and send CertRequest
Step 2.
A
CertRequest
B
1. Check for certificate validity
2. If valid, store certificate and send CertReply
Step 3.
A
CertRequest
B
1. Check for validity
2. If valid, store certificate

Fig.1. Authentication of nodes

authenticated nodes. Encryption of data and the signing of hashes in each node are done with a single symmetric key called a"neighbourhood key". The neighbourhood key is shared with its current authenticated neighbours in network.

A joining node must generate a new neighbourhood key and send it to all of its authenticated neighbours in order to maintain confidentiality in the network.

Therefore every node must use a public key algorithm to encrypt the new neighbourhood key with the public keys of all the authenticated neighbours that are stored in the node during the authentication process as shown in "Fig. 1,". In this phase, nodes update keys only with current neighbours. However updating and exchanging a new neighbourhood key is executed whenever the set of authenticated neighbours are changed or the specified maximum lifetime of the current neighbourhood key is expired. The security issues are exacerbated during failures in reconstruction of the network topology when one or more nodes join and leave the network at the same time.

The neighbourhood Scheme prevents nodes against a DoS attack from a malicious adversary by implementing an integrity test, and also the allowed frequency of transmitted Key Request messages is limited.

### *B. Constructing and Transmitting Keys Using Cluster Head*

Nen-Chung Wang, and Shian-Zhang Fang [3] introduced a hierarchical key management scheme for secure group communications in a mobile ad hoc network. They proposed a new approach with a two-layer structure whereby a cluster head manages information between nodes in the layers as shown in "Fig. 2". Node with the largest weight value in each level is selected to be a cluster head [4]. The key transmission operation between nodes in the same level subgroup and with nodes in other level subgroups is coordinated by the cluster head.

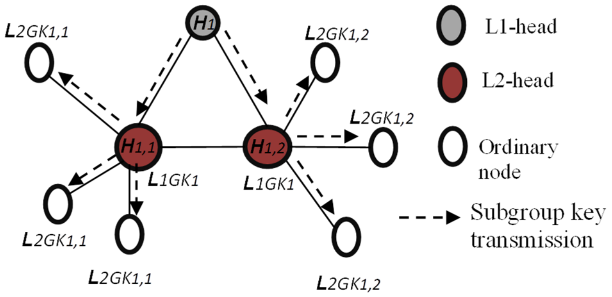


Fig.2. Subgroup key transmission operation between nodes

Level 1 subgroup "L1-subgroup" contains all of the nodes in the subgroup and the level 2 subgroups "L2-subgroup" are selected based on their positions. The node with the largest weight value in every L2-subgroup will be selected as the level 2 cluster head "L2-head" to manage the other nodes of the L2-subgroup.

The Diffie–Hellman "DH" scheme is used for secure transmission between nodes in subgroups, and each subgroup has its own subgroup key. L1-head generates the communication keys that are used between the different subgroups. The encryption and decryption operation during data transmission in different subgroups is only through subgroup keys, which means that packets are transmitted through the cluster heads.

The level 2 cluster head "L2-head" is responsible for a new node joining its subgroup. It initiates the generation of a new subgroup key "$K_{LXGK_S}$" after a new node joins a subgroup.

A node leaving of subgroup [4] falls under three cases: the leaving of ordinary nodes, the leaving of L2-heads and the leaving of L1-heads. These cases for each scheme are explained in the section III.

## III. Security Analysis and Comparison between Neighbourhood and Hierarchical Key Management Schemes

In this section, the neighbourhood key and hierarchical key management scheme for secure group communications are compared under the headings; Security Keys, Node joining process, and Node leaving process.

### *A. Security Keys*

In both schemes the packet during data transmission is encrypted using different cryptographic algorithms depending on the packet function; AES for symmetric encryption, RSA for public key encryption, and Diffie-Hellman scheme to generate keys. Two symmetric keys "128 bits" in the neighbourhood scheme are generated in every node, one is shared with all authenticated neighbours and the other is used to encrypt the payload of the message.

$$\left(\{M\}_{k_s}, \{k_s\}_{k_{nj}}\right) \qquad (1)$$

Where "$M$" is message, "$k_s$" is source key, and "$k_{nj}$" is neighbourhood key of node $j$.

In order to reduce the delay that is incurred by decrypting and re-encrypting the message between forwarded nodes, a node only needs to re-encrypt the source key "$k_s$" with its own neighborhood key before transmission. However the hierarchical key management scheme uses three security keys to deliver data between nodes in a network. Symmetric subgroup keys "$K_{LXGK_S}$" are used for transmission between all nodes that fall under the L1-head subgroup. Additionally, secure data delivery in different subgroups is achieved through symmetric communication keys "$k_c$" and "$k_{DH}$" which only belongs to the source node and the destination node. "$k_{DH}$" is used for the first encrypted packet transmitted.

*Example*: Assume node A in subgroup X would like to send data to node C in subgroup Y:
In node A:

$$\{\mathrm{M}\}_{\mathrm{K_{DH}}} \tag{2}$$

Packet will encrypt and decrypt with "$K_{LXGK_S}$" when transmitted through nodes in the same subgroup. At the first forwarding node in the subgroup Y:

$$\{\{\mathrm{M}\}_{\mathrm{K_{DH}}}\}_{\mathrm{K_C}} \tag{3}$$

After receiving the packet, the first forwarding node in the subgroup Y will decrypt the packet with "$k_C$", then encrypt the decrypted packet with "$K_{LYGK_S}$" and then send the packet to node C

$$\{\{\mathrm{M}\}_{\mathrm{K_{DH}}}\}_{\mathrm{K_{LYGK_S}}} \tag{4}$$

Where $M$ is message, "$k_{DH}$" is the Diffie-Hellman generated key, and "$K_{LYGK_S}$" is the symmetric, Y subgroup key.

The decryption and encryption steps are repeated until the destination node receives this packet.

All nodes in a subgroup have their own public and private keys. In case of any change in subgroup members, the L1-head will encrypt the regenerated Symmetric subgroup keys "$K_{LXGK_S}$" with the public key of each node before sending it to the nodes via the L2-heads in its subgroup.

### B. Node Joining Process

In the neighbourhood key method the authentication process relies on public key certificate that are signed by an offline trusted third party [7][8]. Also nodes can perform authentication with new nodes independently without any coordinate from any other nodes. A new node sends a join request including its own signed certificate to existing nodes. Certificates between nodes will be exchanged after the received node has verified the certificate of the new node. Once the certificates are exchanged, the nodes will exchange symmetric neighbourhood keys using the RSA algorithm. Whenever a node receives request messages from a node for the first time it must update its neighbourhood key store.

Since rebuilding and redistributing of a new neighbourhood key to all nodes is required each time a node joins or leaves the network, the network may take a long time to stabilise. This issue will worsen when many nodes join and leave the network at the same time.

The benefit of the hierarchical key management scheme is mainly based on its hierarchical structure. When a new node joins a subgroup, rekeying is not a global operation. The L2-subgroup head just regenerates the L2-subgroup key "$K_{LXGK_S}$" for this subgroup, which can be relatively few nodes.

### C. Node leaving process

When an authenticated neighbour has not sent a message for a long time in the neighbourhood key scheme it is assumed to have left the network and a new neighbourhood key must be generated and transmitted to its authenticated neighbours. Whereas in the hierarchical key management scheme the leaving of a node falls under three scenarios;

- For the leaving of an ordinary node, the level 2 cluster head regenerates the L2-subgroup key.
- In the case of the L2-head leaving the subgroup. The node with the largest weight value of the remaining ordinary nodes in the subgroup will be selected to be the new L2-head.
- The third case is the leaving of L1-heads; L2-head with the largest weight value of the L2-heads in the subgroup will selected to be the new L1-head.

## IV. Proposed Trust Security Mechanism for Marine Wireless Sensor Networks

This section outlines the proposed smart security technique for Wireless Sensor Network nodes suitable for use in marine coastal environments. The scheme addresses issues highlighted above in [2] and [3] such as the rekeying process and the number of stored keys required in each node. This scheme uses the advanced encryption standard "AES-128" and the public key cryptosystems "RSA-1024" to allow the secure transmission of data between nodes in the network. The pre-distribution of keys is currently been used is the scheme, whereby keys are allocated to all sensor nodes before deployment and securely transferred between nodes using a master key.

### A. General Outline of the Scheme

Each node keeps its own symmetric key, called an adjacent key "$k_{ni}$" that it shares only with its two neighbours in the network "Fig. 3". Also each node must have

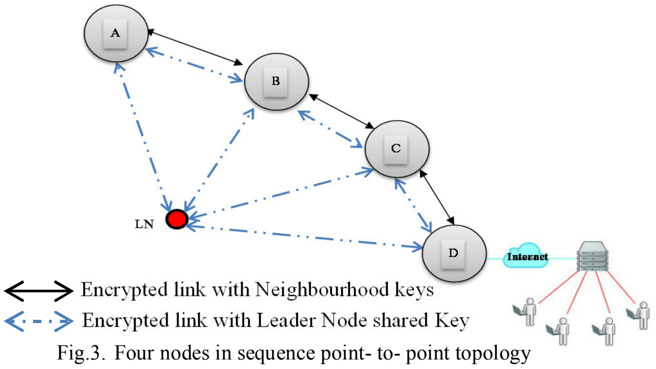

A
B
C
D
LN
Internet
Encrypted link with Neighbourhood keys
Encrypted link with Leader Node shared Key

Fig.3. Four nodes in sequence point- to- point topology

a symmetric master key called the leader node key“$k_{LN}$” that is generated by the master node called Leader Node. The RSA algorithm is used for key management in updating the leader node key.

Each node performs authentication and revocation in coordination with a leader node. In the authentication process a new joining node simply sends a request-to-joint command to any ordinary node included with its Identification. Then the ordinary node sends the received Identification to the leader node. The leader node will verify the Identification based on stored certificates in a trust database of its network members. Once a new node is authenticated from the leader node, adjacent keys are securely exchanged between nodes. These keys are encrypted with the leader node key“$k_{LN}$”.

One of the most important aspects of our security mechanism is the process of revocation and rebuilding the network topology when a node leaves. The leader node monitors the behaviour of all nodes in the network through broadcasting a ‘hello’ message, and all nodes must reply with a response message. If any node does not respond to the ‘hello’ message, the leader will revoke this node and rebuild the network via one of its authenticated neighbours.

As already mentioned, the leader node uses the public key of the authenticated neighbours to securely share a new symmetric leader node key and the identification of the revoked node.

### B. *Key Maintenance and Revocation Process in Proposed Technique*

When a node leaves the network, it should not be able to decrypt the future encrypted traffic [6]. The leader node monitors all nodes’ activities continuously in the network and every node maintains contact with the leader node. In case of any node not responding the leader node will remove this node from its member list and reconnect its neighbours to keeps the network functioning as shows in “Fig. 4”.

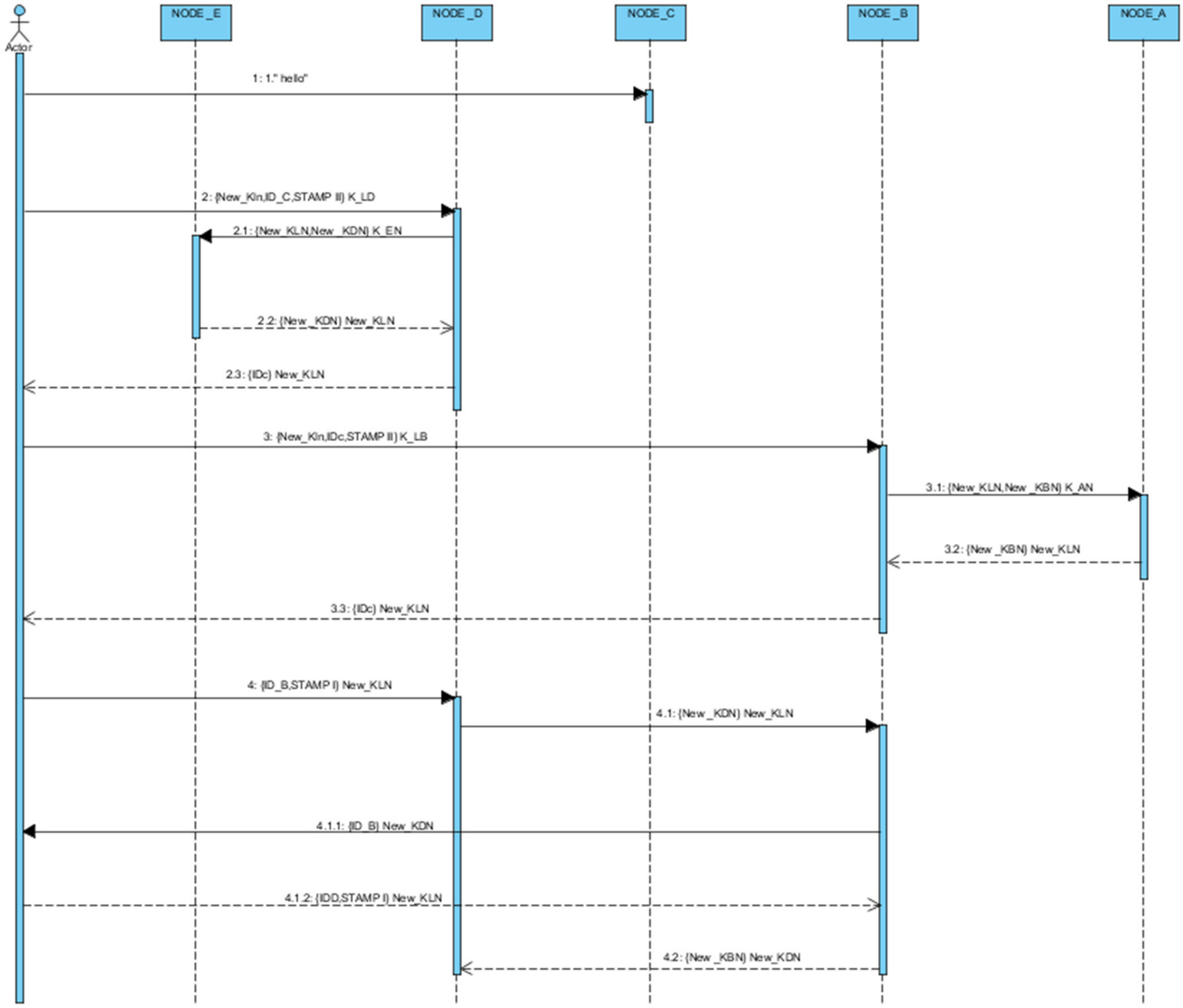


Fig.4. Key maintenance and revocation process: sequence diagram

In "Fig. 4" assume that node C does not response to the 'hello' message and nodes B and D are its two neighbouring nodes.

Notation Used:

New_$K_{LN}$: New shared secret key between Leader and nodes "128bits".
ID*x*: A unique Identification of node x.
$STAMP_{II}$: Part of message means the included ID is revoked.

New_$K_{DN}$: New shared secret key between D and its neighbours "128bits".

New_$K_{BN}$: New shared secret key between B and its neighbours "128bits".
$STAMP_I$: Part of message means the included ID is authenticated.
$K_D$: Public Key of D.
$K_B$: Public Key of B.

Messages Exchanged:
1: "Hello"
2: $\{New\_K_{LN}, ID_C, \mathrm{STAMP}_{II}\}_{K_D}$.
2.1: $\{New\_K_{LN}, New\_K_{DN}\}_{K_{EN}}$.
2.2: $\{New\_K_{LN}\}_{NEW\_K_{DN}}$.
2.3: $\{ID_C\}_{NEW\_K_{LN}}$.
3: $\{New\_K_{LN}, ID_C, \mathrm{STAMP}_{II}\}_{K_B}$.
3.1: $\{New\_K_{LN}, New\_K_{BN}\}_{K_{AN}}$.
3.2: $\{New\_K_{LN}\}_{NEW\_K_{BN}}$.
3.3: $\{ID_C\}_{NEW\_K_{LN}}$.
4: $\{ID_B, \mathrm{STAMP}_I\}_{NEW\_K_{LN}}$.
4.1: $\{New\_K_{DN}\}_{NEW\_K_{LN}}$.
4.1.1: $\{ID_D\}_{NEW\_K_{LN}}$.
4.1.2: $\{ID_D, \mathrm{STAMP}_I\}_{NEW\_K_{LN}}$
4.2: $\{New\_K_{BN}\}_{NEW\_K_{DN}}$.

### C. Description of key maintenance:

After the relationships among the nodes are re-established, all nodes must send a response to the 'hello' message to the leader node. The leader node knows the certificates of all the nodes in the network. Assuming that node C does not response to the 'hello' message, then the LN must remove it from the network. Due to the point-to-point topology the nodes that are positioned before and after the revoked node C must be securely reconnected. The scenario for the revocation of node C "Fig. 4" is described below.

Step 1: Node C does not response to 'hello' message.
Step 2: After the LN has verified all node responses, the leader node will send an encrypted message to node D. This message includes a new leader node master key "$New\_K_{LN}$" and the Identification of the revoked node C. Node D will update and share its adjacent key "New_$K_{DN}$" and new leader node key "$New\_K_{LN}$" with authenticated node E, its neighbour. Node D will confirm this step by sending Identification of the revoked node "C" encrypted with" $New\_K_{LN}$" to leader node.

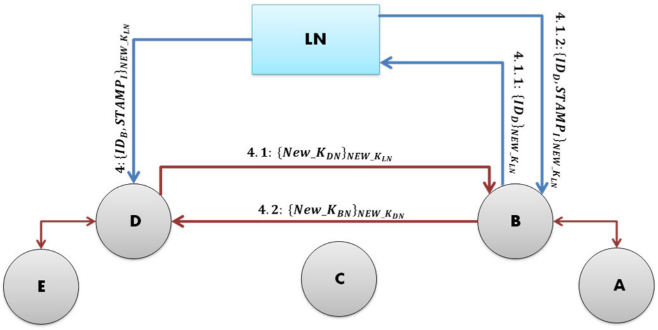

LN
4: $\{ID_B, STAMP_I\}_{NEW\_K_{LN}}$
4.1.1: $\{ID_D\}_{NEW\_K_{LN}}$
4.1.2: $\{ID_D, STAMP_I\}_{NEW\_K_{LN}}$
4.1: $\{New\_K_{DN}\}_{NEW\_K_{LN}}$
4.2: $\{New\_K_{BN}\}_{NEW\_K_{DN}}$
D
B
E
C
A

Fig.5. Rebuilding the network topology after node C revocation

Step 3: After receiving the revoked message of node C, node B will update its adjacent key and share both "$New\_K_{BN}$" and" $New\_K_{LN}$" with its authenticated node A.
Step 4: leader node will coordinate the authentication process between B and D as shown in "Fig. 5". Initially, leader node will send the Identification of B and $STAMP_I$ to node D in order to reconfigure the network. Then nodes D and B will mutually exchange their shared keys using the master key" $New\_K_{LN}$".

### D. Advantages

This approach provides a number of advantages in comparison with [2] and [3]. Firstly, each node in [2] performs authentication independent of and without coordination with other nodes. An exchange and verification of certificates between neighbours in the network occurs only when needed. However management of the authentication process in [3] is occurring via the L2-head and the L1-head.

In our scheme the authentication process is coordinated by the leader node LN. Ordinary nodes store only information of their neighbours which leads to reducing the number of keys stored in every node and also the security risks involved in storing large number of network keys. Furthermore, when a new node joins the network, authenticated nodes do not need to regenerate their keys if they are not a neighbour of the new node. After the new node is verified by LN, it will exchange adjacent keys with its neighbours "Fig. 5". This increases the life of the node as well as lifespan of the entire WSN as it only communicates with its closest nodes. The second advantage is that the encryption and decryption operation during data transmission in [3] is occurring through" $K_{DH}$", subgroup keys and communication keys. Therefore, it has a longer transmission time than our scheme, which encrypts and decrypts only the part of the source key when a message is forwarded. The third and important advantage of this security mechanism is updating the shared key, where only current neighbours of a revoked node will regenerate their symmetric keys. The leader node will distribute a new master key through the trustworthy nodes. In the network, ordinary nodes are deployed in a line topology, and the distance between every two neighbours is around 1500m. In order to cover a range of up to 7000m, in this scheme, we have used XBee-802.15.4-Pro/2.4GHz integrated with Waspmote. This advantage will lead network to securely reconnect in case of three neighbour nodes are revoked.

## V.CONCLUSION

The overall objective of this work is to design a smart security technique for Wireless Sensor Network nodes that can successfully operate in marine coastal environments. We address some potential drawbacks of two existing key management schemes that would be considered suitable and combine their advantages. These protocols used symmetric-key and public-key based key transport protocols for the provision of authentication between nodes. However, both schemes require updating all shared keys whenever the membership in the network changes. The time required to build and distribute new keys will lengthen the time it takes to establish a stable topology in comparison with our proposed scheme which restricts key update to the neighbours of the leaving node. An implementation of the technique is currently being performed on the Waspmote sensor platform and it is hoped that some measurements will be available for the conference presentation.

## VI. REFERENCES

[1] S. Babu, A. Raha, and M. Naskar, "Trust Evaluation Based on Node's Characteristics and Neighbouring Nodes' Recommendations for WSN" Wireless Sensor Network. vol. 6, August 2014, pp. 157-172, http://dx.doi.org/10.4236/wsn.2014.68016.

[2] J. Liebeherr, and G. Dong, "An overlay approach to data security in ad-hoc networks", Ad Hoc Networks, Elsevier, 5th July 2006, pp.1055-1072, http://dx.doi.org/10.1016/j.adhoc.2006.05.017.

[3] V. Wang, and S. Fang," A hierarchical key management scheme for secure group communications in mobile ad hoc networks",Ad Hoc Networks, Elsevier, 23 January 2007, pp. 1667-1677, doi:10.1016/j.jss.2006.12.564.

[4] S. Dhurandher, and G .Singh,. 2005. "Weight based adaptive clustering in wireless ad hoc networks". IEEE Personal Wireless Communications, New Delhi, India, Jan. 2005pp. 95 – 100, doi: 10.1109/ICPWC.2005.1431309.

[5] J. Lopez, R. Roman, I. Agudo, and C. Fernandez-Gago, "Trust Management Systems for Wireless Sensor Net- works: Best Practices". Computer Communications, Elsevier vol. 33, June 2010, pp. 1086-1093, doi: doi:10.1016/j.comcom.2010.02.006.

[6] K. Chauhan, and S. Amit, Singh, "Securing Mobile Ad hoc Networks: Key Management and Routing," AdHoc Networking Systems, (IJANS). vol. 2, April 2012.

[7] Y. C. Hu, and A. Perrig, "A survey of secure wireless ad hoc routing," IEEE Security and Privacy, vol. 2, pp. 28- 39, March 2004, doi: http://doi.ieeecomputersociety.org/10.1109/MSP.2004.1.

[8] Y. Sun, Z. Han, and Liu, K.J.R. "Defence of Trust Management Vulnerabilities in Distributed Networks". IEEE Communications Magazine, vol. 46, February 2008, pp. 112-119, doi: 10.1109/MCOM.2008.4473092.

[9] X.B. Zhang, S.S. Lam, H. Liu, "Efficient group rekeying using application-layer multicast" IEEE Distributed Computing Systems, June 2005, pp. 303- 313, doi: 10.1109/ICDCS.2005.27.

DOI: 10.15439/2015F193
ACSIS, Vol. 5

# A New Energy Efficient Cluster based Protocol for Wireless Sensor Networks

Mohamed Eshaftri, Ahmed Y.Al-Dubai, Imed Romdhani, *Muneer Beni Yassien
School of Computing, Edinburgh Napier University, 10 Colinton Road, Edinburgh EH10 5DT, UK
*Department of Computer Science, Jordan University of Science and Technology
Email: {M.Eshaftri; A.Al-Dubai; I.Romdhani; M.BaniYassein}@napier.ac.uk

***Abstract*—In Wireless Sensor Networks (WSNs), clustering techniques are usually used as a key effective solution to prolong the network lifetime by reducing energy consumption among the sensor nodes . Despite many works on clustering in WSNs this issue is still outstanding. However, the most existing solutions suffer from long and iterative clustering cycles. In an attempt to fill in this gap, we propose a new cluster-based protocol, referred to as Load-balancing Cluster Based Protocol (LCP) that introduces a new inter-cluster approach to increase network lifetime. This new protocol rotates continuously the election of the Cluster Head (CH) election in each cluster, and selects the node with the highest residual energy in each round. Extensive simulation experiments show that our proposed approach effectively balances the energy consumes among all sensor nodes and increases network lifetime compared to other clustering protocols.**

***Keywords*—*WSNs; Distributed clustering; lifetime; Routing; Network*.**

## I. Introduction

THE Wireless Sensor Network (WSN) technology has been one of the major avenues of networking and Internet of Things (IoT) due to their potential role in digitising smart physical environments [1]. WSNs composed of a large number of sensor nodes with limited battery power, which can be either densely or sparsely deployed in harsh and extreme environments, such as wild remote areas, natural habitats, and regions with access risk. On one hand, sensor nodes are usually battery-powered with limited operating time, and therefore they are highly sensitive to failure [2] [3]. On the other hand, the design an energy efficient WSNs protocol to prolonging the network lifetime is a challenging task due to the unique nature and strong networking constraints of wireless sensor networks [4].

The research community proposed different routing protocols to optimise the routing process in WSNs. Typically, the routing protocols for WSNs can be classified into three categories: flat, location and hierarchical based routing [5]. In flat routing, all nodes have identical functionality and they work together to sense and route [6]. Location based routing protocols rely on the position information of each node to discover and build optimal routing paths [7]. Compared to the two previous categories, in hierarchical routing approaches, the sensing field is subdivided into a set of administrative domains called clusters [8]. Each cluster has an organised leader or a root node called the ClusterHead (CH). The primary aim of the CH is to collect data from attached and associated downstream nodes and forward it to the best next well-known hierarchical upper level upstream neighbour node. The data is forwarded in a hop-by-hop manner until it reaches the Base Station (BS). The BS can then send the data, using a wired or wireless Internet connection, to an end user located outside the sensing field [9][10].

A number of cluster protocols based on energy efficient have been proposed in the literature [11]. These approaches attempt to minimise energy consumption by reducing the transmission of redundant data. Clustering approaches focus primarily on the communication process during cluster organisation and CH election and neglect the effect of information processing on energy consumption. Hybrid Energy-Efficient Distributed (HEED) is one of the clustering protocols that uses both energy and communication costs to select CHs in a probabilistic manner. This protocol uses different inter-cluster approach in order to reduce energy consumption and to prolong the network lifetime [12].

In this paper, we present a new energy-aware distributed and dynamic clustering protocol, namely A Load-balancing Cluster Based Protocol (LCP). LCP addresses load balancing issues in cluster-based routing approaches. Given that cluster-based protocols require regular re-clustering for balancing energy consumption. However, re-clustering process in interval time for entire network increases the network overhead and consequently decreases the network operation time. The proposed model provides a pre-defined interval of time at the beginning of every round to select the CH. This delays the frequency of the re-clustering messages received from the BS. If the sensor nodes do not receive the BS message, the CHs continues rotating the leadership among them within the same members of cluster by electing the node with the highest residual energy each round. The performance evaluation of LCP is examined in depth and compared to HEED [12], LEACH [13] and R-HEED [14]. Obtained results demonstrate that LCP enhances the network lifetime by 15%.

The rest of the paper is organised as follows. In Section II, we review a set of up-to-date clustering algorithms proposed for WSNs. Section III presents the features of the new LCP protocol. Section IV presents a detailed description of the simulation environment and the simulation results. Finally, section V reviews the entire study and offers conclusions and recommendations for future work.

978-83-60810-66-8/$25.00©2015, IEEE

## II. Related Work

Different cluster-based approaches have been proposed by the research community to address the challenging issues of WSNs. Some of these approaches are as follows:

### A. Low Energy Adaptive Clustering Hierarchy (LEACH)

Heinzelman, et al. proposed the first well-known clustering LEACH protocol [13]. This protocol was targeted at prolonging the lifetime of WSNs and reducing the energy consumption of sensor nodes. From an algorithmic point of view, LEACH is hierarchical, probabilistic, distributed and single-hop protocol. It forms clusters based on the strength of received signal, while CH nodes act as default gateways to the BS, as illustrated in Fig. 1. In LEACH protocol, nodes make autonomous decisions without relying on a centralised third party entity. In addition, all nodes have an equal opportunity to become CHs. Initially, a node generating a random number between (0-1) to be a CH by comparing it with a threshold value T (n), calculated using Equation (1).

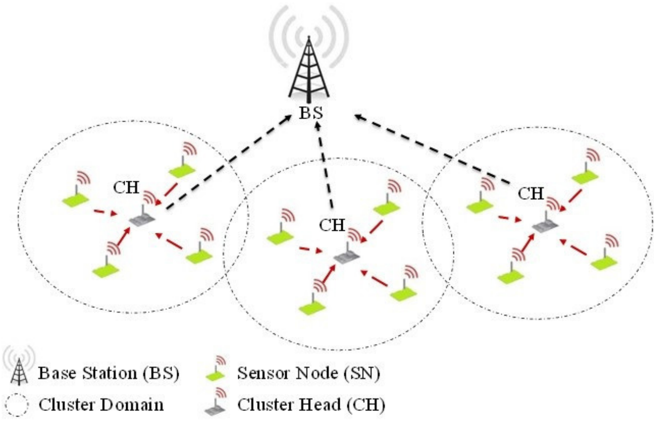


Fig. 1. Basic LEACH topology [11]

Nodes with a random number lower than T (n) then become CHs. Each elected CH broadcasts an advertisement to non-CHs to form a cluster.A non-CH node selects a CH that expending the least energy for communication.

$$T\left(n\right)=\begin{cases}\frac{p}{1-p\left(rmod\frac{1}{p}\right)} & ifn\in G\\ 0 & \end{cases} \quad (1)$$

Where p is the desired percentage of nodes to be CH; r is the current round; G is the set of nodes that have not been cluster heads during the last 1/P rounds.

Generally, LEACH provides a good model for energy consumption while providing an equal probability for node to be elected CHs. Once chosen as a CH, a sensor node cannot be reselected in a subsequent round. Moreover, LEACH avoids unnecessary collisions between CHs because it uses the Time Division Multiple Access (TDMA) protocol. Despite its generally good performance, LEACH also has some clear limitations. It uses single-hop communication which limits its scalability. In addition, the probabilistic election mechanism of CHs may lead to either high concentrations of CHs in one part of the network, or to orphan nodes (nodes without CHs in their neighbourhood.

### B. LEACH-Centralised (LEACH-C)

To address the shortcomings of LEACH with respect to determining each CHs location and number rounds,a new version of LEACH, named LEACH Centralised (LEACH-C) proposed [15]. In the new version, the BS decides which sensor nodes are eligible to become CHs and form a cluster. Each node transmits its location and energy level to the BS, which in return calculates the average energy level for the network and eliminates the nodes with remaining energy levels below this average, to form the set of CHs for that round. In the centralised algorithm the energy load is distributed among all nodes equally, where the numbers of CHs are specified and the network is divided into optimum and equal sized clusters. However, the construction of clusters with an equal number of nodes in each cluster is not guaranteed in this protocol, and it is not always possible for nodes distant from the BS to send information about their status.

### C. CHybrid Energy-Efficient Distributed (HEED)

O. Younis et al. [12] Introduced HEED clustering protocol. In this protocol, the authors enhanced LEACH protocol by introducing two basic parameters to elect the CHs. The first main parameter concerns the remaining energy of each node, and the second parameter is the intra-cluster "communication cost". For example, the cost can be a function of neighbour proximity or cluster density, that can calculated using Equation (2). Unlike LEACH, HEED protocol the CH nodes are not randomly selected. Only nodes with high levels of remaining energy can become CH nodes. In addition, when two nodes are within each other's cluster range, the probability of both becoming cluster heads is negligible. In comparison to LEACH, in HEED, the CHs are well distributed throughout the network. However, this protocol cannot fix the cluster count in each round. In addition, the energy consumption is not balanced, because more CHs could be generated more than expected, which creates massive overheads due to multiple election rounds.

$$CH_{prob}=C_{prob}\times\frac{E_{residual}}{E_{max}} \quad (2)$$

Where: $C_{prob}$ is an initial percentage of cluster heads among all n nodes, $E_{residual}$ is the estimated current energy of the node, $E_{max}$ is the referenced maximum energy (corresponding to a fully charged battery).

### D. Distributed Energy-Efficient Clustering (DEEC)

Li Qing et al. [16] proposed the DEEC algorithm for WSN to improve HEED performances. In DEEC, the CHs are selected with a probability based on the residual energy of each node and the average energy of the network. The authors of this algorithm assumed that nodes would have different amounts of energy. With the adaptive values, the sensor nodes determine their role probabilistically in each round. The main drawback of DEEC is that each node demands global knowledge from the network, which increases the overheads.

### *E. Rotated Hybrid, Energy-Efficient and Distributed (R-HEED)*

W. Mardini et al. [14] introduced R-HEED. With this protocol, the authors improved the performance of HEED by applying a different inter-cluster approach. The new approach conducts the cluster reformation based on certain rules. At the start of setup phase on every round, the CHs node must delay for period of time waiting for a cluster reformation message from the sink .if the cluster reformation message not received, each cluster persevere with rotating the cluster head task in the same cluster. However, randomly rotating the CH does not take into account energy consumption.

### *F. Distributed weight based energy efficient hierarchical clustering protocol (DWEHC)*

P Ding et al. [18] proposed a new protocol called DWEHC that improves HEED performances. Their primary aim was to to improve energy consumption by forming balanced cluster sizes and improving intra-cluster routing. Each sensor node begins broadcasting its (x, y) coordinates to search for its neighbour. After finding neighbouring nodes in its area, each node calculates its weight. Weight is the only parameter calculated locally and used for CH election; it is represented by weight in DWEHC as defined by Equation 3.The node with the largest weight is selected as a CH. The ordinary nodes become child nodes by joining CH. The nodes at this stage, are considered first level members because they have a direct link to the CH. As the child nodes are further divided into levels (level 1, level 2, etc.) the total number of levels is seen to depend on the cluster range and the minimum CH energy. Like HEED, DWEHC is a fully distributed clustering protocol with a more balanced CH distribution. In addition, its clustering process does not rely on network size. However, this protocol cannot increase its energy efficiency give its inter-cluster communication function and the large control message overheads.

$$W_{weight(s)} = \left( \sum_{u \in N\alpha, c(s)} \right) \frac{(R-d)}{6R} \times \frac{Eresidual\,(s)}{Einitial\,(s)} \tag{3}$$

Where: R is the cluster range, d is the distance from node s to neighbouring node; $E_{residual(s)}$ is the residual energy in the nodes; $E_{initial(s)}$ is the initial energy in the nodes.

### *G. Power-Efficient and Adaptive Clustering Hierarchy (PEACH)*

The majority of existing clustering protocols consume large amounts of energy, incurred by cluster formation overheads and fixed-level clustering. This is especially true when sensor nodes are densely deployed. To address this problem, Sangho Yi et al. [17] proposed PEACH protocol to reduce energy consumption, that improve the network lifetime.In PEACH protocol, a node selected as CH when the packet received was for the that node . When the packet is received by a different node and is not the destination for the packet, the node will join the destination of that packet. Simulation results showed that PEACH consumes lower energy and prolonges the network lifetime comparative to the LEACH, and HEED protocols. However, the network is not very scalable, because all the nodes must have global knowledge of the network.

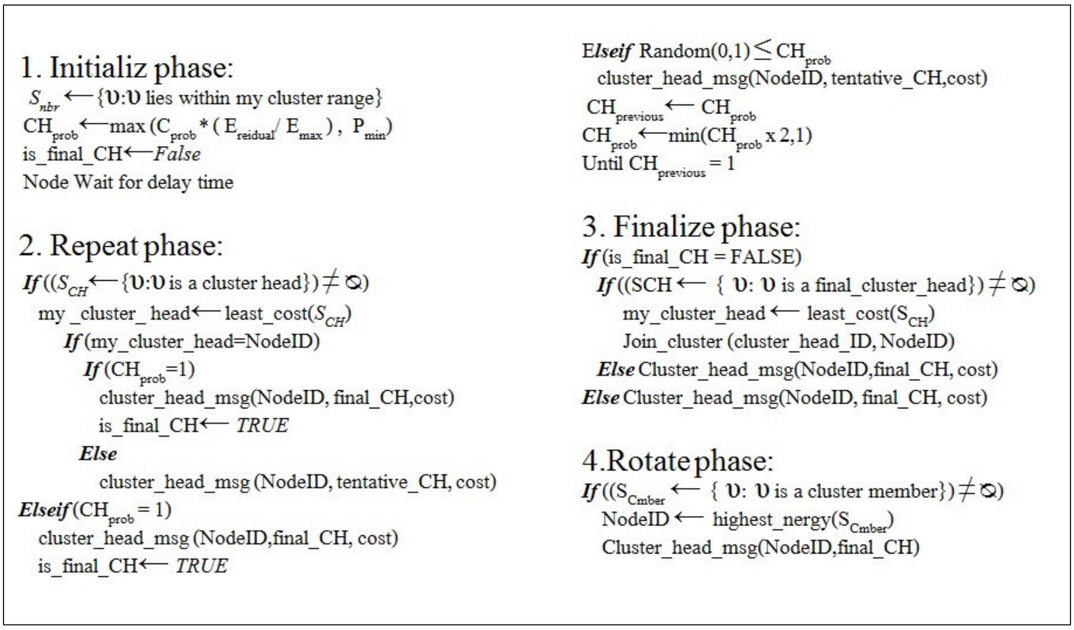
## III. THE LCP CLUSTERING PROTOCOL

The proposed scheme builds on the success of the HEED protocol. The clustering phase of the HEED protocol has been modified to make it more energy-efficient. The modified version is named A Load-balancing Cluster Based Protocol (LCP). In LCP, the clustering operation is divided into several rounds, each round has two phases: the setup and the steady-state phase. LCP is similar to HEED in terms of the following features:

- The elected CHs sent advertisement message within cluster range.
- The cluster formation "setup phase" finish in O(1) iterations.
- Each node become member only to one cluster and communicates directly with its CH.
- Through the cluster formation process, Nodes can become either a tentative_CH or a final_CH, or it can be covered.
- At the end of the clustering procedure, CHs node forms a network backbone. Thus, the data is forwarded in hop-by-hop through CHs until it reaches the BS.
- The steady state phase for LCP protocol is alike HEED, and CH election is done as part of an iterative process.

The setup phase is divided into four phases: 1) Initialise phase, 2) Repeat phase, 3) Finalise phase and, 4) Rotation phase. The following steps describe the proposed phases, which are illustrated in Fig. 2.

1) Initialise phase: At the beginning of this phase, nodes exchange their information with neighbours in order to computes its cost. Unlike HEED, the costs are exchanged through the cluster head message. The LCP algorithm sets an initial percentage of node become Cluster head $C_{prob}$. Thus, each sensor node establishes its probability of becoming a CH based on the reaming energy $CH_{prob}$ according to HEED.
2) Repeat phase "Main Processing": Each node in this phase is subject to a delay time, in which it can decide whether the node will be elected as candidate CH node "tentative_CH". If the node not elected as tentative_CH, it will declare itself a cluster head node "final_CH". The final_CH node broadcast a cluster_head_msg (Node ID, tentative, cost) within cluster range.
3) Finalize phase: During this phase, most sensor nodes declare itself either a cluster head node or a member node. If node received a cluster head advertise message, it will join the final_CH with the lowest cost. The node neither final_CH or has not received cluster head advertise message, it will declare itself a final_CH node.



```
1. Initializ phase:
  $S_{nbr} \leftarrow \{\upsilon : \upsilon$ lies within my cluster range$\}$
  $CH_{prob} \leftarrow \max(C_{prob} * (E_{reidual} / E_{max}),\ P_{min})$
  is_final_CH ← False
  Node Wait for delay time

2. Repeat phase:
  If (($S_{CH} \leftarrow \{\upsilon : \upsilon$ is a cluster head$\}) \neq \emptyset$)
    my _cluster_ head ← least_cost($S_{CH}$)
      If (my_cluster_head=NodeID)
        If ($CH_{prob}$=1)
          cluster_head_msg(NodeID, final_CH,cost)
          is_final_CH ← TRUE
        Else
          cluster_head_msg (NodeID, tentative_CH, cost)
  Elseif ($CH_{prob}$ = 1)
    cluster_head_msg (NodeID,final_CH, cost)
    is_final_CH ← TRUE
  Elseif Random(0,1) ≤ $CH_{prob}$
    cluster_head_msg(NodeID, tentative_CH,cost)
  $CH_{previous} \leftarrow CH_{prob}$
  $CH_{prob} \leftarrow \min(CH_{prob} \times 2, 1)$
  Until $CH_{previous}$ = 1

3. Finalize phase:
  If (is_final_CH = FALSE)
    If ((SCH ← { $\upsilon$: $\upsilon$ is a final_cluster_head}) $\neq \emptyset$)
      my_cluster_head ← least_cost($S_{CH}$)
      Join_cluster (cluster_head_ID, NodeID)
    Else Cluster_head_msg(NodeID,final_CH, cost)
  Else Cluster_head_msg(NodeID, final_CH, cost)

4.Rotate phase:
  If (($S_{Cmber} \leftarrow \{\upsilon : \upsilon$ is a cluster member$\}) \neq \emptyset$)
    NodeID ← highest_nergy($S_{Cmber}$)
    Cluster_head_msg(NodeID,final_CH)
```

Fig. 2. A Load-balancing Cluster Based algorithm.

4) Rotate phase: After elect the CHs node and form clusters in the first round, each CH constructs a turning schedule for its member when it becomes a CH. The turns are sorted based on residual energy in the sensor node. Node with the highest residual energy will be the first candidate to become a CH for next round. Therefore, at begging of the next round unlike HEED protocol is not necessary to re-cluster the network. Node within the same cluster in subsequent rounds continues rotating the CH role between them, by selecting the node with the highest residual energy every round. When first cluster finishes the rotating process,it inform the BS by sending re-form cluster message via multi-hop route. BS re-broadcast the message among the nodes inform them to start a new cluster process, See Rotate phase in Fig. 3.

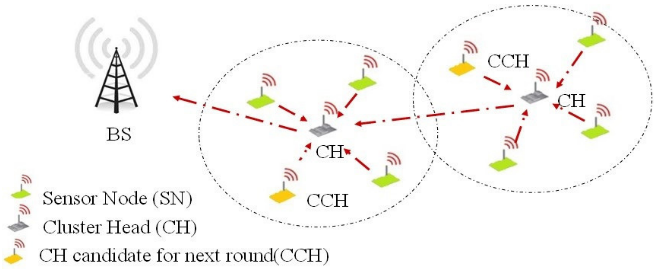


Fig. 3. A Load-balancing Cluster Based Protocol Rotate phase.

## IV. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the LCP mechanism by using open source Castalia simulator [19]. We consider a sensor network, composed of (100-350) sensor nodes, which are randomly deployed in a playground of 200mX200m square region. All sensor nodes are fixed and homogeneous and with limited stored energy. Nodes are not equipped with GPS-capable antennae. The BS is placed at the center of the sensor field. The energy consumption for each sensor node is calculated by data transmission and aggregation per round. The energy efficiency of LCP is compared against LEACH, HEED, R-HEED. Simulation parameters are given in Table I.

TABLE I
PARAMETER SETTINGS

| Parameter | Value |
|---|---|
| Deployment field | 200 X 200 m |
| Data packet size | 200 bytes |
| Control packet | 25 bytes |
| Number of node | 100-350 |
| Initial cluster radius (RC) | 25m |
| Sink position | (0,0) |
| Initial energy | 25J |
| Threshold distance ($d_0$) | 75m |
| Deployment method | Uniform, Random |
| Rotated time ($T_r$) | 20 Sec |
| Radio model | CC4220 |

We use in this paper the residual energy matric,and the network lifetime metric to evaluate the performance of our protocol. The residual energy metric is computed by the average energy remaining in all nodes at a specific round. The network lifetime metric is based on WSN applications require.For example , applications require that all node must work to ensure the network has good coverage. Thus, the network lifetime metric for these applications should be measured according to the lifetime of the shortest-living node. Some Other applications they only need a specific percentage of nodes have to remain alive to achieve the applications requirement [17]. Therefore, the network lifetime in our protocol ,is measured by following three different metrics.

1) First Node Die (FND): is defined as time elapsed in rounds until the first node has consumed all available energy.
2) Half Nodes Die (HND): is defined as time elapsed in rounds until half of the nodes have consumed all available energy stores.
3) Last Node Dies (LND): is defined as time elapsed in rounds until all the nodes have exhausted their entire energy supply.

The "round" definition in our paper refer to the time interval in seconds befor the network statr a new cluster process.Therfore , no difference between the round concepts in LCP ,and HEED in terms of time.In LCP protcol we specified a round time of 20 seconds.

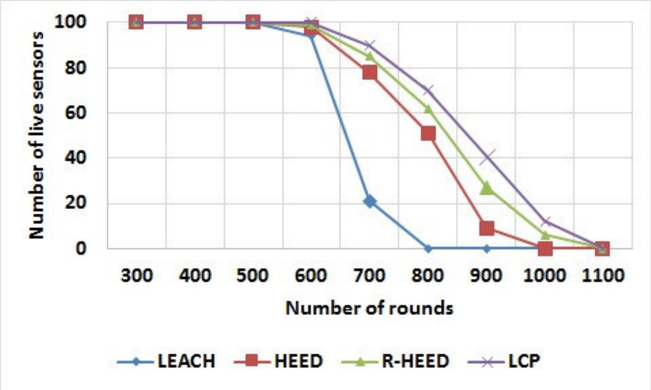


Fig. 4. Number of alive sensors Vs numbers of rounds for LEACH, HEED, R-HEED and LCP.

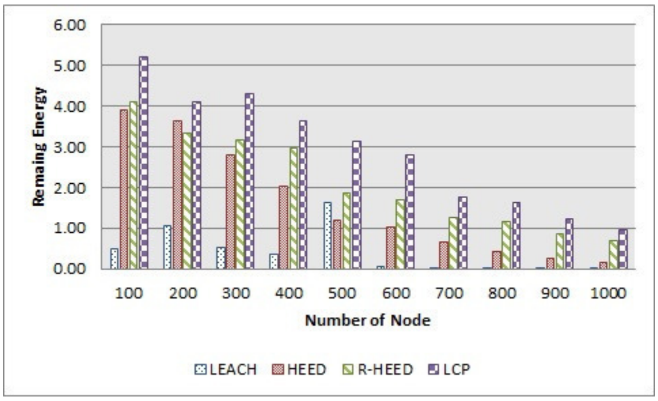


Fig. 5. Total remaining energy in LCP in comparison with HEED, R-HEED.

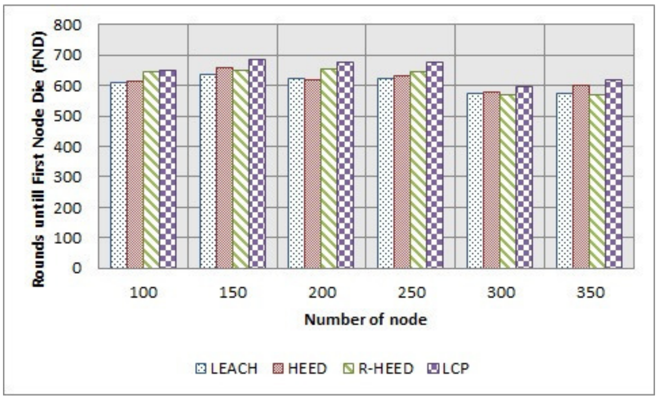


Fig. 6. Comparing LEACH, HEED, R-HEED and LCP using different number of node for FND metric.

Fig. 4 . demonstrates the total number of nodes remaining alive following the simulation round. LCP increases the network lifetime compared to its peers. Fig. 5. demonstrates the relationship between the remaining energy and the number of nodes. It is evident that LCP consumes the least amount of energy. Furthermore, how the increasing number of the nodes affects the lifetime of each protocol has been evaluated. . Fig.6 demonstrates the network lifetime until the first node dies, when the number of nodes varies between (150- 350).

Figure.7 and 8 also reveals the same comparisons to define the network lifetime.In Fig.7 shows the number of rounds until half of the nodes die, while Fig.8 show the number of rounds until the last node dies. It is evident that the network lifetime improves when the number of nodes increases in all protocols. Figs 6, 7 and 8 show that in all three cases LCP protocol performs better than the rest of the protocols. This advancement is caused by the rotating process of the cluster heads within the same cluster.

Consequently, the rotating process leads to reduc the energy consumption among the nodes, and increasing the network lifetime. It can be easily observed from the simulation results that when the number of the nodes increases the percentage improvement also increases. Therefore, it can be reasoned that when Increase the amount of nodes it reduce the energy consumed during the setup phase.Thus, the energy saved as a result of this new clustering scheme will be maximised, which will improve the networkâĂŹs lifetime.

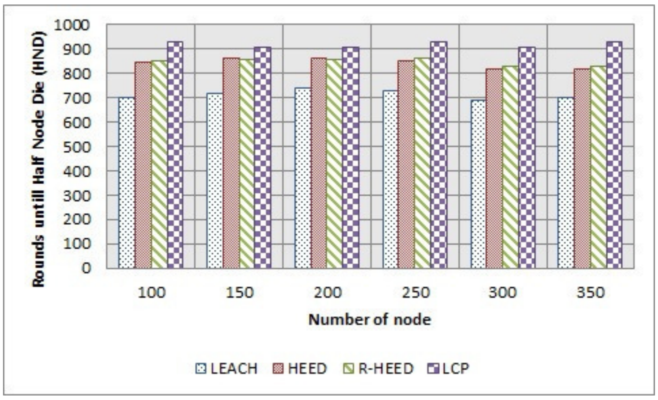


Fig. 7. Comparing LEACH, HEED, R-HEED and LCP using different number of node for HND metric.

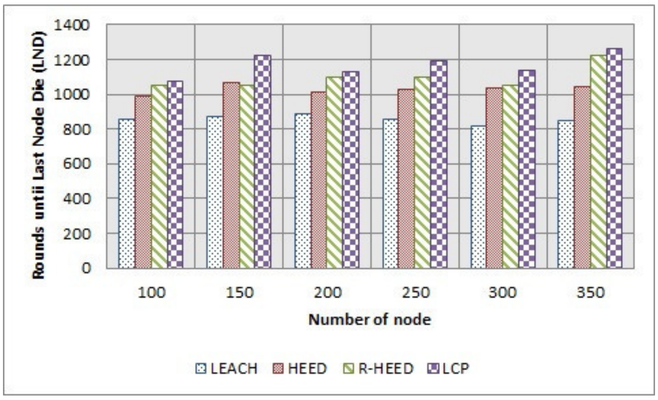


Fig. 8. Comparing LEACH, HEED, R-HEED and LCP using different number of node for LND metric.

## V. Conclusions and Future Work

In this paper, the clustering scheme Load-balancing Cluster Based Protocol (LCP) for wireless sensor networks was proposed as a more energy-efficient protocol. The main contribution of the LCP protocol is its ability to continue rotating the cluster head (CH) role between nodes within the same cluster, by selecting the node with the highest residual energy to become a CH for the next round. We compared and evaluated the LCP protocol performance with well-known Energy Efficient clustering protocols, which is have the same aim increase the network lifetime. The simulation results showed that LCP protocol significant balance the energy consumption among the entire node and achieves an obvious improvement to the network's lifetime by 15%.

Finally, we evaluated our protocol performance in term of energy consumption. Hence our future work we plane to investigate the performance of LCP according to other networking metrics, such as packet delivery ratio and end-to-end delay.

## References

[1] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos,*"Context aware computing for the internet of things: A survey,"* IEEE Commun. Surv. Tutorials, vol. 16, no. 1, pp. 414-454, 2014. http://dx.doi.org/10.1109/SURV.2013.042313.00197

[2] Z. Manap, B. M. Ali, C. K. Ng, N. K. Noordin, and A. Sali,*"A review on hierarchical routing protocols for wireless sensor networks,"* Wireless Personal Communications, vol. 72, no. 2, pp. 1077-1104, 2013. http://dx.doi.org/10.1007/s11277-013-1056-5

[3] R. N. Enam , M. Imam and R. I. Qureshi *"Energy Consumption in Random Cluster Head selection Phase of WSN"* 2012 IACSIT Hong Kong Conferences IPCSIT vol. 30 (2012) IACSIT Press, Singapore.

[4] M. Shokouhifar and A. Jalali, *"A new evolutionary based application specific routing protocol for clustered wireless sensor networks,"* AEU - Int. J. Electron. Commun., vol. 69, no. 1, pp. 432-441, 2015. http://dx.doi.org/10.1016/j.aeue.2014.10.023

[5] A. Rahman, S. Anwar, I. Pramanik, and F. Rahman, *"A Survey on Energy Efficient Routing Techniques in Wireless Sensor Network,"* 15th International Conference on Advanced Communications Technology (ICACT 2013). Pyeong Chang, South Korea, pp. 200-205, January 2013.

[6] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, *"Directed diffusion for wireless sensor networking,"* Networking, IEEE/ACM Trans., vol. 11, no. 1, pp. 2-16, 2003. http://dx.doi.org/10.1109/TNET.2002.808417

[7] Y. Xu, J. Heidemann, and D. Estrin, *"Geography-Informed Energy Conservation for Ad Hoc Routing,"* Proc. ACM/IEEE Intl Conf. Mobile Computing and Networking (MOBICOM), pp. 70-84, July 2001. http://dx.doi.org/10.1145/381677.381685

[8] S. Gupta, N. Jain, and P. Sinha, *"Clustering Protocols in Wireless Sensor Networks: A Survey,"* in International Journal of Applied Information Systems (IJAIS) âĂŞ ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA vol. 5, no.2, pp.41-50, January 2013.

[9] S. P. Singh and S. C. Sharma, *"Cluster Based Routing Algorithms for Wireless Sensor Networks,"* International Journal of Engineering & Technology Innovations (IJETI), vol. 1, no 4, November 2014.

[10] A. Marya, A. Kumar, and C. Mohan, *"Energy efficient heterogeneous leach with enhanced stability for wireless sensor network systems,"* International Journal of Applied Science and Engineering Research (IJASER), vol. 3, no. 5, pp. 920-931, 2014. http://dx.doi.org/10.6088/ijaser.030500002

[11] D. J. Dechene, A. El Jardali, M. Luccini, A. Sauer, *"A Survey of Clustering Algorithms for wireless Sensor Networks,"* Information and Automation for Sustainability (ICIAFS), 4th International Conference , Publication Year: 2008 , pp. 295-300.

[12] O. Younis and S. Fahmy, *"HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks,"* IEEE Trans. Mobile Computing, vol. 3, no. 4, pp. 366-379, Oct.-Dec. 2004. http://dx.doi.org/10.1109/TMC.2004.41

[13] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, *"Energy-Efficient Communication Protocol for Wireless Microsensor Networks,"* Proc. 33rd Hawaii Intl. Conf. Sys. Sci., Jan. 2000. http://dx.doi.org/10.1109/HICSS.2000.926982

[14] W. Mardini, M. B. Yassein, Y. Khamayseh, and B. a. Ghaleb, *"Rotated hybrid, energy-efficient and distributed (R-HEED) clustering protocol in WSN,"* WSEAS Trans. Comm , vol. 13, pp. 275-290, 2014.

[15] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, *"An Application-Specific Protocol Architecture for Wireless Microsensor Networks,"* IEEE Trans. Wireless Comm., vol. 1, no. 4, pp. 660- 670, Oct. 2002. http://dx.doi.org/10.1109/TWC.2002.804190

[16] L. Qing, Q. Zhu, and M. Wang, *"Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks,"* Computer Commun., vol. 29, pp. 2230-2237, 2006. http://dx.doi.org/10.1016/j.comcom.2006.02.017

[17] S. Yi, J. Heo, Y. Cho, and J. Hong, *"PEACH: Power-efficient and adaptive clustering hierarchy protocol for wireless sensor networks,"* Comput. Commun., vol. 30, no. 14-15, pp. 2842-2852, Oct. 2007. http://dx.doi.org/10.1016/j.comcom.2007.05.034

[18] P. Ding, J. Holliday, and A. Celik, *"Distributed energy-efficient hierarchical clustering for wireless sensor networks,"* in Proceedings of the IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS âĂŹ05), vol. 3560, pp. 322-339, June 2005. http://dx.doi.org/10.1007/11502593_25

[19] A.Boulis, *"A Simulater for Wireless Sensor Networks and Body Area Network,"* (Castalia), March 2011, https://castalia.forge.nicta.com.au, (Accessed: 9 February 2014).

DOI: 10.15439/2015F356
ACSIS, Vol. 5

# Application of WSN for Smart Power Metering to Avoid Cheating on Electric Power Consumption at Places with Shared Power Sources

Michal Hodoň
University of Žilina, Univerzitná 8215/1,
01026 Žilina, Slovakia, Email:
michal.hodon@fri.uniza.sk

Samuel Žák, Martin Kopkáš,
Peter Ševčík, Martin Húdik
University of Žilina, Univerzitná
8215/1, 01026 Žilina, Slovakia,
Email: {samuel.zak, martin.kopkas,
peter.sevcik, martin.hudik
}@fri.uniza.sk}

***Abstract* — Application scenarios of WSN implementations are really broad. Due to their typical characteristics, such advantageous power consumption, wide communication coverage and relatively broad sensing possibilities, many application fields can be covered. In this paper, the problematic of power metering is targeted. Places with shared power sources, where electricity usage is calculated equally according to the number of users, are often misused by single users for illegitimate power usage. Special WSN was developed and implemented in a scenario of multi-store parking garage to solve this problem.**

## I. Introduction

Wireless sensor networks (WSNs), as the front-end representative of Internet-of-Things framework, are still more and more implemented in real applications of our present life. When considering the typical WSNs implementation advantages, as their scalability, capability, integration simplicity or operability easiness, their application usage is really broad. The typical applications cover mainly the problematic of environment monitoring [1], [2], [3], e.g. with particular focus on the specific parameters of the nature [4], [5], [6], [7]. Other typical abundant application field is the area of Intelligent Transportation systems, where WSNs already got used to the climate, see [8], [9], [10], [11]. Secondary areas, with no such a high WSN implementation capabilities, are military [12] and eHealth [13], where some strict security/safety rules make the accommodation of central applications difficult. However, by implementation of specific approaches, as [14], [15], [16], [17] and [18], the application fields covered by WSNs could be wider, e.g. agriculture, industry, home automation… Especially the last two areas offer interesting opportunities of WSN indoor implementations, the application problematic on which this paper is geared to.

Real applications of WSNs indoors are the challenging tasks, which due to the certain specific conditions - stronger multipath, blocked line-of-sight, interference,..- attract not so many realizations. The possibility of energetic self-sufficiency - as of the WSN biggest asset - omission, due to the unlimited energy sources indoors, makes utilization of WSNs indoors a lot ineffective. However, there exist some indoor applications, where utilization of application-specific WSNs can be advantageous.

In [19], authors introduced a smart home system which could supervise household appliances remotely and realize real-time monitoring of home security status through mobile phone offering the possibilities of real-time monitor of the house status. In [20], the paper described a practical design and implementation of WSN for controlling and monitoring system in multi-storey building implemented in the university building. Authors in [21] proposed a service oriented architecture for development of an enterprise networking environment that was used for integrating facilities management applications and building management systems for the purpose of the enterprise environment control. The application of WSN brought in this case significant benefits due to the lack of wiring installations allowing flexible positioning of the sensors, especially when building retrofitting was concerned. In [22], the paper presented a WSN-based smart monitoring and control system for building automation. The system was developed with focus on significant reduction of energy consumption of the building under control. The data gathered by WSN were used in order to calculate and estimate the requirements for heat corrections, with respect to ventilation and weather predictions. Non-traditional usage of Facebook as a platform for such kind of indoor sensor network monitoring system was examined in [23]. For the purposes of building emergency-management system, an indoor emergence guidance algorithm based on WSN to guide people to a building exit gate while helping them avoid a hazardous area was introduced in [24]. WSN was responsible for real-time monitoring of the environment, when emergency event was detected the proposed algorithm was applied. The proposed guiding mechanism was successfully examined and verified through the real test bed implementation.

978-83-60810-66-8/$25.00©2015, IEEE

Developments in the field of WSNs, coming out from the MEMS system designs advancements, allow realizing special WSN-based-systems for smart grids too. Smart grid technology is one of the recent developments in the area of electric power systems that aid the use of non-conventional sources of energy in parallel with the conventional sources of energy [25]. By realizing the two-way communication between the utility and the smart meters in the houses, smart grid enables a time-of-use tariff to reduce peak load through incenting residents to adopt a more efficient usage of domestic appliances [26]. However, by application of specific WSN-based smart grids, implementation of local networks, allowing bidirectional communication of local distribution grid with related electric meter. Reliable and, especially, more efficient power metering could be in this case reached. Some applications of WSN-based smart grids can be seen in [25], [26], [27] or [28], where different application scenarios were discussed: e.g. the designs of power sensing nodes to calculate the power for any kind of loads; the designs of special energy management schemes connecting non-urgent appliances to the smart meter through the wireless sensor networks; the architecture of reliable communication protocol for the harsh WSN-based smart grids environment; the architecture of WSN as a platform to enable detailed household monitoring interfaced with smart-grid through a single input to the building control system.

## II. Shared-Power sources Metering

In this paper, the problematic of local power metering at special places with the shared power meters and, therefore, shared bills for power sources, is targeted. Places with shared power sources, where electricity usage is calculated equally according to the number of users, are often misused by single users for illegitimate power usage. The creativity of the individuals is unlimited. The people are able to risk their health to spare a little amount of money modifying the installations to provide the possibilities of power feeding for their gadgets. The most common case is to extend the pure lighting installations with sockets, where their rarely used devices, which have to run permanently, are attached, e.g. freezers, wine coolers, servers, battery chargers, heaters(!)... These devices do not require such a high power, so the circuit breakers can cope with this load. In the Fig. 1 can be seen the typical "cheating" scenario.

The problem, which was described above, was bother the users of public garages in Slovakia, housing estate "Vlčince" in Žilina (Fig. 2), where people paid exceptional amount of money for their garages "taking advantage" of shared power metering. The particular garages were equipped just with garage door openers - power consumption under load around 400W, in standby only 1W - and with neon-tube lamps - 20W - but the average daily power consumption was after re-counting on each user 3kWh$^{-1}$/year, what means 830Wh$^{-1}$/day per common user which opens and closes the garage doors in 30s two times a day.

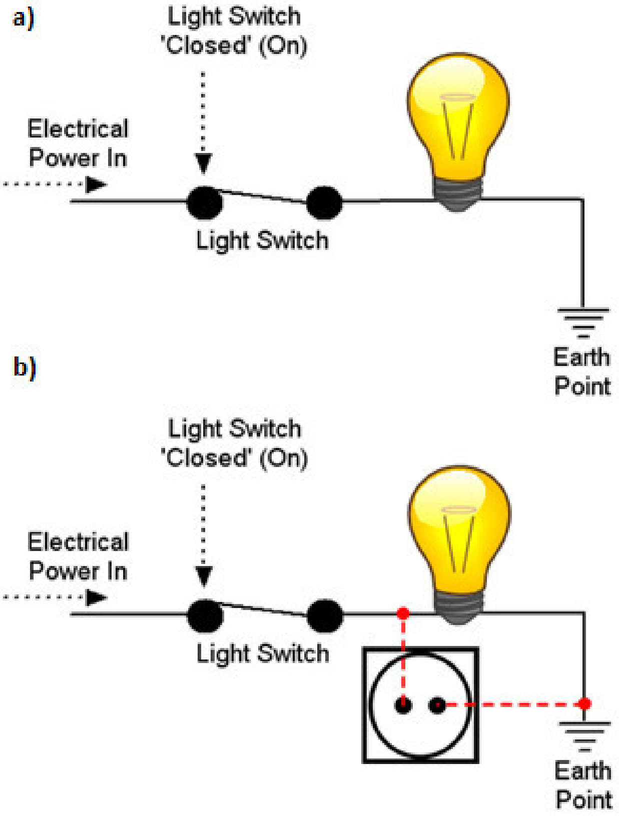


Fig. 1 a) normal lighting installation b) modification of normal lighting installation extended by power socket for "black" power consumption

Therefore the inspection of particular garages was performed to investigate the quality of electrical installations - not every user wanted to make the garage accessible.

Fig. 2 Investigated garage house

The reason was trivial, "black" power sockets were presented in some garages. The garage owners modified the installations and used the power sockets for illegitimate power usage. In the figures below can be seen the installation modifications.

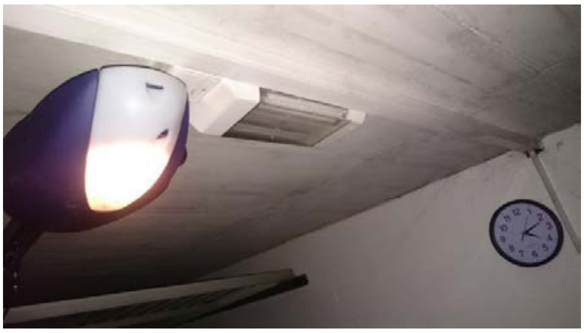


Fig. 3 Lighting installation as the source of shared power

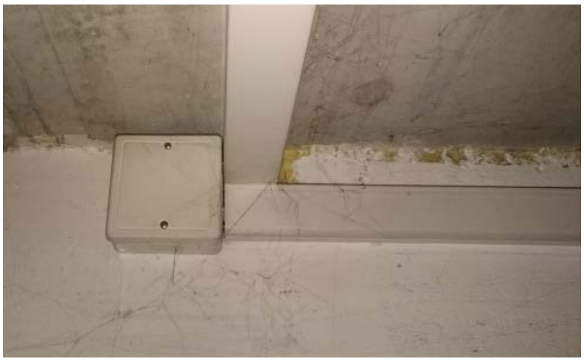


Fig. 4 Modification of lighting installation by screw terminal

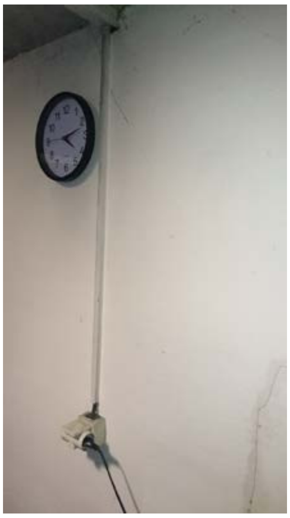


Fig. 5 Illegal socket as a “black” extension of lighting installation

To help people to overcome this problem, the special WSN-based power grid monitor was introduced. The power grid monitor is a system for energy consumption monitoring in the electricity grid. It consists of a measuring device in the form of the outlet adapter and the display device. The outlet adapter task is to measure the mains voltage and current that passes through it. The measured data are sent wirelessly to the display device. The outlet adapter consumes no more than 30mW during the measurement and it is able to measure with 15 kHz sampling rate (it means 300 samples per period (20 ms)). The outlet adapter is able to turn on / off the output which allow controlling remote devices, e.g. external lighting, or black power consumption characterized by increased power consumption. Moreover, switching can be automated on the time basis, power consumption or the outlet adapter temperature. So it can be used as a fuse.

As a display device with API, there is currently used a personal PC with its own wireless module in the form of a USB key. Wireless module communicates with the outlet adapter in unlicensed bands at frequencies around 868 MHz. Wireless range is 200m in ideal conditions, in real about half. The outlet adapter can be upgraded to have wired data connection, but currently does not have this opinion. USB acts in this case as an interconnection gateway providing connection to the other users (admin) with particular nodes employing the star architecture.

### *A. The outlet adapter*

The outlet adapter device is capable of measuring mains voltage and current drawn from the output of the outlet adapter. This output can be turned on / off via the communication interface. The device has a wireless communication interface for unlicensed bands around 868MHz. Wired interface is not implemented yet, but there is this option. Also "Power-line" communication can be added, it means to transfer data directly via the electricity grid.

Possible use of the outlet adapter:

- Electricity meter - The device can measure the power consumption and report it to the remote computer. The graphical application on the computer evaluates the power consumption of multiple outlet adapters.
- Electronic fuse - The device can monitor current limit and turn off output when the limit is exceeded. Turn off limit can be set in application. The device should be turn on again after a time period so no manual turn on is needed.
- Remote device switching - User can through a computer with a wireless module freely turn on / off the device plugged into the outlet adapter. This is an advantage especially for the inaccessible devices e.g. Christmas lights in an exterior or on the roof.
- Power inverter management - Modern household equipped with a solar panel and power inverter can use consumption information so that the inverter will only supply the power that the household consummates (and charge the batteries with the rest power). So the household wouldn´t become energy producer and could be resistant to short power dropout. During the power dropout the device will separate the household and the rest of the electricity grid. The inverter would then supply the household or a part of it.

Present version of the socket, which will be in our case installed at each garage doors power input is shown in a Fig. 5. To secure the socket, the socket cover is connected to the I/O pins of the socket, so the operator knows when the cover was removed. Therefore, if the cheater would like to somehow manipulate with the particular socket, it could be easily identified, since each socket will be accessible only from particular garage interior. A newer version of the device could be in a DIN rail housing thus could be put between circuit breakers.

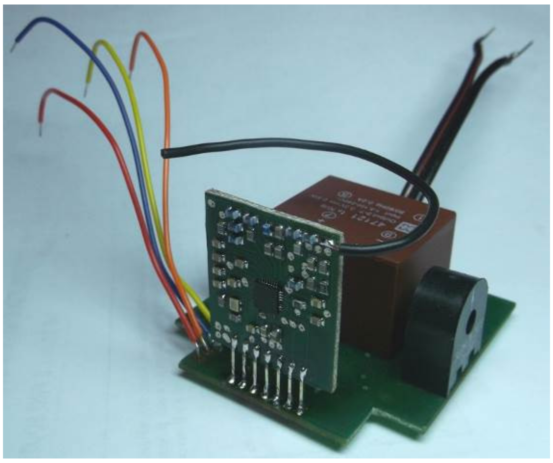


Fig. 6 The outlet adapter, system socket

*B. Wireless Module for PC*

Small module as a USB key contains a wireless transceiver with built-in antenna. The module transmits in the unlicensed bands, the same as the outlet adapter. Communication radius is approximately 200 meters in ideal conditions; in the building it is a bit less (depending on the particular location). The primary usage of the module is for receiving measured data from the individual outlet adapters. User application installed on a computer can interpret this data (e.g. consumption graph plotted for each outlet adapter for a certain period, displaying current / average consumption…). The USB module is shown in the Fig. 6.

The modules are able to communicate well with each other, allowing the interconnection of computers over a greater distance as Wi-Fi network. The transmitter can be improved in order to reach the communication radius near to 1km. The disadvantage is the low communication speed - only 12kbps. Other possible applications could be for warehouses or logistics centers where computers are far apart and need to transfer only a small amount of data.

## III. Description of the System HW Designs

Outlet adapter is using MSP430AFE252 microcontroller, which has a differential 24-bit sigma-delta converter with two input channels (there are also versions with more channels). Input voltage range for the converter is +-0,5V. One channel measures the electric grid voltage on the voltage divider adapted for its input range. The second channel measures the voltage at the current transformer "AX-1000" from the manufacturer "Talema". Mains conductor passes through the transformer to the device connected to the outlet adapter. The voltage at the transformer is thus proportional to the current flowing through the device (current includes also its own consumption). The maximum operating current which can microcontroller converter measure is 10A. Output from the power latching relay flows "JSL-D5N-K" manufacturer "Fujitsu". The relay is used to control the output of the two coils, which are activated by a voltage of 5V. The maximum power that can load the component part was set by the manufacturer to 2000W.

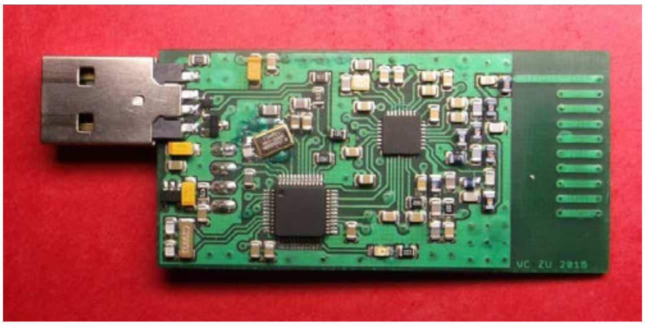


Fig. 7 USB wireless module for PC, interconnection gateway

Outlet adapter electronics is powered from the switching power supply from "Myrra" producer. The switching power supply changes the AC mains to stabilised operating voltage 3,3V with a maximum current consumption 750 mA. Producer guaranteed that average efficiency is 65% for this switching power supply. We did not observe any malfunctioning during its working. The outlet adapter includes a charge pump "TPS60401" manufacturer "Texas Instruments". The charge pump is connected as a voltage inverter and serves as a power supply for the bistable relay.

Changing the voltage using the charge pump is advantageous for this application. The main reason is its high efficiency and low circuit complexity. Low output current may seem like a disadvantage because of the higher current necessary to change the relay status. Usage of the bistable relay however means that the control current in the coils may not stay for the conservation of the relay. Its switch requires only short pulse. Since the relay is the only component that uses a negative potential, it does not matter if it is this tension fluctuate. To switch the relay is sufficient to have a negative voltage branch of a sufficiently large capacity of which short pulse will pass.

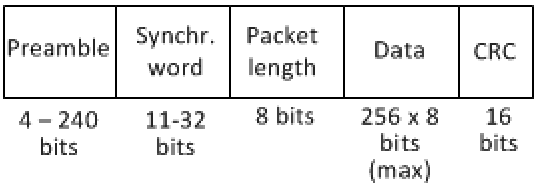
The wireless transmitting and receiving is provided through "CC1120" device from "Texas Instruments". The module combines transmitter with maximum output of 16dBm and receiver sensitivity of -123dBm at a 1,2kbps. Important features are summarized in the following points:

- Operating voltage is from 2V to 3.6V.

- 4-wire communication SPI bus and an interrupt signal.
- Special FIFO memory for transmitting and for receiving, both 128 bytes.
- Maximum transmission power to 50 Ohm antenna varies depending on the supply voltage. For 2V power supply is 12dBm and for 3.6V is 16dBm
- Power consumption when in max+imum power transmitting mode (+ 15dBm) is 50 mA. Power consumption for receiving mode is approximately 22,6 mA.
- The transmission frequency rate can be 164-192MHz, 274-320MHz, 410-480MHz or 820-960MHz
- Width of the transmission channel can range from 7812 Hz to 200 kHz.
- Circuit supports the following modulation: FSK-2, 2-GFSK, 4-FSK, 4-GFSK, ASK, OOK, DSSS and analog FM
- The maximum bit rate is half of the width of the transmission channel. For the two-state frequency modulation is the highest bit rate of 100ksps. When four state modulation it can be equal to the width of the channel - thus maximum 200kbps.

Wireless module offers the useful function "Enhanced Wake on Radio" (eWOR). This is a power-saving mode of receiving when the module turns on and off wireless module at regular intervals to detect the presence of the transmitter. This process reduces the average current consumption during receiving. It is possible to make the communication protocol in order to use this function by simply increasing the preamble size before the data packet so the receiver had more opportunities to capture the packet.

Wireless module CC1120 supports hardware processing of the packet that is shown in the figure below.



| Preamble | Synchr. word | Packet length | Data | CRC |
|---|---|---|---|---|
| 4 – 240 bits | 11-32 bits | 8 bits | 256 x 8 bits (max) | 16 bits |

Fig. 8 Hardware packaging data into packets

The preamble does not contain any valuable data, usually only alternation of ones and zeros. The task of the preamble is to give remote devices possibility to capture the presence of a transmitted signal to wake up from standby. The data transfer can be realized even without the preamble. Next important part is the synchronization word. Its purpose is to indicate the start of transmission of important data. If the receiver does not receive proper synchronization word transmission failed. Next part is the length of the packet. Length of the packet means only the size of user data that follow this part. This part may be omitted and sets the protocol for constant packet length. The last addition is the CRC code through which the correctness of the transferred data can be checked. Wireless module calculates its own CRC code from the received data and compares it with the received CRC code. If during transmission, there are some interference sources, which caused data defection, data are automatically declared non-conformity and will be erased.

The transmitted data encapsulates also two bytes for addressing devices and two bytes for the security key. The current implementation of the key can be understood as a password by which the system will or will not respond to the packet.

Smart metering units have their own standardized protocol that ensures compatibility among devices from different manufacturers. It is described in the standard "EN 13757", developed by "CENELEC". Wired communication protocol is commonly referred to as "Mbus" and wireless version as "WMbus". Standard "EN 13757" contains six documents. The latest ones are from 2013. [29], [30], [31], [32], [33]

The current weaknesses of the system are mainly related to the crisis cases. As an example, a power failure in the electric grid can be considered. Power failure causes loss of measured data which have not been sent to headquarters. The biggest problem is the loss of measured consumption. There are two ways to prevent it. The first is to use the backup battery, which provides data transmitting after power failure. This case assumes that the central has its own backup battery. The advantage of this method is the possibility of error detection in electrical distribution. If the mains voltage re-starting, headquarter can detect the status of the measuring devices. If the device reports the absence of voltage, it is situated in the faulty branch. The second way how to deal with data loss in case of power failure is using of non-volatile memory in which can the node hold the regularly stored data. In case of failure by using of such equipment, only a small part of data record will not be recorded.

Headquarter for the measuring device presents a personal computer. Ability to communicate in the ISM band at a frequency below 1 GHz gives the USB device. For proper operation of the device to a computer it is necessary to add a custom driver. Subsequently, the device appears as a virtual serial link. The easiest way is through the usage of direct communication from serial line.

The main scope of the WSN implementation was to distinguish the difference between authorized and unauthorized power consumption. An analysis of power characteristic waveforms will help with this problem, since it provides the information about the load type.

## IV. Power Factor Estimation

The definition of the power factor (PF) in AC circuits is well-known as the ratio of the real power ***P*** that is used to do work and the apparent power ***S*** that is supplied to the circuit:

$$PF = \frac{P}{|S|}\,, \qquad (1)$$

where ***PF*** is power factor, ***P*** is real power in watts [W] and |***S***| is apparent power - the magnitude of the complex power in volt·amps [VA].

The power factor can get values in the range from 0 to 1. When all the power is reactive power with no real power (usually inductive load) - the power factor is 0. When all the power is real power with no reactive power (resistive load) - the power factor is 1.

For **sinusoidal current**, the power factor ***PF*** is equal to the absolute value of the cosine of the apparent power phase angle φ:

$$PF = |\cos\varphi|, \qquad (2)$$

where **φ** is the apparent power phase angle.

In order to estimate power factor, we can compute real energy ***E<sub>PP</sub>*** consumed during one period by formula:

$$E_{pp} = \left(\sum_{i=0}^{n-1} U_i \,.\, I_i.T_s\right), \qquad (3)$$

where $\boldsymbol{U_i}$ in [V] and $\boldsymbol{I_i}$ in [A] are sampled values during one period of mains voltage, $\boldsymbol{T_s}$ is the sampling period in [s] and n is number of samples sampled during one period of mains voltage.

Now we can estimate apparent energy $\boldsymbol{E_{SP}}$ as:

$$E_{sp} = U_{RMS} \cdot I_{RMS} \cdot T_s\,, \qquad (4)$$

We can get $\boldsymbol{U_{RMS}}$ and $\boldsymbol{I_{RMS}}$ as the maximal sampled value divided by the square root of two:

$$U_{RMS} = \frac{|U_{MAX}|}{\sqrt{2}}\,, \qquad (5)$$

$$I_{RMS} = \frac{|I_{MAX}|}{\sqrt{2}}\,, \qquad (6)$$

where $\boldsymbol{U_{MAX}}$ and $\boldsymbol{I_{MAX}}$ are absolute maximums of sampled waveforms.

When we have all this values we can compute power factor ***PF*** by the formula:

$$PF = \frac{E_{pp}}{E_{sp}}\,, \qquad (7)$$

In normal case the light bulb has resistive character. Therefor the power factor is very close to 1. In case of inductive (or capacitive) load the power factor is lower. This indicates that the lessee uses forbidden type of appliance. Differences between reactive and resistant power waveforms can be seen in the figures below.

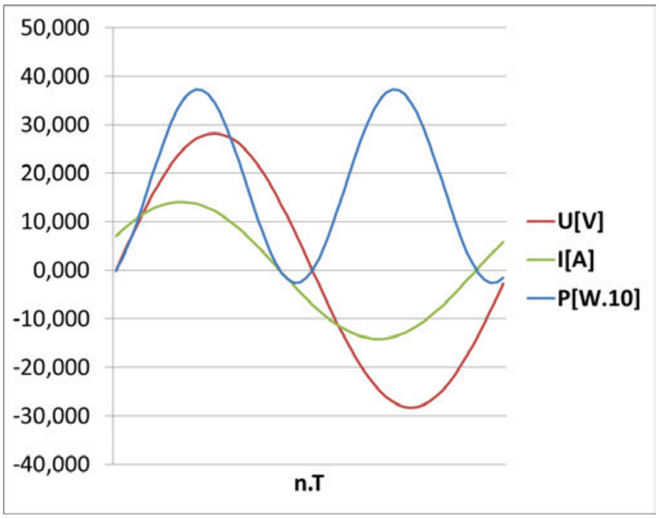


Fig. 9 Waveform of reactive power characteristics

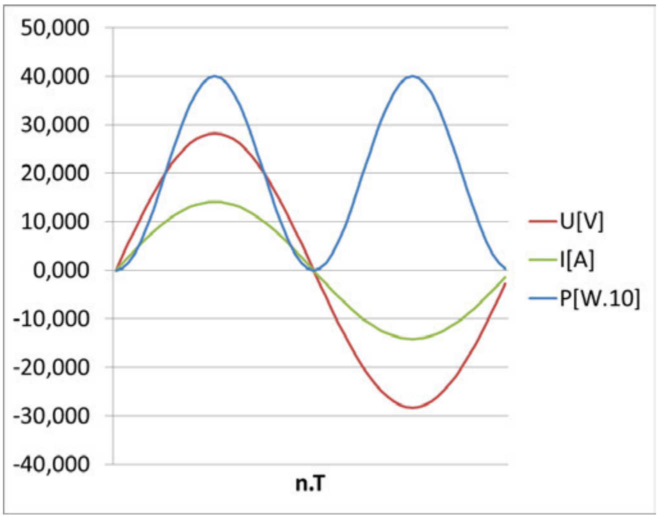


Fig. 10 Waveform of resistant power characteristics

## V. Conclusion

The paper described the first try to implement the application-specific indoor WSN for power metering at places with shared power sources. Practical measurement shows, that the network is ready to be finally applied. However, approval of all garage owners is necessary, what is at the moment an irresolvable problem. Other concrete application scenario has therefore to be found.

## Acknowledgment

The research is supported by the European Regional Development Fund and the Slovak state budget for the project ”Research Centre of University of Žilina”, ITMS 26220220183.

RESEARCH CENTRE UNIVERSITY OF ŽILINA

”We support research activities in Slovakia/project cofounded from the resources of the EU.”

## References

[1] M J. Miček, J. Kapitulík: WSN sensor node for protected area monitoring, *FedCSIS,* 2012: IEEE. - ISBN 978-83-60810-51-4.

[2] Chakravarthi, V. S.; Bhaskar, R. S.; Kusanur, V. - Conceptual Frame Work of Smart WSN for Bangalore Urban Environment Monitoring, *4th International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN),* 2012, Pages: 59-63, DOI: 10.1109/CICSyN.2012.21

[3] Peng Yu; Xu Yong; Peng Xi-yuan - GEMS: A WSN-based greenhouse environment monitoring system - *IEEE Instrumentation and Measurement Technology Conference (I2MTC),* 2011, Pages: 1 – 6, DOI: 10.1109/IMTC.2011.5944132.

[4] M. Hodon, P. Šarafín and P. Ševčík - Monitoring and Recognition of Bird Population in Protected Bird Territory, ISCC 2015, *The Twentieth IEEE Symposium on Computers and Communications,* Larnaca, Cyprus, 06.-09.July.

[5] J. Papán, M. Jurecka, and J. Púchyová - WSN for forest monitoring to prevent illegal logging, In: *FedCSIS: Proceedings of the Federated conference on computer science and information systems.* pp. 809 – 812. Wroclaw, 2012, Poland.

[6] Ning Jin; Renzhi Ma; Yunfeng Lv; Xizhong Lou; Qingjian Wei – A novel design of water environment monitoring system based on WSN, *International Conference on Computer Design and Applications (ICCDA),* 2010, Year: 2010, Volume: 2, Pages: V2-593 – V2-597, DOI: 10.1109/ICCDA.2010.5541305

[7] J. Miček, O. Karpiš, V. Olešnaníková and M. Kochláň - Monitoring of Water Level Based on Acoustic Emissions, ISCC 2015, *The Twentieth IEEE Symposium on Computers and Communications,* Larnaca, Cyprus, 06.-09.July 2015.

[8] O. Karpiš, J. Juríček and J. Micek - Application of wireless sensor networks for road monitoring, In *10th IFAC workshop on programmable devices and embedded systems,* 2015, Vol. 3. pp. 611–617.

[9] M. Hodoň, M. Chovanec and M. Hyben -Intelligent traffic-safety mirror, In: *Studia informatica universalis* - ISSN 1621-7545. - Vol. 11, no. 1 (2013), online, pp. 87-101.

[10] R. Žalman, J. Kapitulík and M. Kochláň - Distributed Sensor Network for Vehicles with Prior Right Detection, *The Twentieth IEEE Symposium on Computers and Communications,* Larnaca, Cyprus, 06.-09.July 2015

[11] R. Žalman, J. Milanová - Analysis and synthesis of acoustic signal in transport systems, *CSIT 2014,* Lviv, Ukraine. ISBN 978-617-607-669-8. - p. 166-167.

[12] J. Furtak, J. Chudkiewicy - The concept of authentication in WSNs using TPM, *Proceedings of the 2013 IEEE Federated Conference on Computer Science and Information Systems,* pp. 183 - 190, 978-1-4673-4471-5.

[13] J. Púchyová, M. Kochláň, and M. Hodoň - Development of Special Smartphone-Based Body Area Network: Energy Requirements, *Proceedings of the 2013 IEEE Federated Conference on Computer Science and Information Systems,* pp. 895–900, 978-1-4673-4471-5.

[14] M. Kochlan, and P. Sevcik - Supercapacitor power unit for an event-driven wireless sensor node, In: *Federated Conference on Computer Science and Information Systems (FedCSIS),* 2012 Publication Year: 2012 , Page(s): 791 – 796.

[15] O. Karpiš: *Solar-cell based powering of a node for traffic monitoring, IOSR journal of engineering,* 2013. - ISSN 2278-8719.

[16] T. Bernard, and H. Fouchal - Slot Scheduling for Wireless Sensor Networks, In: *IOS Press, Journal of Computational Methods in Science and Engineering,* doi:10.3233/JCM-2012-0432

[17] V. Olešnaníková and J. Púchyová - Analysis of voice activity detection for implementation into WSN. *CSIT 2014,* Lviv, Ukraine. - ISBN 978-617-607-669-8. - p. 75-76.

[18] M. Hodoň et al. - Maximizing Performance of Low-Power WSN Node on the Basis of Event-Driven-Programming Approach, ISCC 2015, *The Twentieth IEEE Symposium on Computers and Communications,* Larnaca, Cyprus, 06.-09.July 2015.

[19] Y. Zhai, X.Cheng - Design of Smart Home Remote Monitoring System Based on Embedded System, *2nd International Conference on Computing, Control and Industrial Engineering (CCIE),* 2011 IEEE(Volume:2 ), 20-21 Aug. 2011, Page(s): 41 - 44, ISBN: 978-1-4244-9599.

[20] Vo, Minh-Thanh, Tran, Van-Su ; Nguyen, Tuan-Duc ; Huynh, Huu-Tue - Wireless Sensor Network for Multi-Storey Building: Design And Implementation, Published in: *International Conference on Computing, Management and Telecommunications (ComManTel),* 2013, 21-24 Jan. 2013, Page(s): 175 - 180, ISBN: 978-1-4673-2087-0

[21] Malatras, A.; Asgari, A.; Bauge, T. - Web Enabled Wireless Sensor Networks for Facilities Management, *IEEE Systems Journal,* Volume:2, Issue: 4, Page(s): 500 - 512, ISSN: 1932-8184

[22] Skeledzija, N., Cesic, J. ; Koco, E. ; Bachler, V. ; Vucemilo, H. N.; Dzapo, H. - Smart Home Automation System for Energy Efficient Housing, *37th International Convention on Information and Communication Technology,* Electronics and Microelectronics (MIPRO), 2014, 26-30 May 2014, Page(s): 166 - 171, Print ISBN: 978-953-233-081-6

[23] Youngjin Choi et al. - Monitoring System Employing Facebook Platform for WSN, *IEEE 15th International Conference on Advanced Communication Technology (ICACT),* 2013, 27-30 Jan. 2013, Page(s): 1037 - 1041, ISSN: 1738-9445, ISBN: 978-1-4673-3148-7

[24] A. Adel Ali, M. Al-Shaboti, and A. Al-Zubairi - An Indoor Emergency Guidance Algorithm Based on Wireless Sensor Networks, *International Conference on Cloud Computing (ICCC),* 2015, 26-29 April 2015, page(s): 1 - 5, ISBN: 978-1-4673-6617-5

[25] Yerra, R. V. P.; Bharathi, A. K. ; Rajalakshmi, P. ; Desai, U. B. - WSN Based Power Monitoring in Smart Grids, *7th International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 6-9 Dec. 2011, p: 401 - 406, Print ISBN: 978-1-4577-0675-2

[26] Peng Han ; Jinkuan Wang ; Yinghua Han ; Qiang Zhao - Novel WSN-Based Residential Energy Management Scheme in Smart Grid, *IEEE International Conference on Information Science and Technology (ICIST),* Page(s): 393 - 396, 23-25 March 2012, ISBN: 978-1-4577-0343-0

[27] Sahin, D.; Bulbul, S.; Gungor, V.C.; Kocak, T. - Reliable Routing in Wireless Sensor Networks for Smart Grid Environments, Published in: *20th Signal Processing and Communications Applications Conference 2012,* 18.04.-20.04 2012, p(s): 1 - 4, E-ISBN: 978-1-4673-0054-4, Print ISBN: 978-1-4673-0055-1

[28] A. Marchiori - Enabling Distributed Building Control with Wireless Sensor Networks, *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks,* Page(s): 1 - 3, 20-24 June 2011, ISBN: 978-1-4577-0352-2

[29] EN 13757-4:2013: *Communication systems for meters and remote reading of meters - Part 4: Wireless meter readout* (Radio meter reading for operation in SRD bands)

[30] UM1759 User manual Wireless M-Bus firmware and application, June 2014 DocID026279 Rev 1, available on-line at http://www.st.com/st-web-ui/static/active/jp/resource/technical/document/user_manual/DM00115100.pdf

[31] http://www.ti.com/tool/WMBUS

[32] EN 13757-2:2004: *Communication systems for and remote reading of meters - Part 2: Physical and link layer*

[33] EN 13757-3:2004: *Communication systems for and remote reading of meters - Part 3: Dedicated application layer*

DOI: 10.15439/2015F67
ACSIS, Vol. 5

# Energy Balancing Algorithms in Wireless Sensor Networks

Anne-Lena Kampen
Bergen University College, Bergen
ITEM NTNU, Trondheim
Norway
Anne-Lena.Kampen@hib.no
anneleka@stud.ntnu.no

Knut Øvsthus
Bergen University College, Bergen
Norway
knut.ovsthus@hib.no

Øivind Kure
NTNU, Trondheim
Norway
okure@item.ntnu

***Abstract*—The energy consumption in Wireless Sensor Networks, WSN, need to be balanced in order to avoid early depletion of nodes. In this paper we use a common context to analyze a broad range of the energy balancing algorithms suggested in literature. In addition we suggest three new algorithms to complete the range. Altogether, nine different balancing techniques are analyzed. We focuses on networks running the IPv6 Routing Protocol for Low-Power and Lossy Networks (RPL) routing protocol. Our simple change in RPL's parent selection procedure can give a significant balancing effect without any increase in management cost. However, the best balancing algorithm is when the nodes exchange residual-energy information to ensure forwarding through the highest residual-energy next-hop node. The increased information exchange implies increased management cost due to the amount of information transmitted and added computational load.**

## I. Introduction

WIRELESS sensor networks (WSN) generally consist of wireless nodes with a collective objective of gathering measured information at the sink [1]. The monitored area may be large compared to the nodes' transmission range. Hence, the information needs to be relayed to reach the sink. The topology of the relaying paths may create imbalance in the traffic share, and therefore the energy consumption, between the nodes. Energy imbalance results in lifespan variation between the nodes. Observations of real networks in [2] and [3] show that some nodes relay a substantial portion of the traffic, thus they become hot-spot-nodes having a high energy consumption rate.

Depleted or dead nodes make the gathered data incomplete and, more important they may cause network partitioning. Applying energy balancing routing algorithms levels the traffic load, hence lifetime, between the nodes. The ideal situation is long living WSNs where all nodes have equal lifetime. However, this ideal situation not feasible due to the increased traffic density toward the sink in multihop networks. The goal in multihop networks is instead to balance the energy consumption between nodes at equal hop distance from the sink. Network management will be simplified if the nodes at each rank have similar lifetime. Balancing algorithms are the topic of this paper.

Our contribution is threefold. First we present a methodical review of a broad range of energy balancing algorithms. The algorithms range from approaches requiring simple changes of the applied routing algorithms, to approaches that require complex add-ons. Second we use a common context to compare these algorithms. Third, we suggest three new balancing algorithms to complete the collection of balancing algorithms found in the literature. The tree suggested algorithms are random selection of preferred-parent, conserving of Single Point of Failure (SPOF) parent and energy balancing based on eavesdropping.

To get a good estimate of the energy pattern in the network we use the nodes residual-energy. The residual-energy gives the true picture of the energy variation that appears between the nodes. In addition, residual-energy is directly related to the nodes lifetime.

To evaluate the impact of different energy balancing techniques, we use the routing protocol suggested by Internet Engineering Task Force (IETF) for use in WSN, IPv6 Routing Protocol for Low-Power and Lossy Networks (RPL) [4]. RPL creates routing entries in the nodes which forms an overall destination oriented directed acyclic graph (DODAG) rooted at the sink. The graph is created by broadcasting of DODAG Information Object (DIO) messages. The sink initiates the transmission, and the messages are further broadcasted throughout the whole network. The DIO includes the senders' rank information. The rank indicates a node's distance to the sink. The rank increases as the distance to the sink increases. The sink is at rank 0 and the sink's one-hop neighbor defines the rank-one nodes and so forth. Each node caches a parent-list containing all neighbors that report a rank equal to the lowest rank heard. A preferred-parent is selected among the nodes in the parent-list. The preferred-parent is used as the current next-hop node on the path toward the sink. To maintain the DODAG, the nodes transmit DIO messages periodically at intervals decided by a trickle timer [5].

The rest of the paper is structured as follows. In Section 2 we present related work, in Section 3 introduces the different energy balancing approaches to be analyzed, the

978-83-60810-66-8/$25.00©2015, IEEE

simulation is presented in Section 4, and Section 5 comprises the conclusion.

## II. Related work

Several energy balancing approaches are suggested in the literature. Energy balancing based on selecting the most energy optimal path is suggested in [6] - [13]. In all but the two latter of these algorithms are energy information exchanged through DIO messages. Applying DIO messages to exchange energy information means the energy balancing depends on the number of data packet transmitted per transmitted DIO message. Thus, increased energy balance is paid by increased DIO transmission frequency, which means increased average energy consumption. Further, the trickle timer [5] decides the DIO emission frequency such that the emission frequency decreases exponentially with time in converged networks. Hence, the balancing effect of the algorithms discussed in the next paragraph will decline with time.

The object function (OF) for RPL suggested in [6] defines the path cost as the energy level of the node on the path with lowest residual-energy. The node that advertises the highest path cost is preferred as selected parent, and the lower energy nodes are spared. The authors of [8] suggest that the node with the highest remaining energy among the nodes with the lowest expected transmission count (ETX) is chosen as the preferred-parent in network running RPL. Both ETX and node energy is used to select between parent nodes of equal hop-count in [9]. The residual-energy is included as a denominator in the additive distance metric in [10]. Using it as in the denominator makes the cost of a node increase toward infinity as energy approaches zero. Hence, the paths including low energy nodes is avoided due to their high cost. A routing metric that calculate the expected lifetime of the nodes is defined in [11]. The expected lifetime is calculated as the ratio of the node's residual-energy over the total energy spent to transmit data. The paths including the most constrained nodes are avoided by defining the path weight as the minimum expected lifetime along the path. An approach similar to RPL is used to create paths for networks with multiple sinks in [7]. Several equal-rank nodes are cached as potential parent based primary on hop-count metric, secondary the nodes energy metric and third on the highest link-quality-indication. Algorithms where the highest residual-energy path is selected or the lowest residual-energy paths are avoided are part of our analysis.

Energy information is exchange through the ACK packet in the approach presented in [13]. The network run RPL, and the nodes perform a weighted selection to choose among its available next-hop nodes. The selection is weighted between distributing the traffic through the lowest delay path and distributing the traffic to nodes with higher remaining energy. In the routing protocol suggested in [12], the energy information is both piggybacked on data packets and included in the ACK packets. Hop-count is used as the metric to generate parent-list. Data packets are transmitted to the highest energy member of the parent-list. If there is no parent node available, the packet is transmitted to the sibling node with the highest amount of energy. Our analysis includes an algorithm where the ACK message exchange the energy information.

Energy consumption can be balanced by continuously spreading the transmitted data over multiple paths, and such methods are also part of our analysis. Approaches using multiple paths are suggested in [14] [15]. RPL is used as the routing protocol in [14] where the forwarding load is weighted between the members of the parent-list. The weighting is based on the members' residual-energy. The transmission range dynamically adjusted to maintain k parents. Energy information is exchanged between the nodes through ACK and DIO packet. In addition is also hello packets mentioned as possible information carriers. The approach in [15] enables multipath data forwarding through energy-sufficient paths, as opposed to minimum-energy-cost paths. They propose a routing algorithm which makes a hierarchical routing graph similar to RPL. The nodes forward packets through alternate paths to extend the network lifetime. The conditions of the paths are monitored by the sink which re-initiate path search if the number of working paths gets lower than two. Multiple paths are also discussed in the surveys presented in [16] [17]. Survey [16] cites an algorithm presented in [18], which takes both the energy level and hop distance into account to allocate different data rates to multiple disjoint paths. The sink decides the rate of the different paths and assign messages are sent form the sink to the source nodes to inform about the path rates. The top-down survey paper [17] cites an interesting improved cost function used to balance the energy consumption among the nodes [19]. The improved cost calculation algorithm makes the cost increase rapidly with decrease in the nodes remaining energy. Hence, traffic is directed away from hot-spot nodes. The approach requires that the nodes cache several states for each neighbor and that energy information is exchanged periodically. The survey paper [17] also discusses energy balancing by using a few relay nodes with enhanced capabilities. In addition they discuss use of mobile sinks. These algorithms increase the start-up management cost of the networks, and increases the network cost.

Clustering is among the energy efficient algorithm discussed the survey presented in [20] and suggested to improve energy utilization in [26]. The basic idea of energy efficient clustering is to perform energy efficient rotation of the clusterhead assignment and let the clusterhead perform energy efficient management of the local cluster traffic. Clustering is not part of our analysis as it is not very well fitted for RPL running network.

Balancing the energy consumption by making the nodes alternate between direct transmissions to the sink and using multi-hop transmissions is suggested in [21].

The protocol is used as an extension to RPL in [22] which presents a smart/green test-bed of nodes spanning across several smart offices. The findings of [22] shows that the protocol suggested in [21] balances the network energy consumption compared to classic RPL. However, it is more energy expensive giving an overall increased energy dissipation. This algorithm is only considered for one-hop networks, while we are considering multihop networks.

## III. Balancing network energy consumption

In this section we present the different energy balancing algorithms that are analyzed. Our hypothesis is that introducing small changes in the parent selection procedure improves the WSN energy balance, while substantial enhancements come at a cost of increased management complexity and information exchange between nodes. Further, efficient energy balance is achieved when focusing on reducing the load of the hot-spot nodes.

The following text lists nine algorithms. The tree new algorithms that we suggest are A: Randomize parent selection, D: Weighting round-robin based on SPOF-parent energy level and H: weighting round-robin based on eavesdropping.

### *A. Randomize parent selection*

As a first approach to enhance the energy balance in WSN, we suggest a simple change in the preferred-parent selection algorithm. The aim of the suggested algorithm is to reduce the probability of creating hot-spot nodes. The probability is reduced by preventing that several child nodes select the same preferred-parent if other potential parents exist. According to the RPL algorithm, all nodes cache a parent-list containing all candidate parent nodes. A preferred-parent are selected among the parent-list nodes, using a specific parameter as tiebreaker. Hence, nodes with globally good tiebreaker value will be selected by all potential child nodes and may therefore become hot-spot nodes.

Our suggested algorithm creates a small change in the preferred-parent selection procedure to reduce the probability of creating such hot-spot nodes. The nodes randomly select a preferred-parent among the nodes in the parent-list instead of using a preordain tiebreaker parameter value. Hence, the probability that several potential child nodes select the same node as preferred-parent is reduced. The forwarding load is therefore more balanced. The weakness of the algorithm is that it can give energy consumption imbalance if selected parents are located such that they represent single paths for other nodes.

### *B. Round-robin through multiple paths*

Selecting a single preferred-parent may overload some potential parents while leaving some potential parents unused. Thus, our analysis comprises an approach where this imbalance is alleviated by making the nodes transmit data packets to all nodes in their parent-list in a round-robin fashion. The approach shares the forwarding load equally between all members of the nodes' parent-lists. The main weakness of the approach is the load imbalance that is created between nodes with different number of child nodes.

### *C. Weighted round-robin based on energy information in DIO messages*

To level the energy imbalance that may appear using the round-robin approach we implement algorithms in which the nodes exchange energy information during DIO transmission. The information is used to perform a weighted-fair-sharing between the parent nodes. Thus, the nodes share the traffic load among the nodes in the parent-list according to their relative residual-energy level. The energy-balancing effect of the weighted algorithms depends on the freshness of the energy information cached for the nodes in the parent-list. Hence, increased DIO exchange frequency means improved energy balance. However, increased DIO exchange frequency increases the energy consumption in the network. Thus, there is a tradeoff between energy balance and average energy consumption.

Weighted round-robin ensures that the energy depletion rate of the low energy parents is reduced, hence the energy balance is improved. However, the algorithm preserves the existing energy imbalance relationship between parent nodes.

### *D. Weighting round-robin based on SPOF-parent energy level*

The goal of our single point of failure (SPOF) algorithm is to prevent early depletion of SPOF nodes. We define SPOF nodes as nodes that are part of one or more parent-lists containing only one member. In other words, a child that has a SPOF parent is disconnected from the routing graph if the parent node dies. Child of SPOF parent forwards all data through the SPOF parent. Even when the SPOF parent has a very low energy level, the child has no other option than continue forwarding through the SPOF parent. Hence, depleting of the SPOF node is continued.

To reduce the depletion rate of the SPOF nodes we suggest to direct traffic originating from higher rank nodes away from the SPOF nodes. In order to do so, we let nodes with a SPOF parent advertise the energy level that is the lowest of its own and its SPOF-parent's residual-energy level. Thus, traffic is directed away from the paths including the SPOF node.

Directing the traffic away from the SPOF nodes may come at an expense of other low energy nodes on the same rank as the SPOF node. However, child with SPOF parent continue to transmit their own generated data to their SPOF parent, while other nodes only get a weighted amount of traffic from their respective child nodes.

The DIO is used to exchange energy information.

### *E. Weighted round-robin based on prediction parents energy consumption*

The energy information gained through received DIO can be used to predict the energy consumption pattern in between DIO updates. To test such energy prediction algorithms, we implement an algorithm that estimate parents current energy level based on statistics of former energy consumption. The algorithm is as follows. The residual-energy a node advertises in consecutive transmitted DIOs is cached at the receiving nodes. The timespan between the consecutive DIO is further used to estimate the depletion rate of the transmitting node. The current energy level is estimated using the individual parents' energy drain rate and last advertised energy level. The estimated energy level is used to perform weighted-fair-sharing between the parent nodes.

### *F. Weighted round-robin while avoid lowest energy parent*

In order to focus on the hot-spot nodes we suggest a partly weighted algorithm. The data is weighted between the parent nodes. However, no data is transmitted to the parent with the lowest residual-energy. Hence, the load on the hot-spot nodes is reduced. This algorithm requires that the nodes exchange residual-energy information through the DIO message.

### *G. Use the highest energy parent node*

In the multiple path approaches, although weighted, each parent receives data for forwarding from their child nodes. This applies even if the residual-energy level of the parent is low. Hence, if all nodes transmit approximately equal amount of traffic, the nodes that are members of several paths are depleted faster than other nodes.

The depletion pace of low energy nodes is reduced in approaches where the lowest energy parent is avoided such as the approach presented in subsection III.F. However, nodes forces parents with second lowest energy level to forward traffic. Hence, the lowest energy nodes alternates their states with next lowest energy nodes.

A simple solution is to use only the highest energy parent node as the next-hop node. This algorithm is similar to the algorithm used in [8]. We implemented this algorithm and used DIO to exchange energy information.

### *H. Weighting round-robin based on eavesdropping*

Utilizing information conveyed in DIOs may give an incomplete view of the current energy levels of the nodes in the parent-list. The reason is that traffic imbalance, and associated energy consumption imbalance that occurs between DIO transmissions are not taken into account.

In order to predict parents' energy consumption between DIO updates, we suggest that nodes eavesdrop on the traffic transmitted in the area. The algorithm operates as follows. Nodes read the source and destination address information in the eavesdropped traffic. The address is matched against the content in the parent-list of the eavesdropping nodes. When a match is found, the energy level of the associated parent-list entry is reduced according to the eavesdropped information. The energy level of the nodes in the parent-list is then used to perform weighted-fair-sharing.

Eavesdropping does not significantly influence on the nodes energy consumption. The reason is that overheard packets destination address has to be read anyway to determine the intended receiver of the packet. The energy consumption due to overhearing is not taken into account when comparing the different balancing techniques. The reason is that the extent of overhearing energy consumption is mainly decided by the energy saving approach chosen at the MAC layer, while we are concentrating on the routing layer algorithms' impact on energy consumption.

The eavesdropped traffic may not give a complete overview of the parent nodes traffic load. For instance, child nodes of a common parent may not receive each other's packets due to hidden node. Thus, the calculations of the energy consumption of the parent-list nodes may be imprecise.

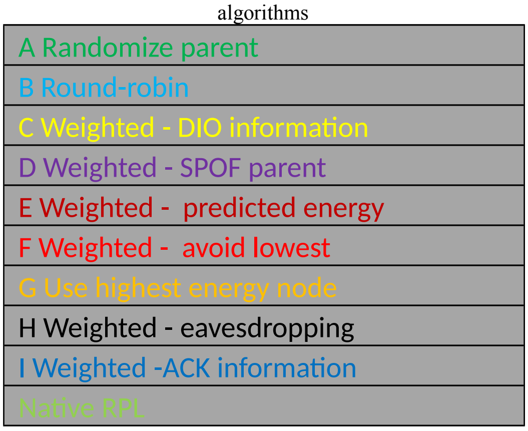
### *I. Weighting round-robin based on energy information conveyed in ACK packets*

Lastly, we implement an algorithm in which information about the nodes' energy variation in between DIO transmissions is exchanged through ACK packets. ACK packets are sent as a response of received data packet. Hence, the nodes achieve a complete overview of the diverse energy levels of the nodes in their parent-list as each parent relays a packet.

However, the energy information of parents with low residual-energy is less current. The reason is that low energy nodes seldom forward data as they have low weight. In addition, nodes that rarely transmit data can have stale energy information for the nodes in their parent-lists. This may give temporary screwed forwarding load among parent nodes. However, the energy levels are continuously

TABLE 1.

Color-codes used in the figures to define the different energy balancing algorithms



| |
|---|
| A Randomize parent |
| B Round-robin |
| C Weighted - DIO information |
| D Weighted - SPOF parent |
| E Weighted - predicted energy |
| F Weighted - avoid lowest |
| G Use highest energy node |
| H Weighted - eavesdropping |
| I Weighted -ACK information |
| Native RPL |

balanced as energy information is updated, smoothing the discrepancy over time.

Weighted-fair-sharing is performed based on the energy information.

## IV. Simulations

We perform simulations to evaluate the different discussed algorithms. The energy consumption in WSN increases toward the sink as the inner nodes are obligated to relay traffic for outer nodes. Thus, we mainly present simulations segregated on the node's rank. Applying an energy balancing algorithm will not change the average energy consumption for the nodes at the different rank since the total number of packet transmitted through each rank is unchanged.

We concentrate on transmitting and receiving energy consumption. Overhearing energy is not taken into account. The reason for omitting the overhearing energy consumption is that we concentrate on energy balancing at the network layer, and overhearing energy consumption is strongly dependent on the energy saving approach applied at the MAC layer. Overhearing may give a small variation in the average energy due to the chosen path. However, the difference between the energy consumed due to overhearing become negligible because the algorithms are compared at given average node density.

As discussed above, the average energy consumption at each rank is consistent regardless of applied balancing algorithm. However, an efficient energy balancing algorithm makes the residual-energy of the most depleted node approach the average residual-energy at the given rank. Hence, we present the average and minimum residual-energy after each node has generated 100 data packets.

Energy information is used to tune the traffic load between parent nodes in some of the evaluated algorithms. In networks running these algorithms, each node caches residual-energy information of its parents. The accuracy of the cached information depends on the update interval. For the algorithms that exchange energy information through DIO messages, the accuracy is improved by reducing the number data packets exchanged per DIO transmitted. The residual-energy information accuracy approaches the accuracy of the ACK algorithms if the DIO emission frequency approaches the data rate. However, increased DIO exchange frequency increases average energy consumption. Further, the DIO exchange frequency is decided by the trickle timer such that the exchange frequency is strongly reduced in converged stable networks. Thus, the energy balance is declining over time when the balance depends on DIO exchanged information.

In our initial simulations, 100 data packets create sufficient network traffic to discriminate the balancing effect of the different category of balancing algorithms. However, each node transmit a total of 16 DIO messages during the simulation runs, hence the number of data packets transmitted for each DIO transmission is low. Thus, to improve the basis of comparison we present additional simulation results for the weighting algorithms. Only two DIO messages are transmitted during the simulation run in these additional simulations, and the number of transmitted data packets is increased to 300. This gives a more fair comparison between the algorithms relying on DIO to exchange energy information, relative to the algorithms that use additional means to exchange energy information.

We evaluate the different energy balancing approaches by performing simulations in OMNET++ [23], using the MiXiM module for wireless communication. The nodes' energy consumption is calculated based on traffic load. Based on the observations and references in [24], we assume that receiving and transmission of data packets consume the same amount of energy. The different types of packets have different packet sizes. Management packets are assumed to be half the size of data packets, while ACK packets are one tenth of the size of the data packets. These relative values are chosen based on an assumption that data packets never need the maximum allowed packet sizes as they are mainly limited to carry only measured data, while management packets only carries strictly needed information. Maximum data frame sizes and ACK frame sizes information extracted from 802.15.4 datasheet [25].

The nodes are randomly distributed in an 800m times 800m area. The nodes transmission range is 141m. The number of nodes is varied such that the node density changes from 8 to 20. The node density is defined as the number on nodes inside a circle with radius equal to the nodes transmission range. Every simulation point presented represents the average value of 30 simulation runs with different seeds for random deployment of nodes.

As discussed above, the average energy consumption is equal for each rank over all energy balancing algorithms. However, the residual- energy values of the most depleted nodes change with the applied balancing algorithm. The most optimal energy balancing algorithm is the algorithm in which the residual-energy of the most depleted nodes converges to the average value. Therefore, we compare the algorithms with respect to their ability to make the nodes with lowest residual-energy approach the average value of their associated rank. Figure 1 shows simulation results that demonstrate to what extent the different algorithms make the lowest and average values converge. All the algorithms discussed in Section 3 are presented in the figure.

The nodes' average residual-energy, as well as the residual-energy of the most depleted nodes are presented in Figure 1. This is the residual-energy level of the nodes after each node has generated and transmitted 100 data packets. In addition to data, management traffic has been exchanged to make the network converge. Further has periodic DIO updates been transmitted. The circled shaped

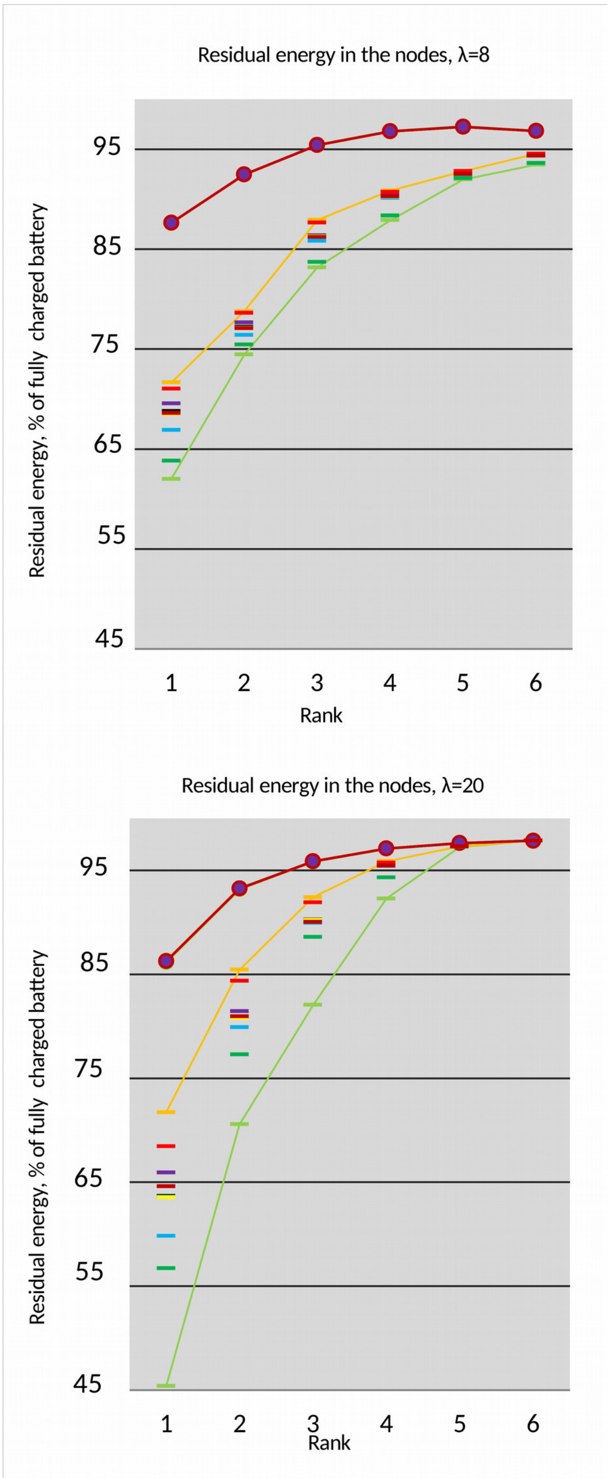


Fig 1. Residual-energy in the nodes after each node has generated 100 data packets.

markers with the corresponding lines show the average residual-energy values. The energy levels for the most depleted nodes are shown as short horizontal markers. In order to clarify the information displayed in the figure, the best and the worst of the residual-energy levels are displayed with solid curves through their associated markers. Hence, the orange curve cut through the markers representing the highest residual-energy level and light green curve cut through the markers representing the lowest residual-energy. The colors of the markers and lines indicate the corresponding energy-balancing algorithm as defined in Table 1. To prevent that the important information gets hidden in an overloaded display, the 95% confidence interval is not shown in the figures. However, the 95% confidence interval is always within 7% of the average values. Simulations performed for node densities of λ = 10 and λ = 15 show the same trends as shown for λ = 8 and λ=20 in Figure 1.

The simulation results displayed in Figure 1 shows that native RPL creates energy imbalanced networks. The native RPL simulation results are represented by the light green marks and the light green curve. Native RPL gives the overall lowest residual-energy for all ranks and all node densities. The reason is that a fixed parent is used throughout the whole simulation scenario, further is lowest node-id used as a tiebreaker when choosing between potential parent nodes. The latter means that several nodes choose the same preferred-parent node.

The difference between the lowest residual-energy node and the average value increases toward the sink. The reason is the increased traffic density. A given imbalance in traffic share causes an increase in the real traffic load difference as the total traffic increases.

The energy imbalance increases rapidly with node density for approaches where the parent node is fixed. This is observed in Figure 1 where the light green native RPL line rapidly moves away from the average line as the node density increases. The reason is the increased number of neighbors. Increased number of neighbors increases the number of child nodes for the fixed parent.

Based on the discussion above it is clear that some kind of energy balancing techniques should be added to networks running native RPL. Our suggested random preferred-parent selection algorithm presented in subsection III.A, demands a minor change in the RPL implementation. Nevertheless, the residual-energy of the most depleted node is reduce by over 10% compared to native RPL for high density networks. This is seen in the Figure 1 comparing the light green native RPL marks with the dark green marks.

However, the most efficient energy balancing algorithm is the one presented in subsection III.G, in which the parent with the highest residual-energy is selected as the next-hop node. The algorithm is represented by the orange marks in the Figure 1. Using highest energy parent increases the residual-energy of the most depleted node more than 25% compared to the native RPL. The merit of the algorithm is that the residual-energy nodes are avoided. This result corresponds to the results presented in [8] and [6] where ETX is used as a metric to populate the parent-list. However, using the parent with the highest residual-energy means that energy information has to be exchanged between the nodes. Randomizing parent selection, presented in subsection III.A, does not add any overhead.

The weighted share algorithms increase the residual-energy of the most depleted nodes with 15 to 20% compared to native RPL in high density networks. The weighted share algorithms are presented in subsection III.C-F and III.H-I. The improved performance complies with the studies performed in in [14] [15]. However, the poorer performance of these algorithms compared to the highest energy parent algorithm, III.G, is due to the fact that the lowest energy parent is still used, although rarely.

In subsection III.F we suggest the improved weighting algorithm, where the most depleted node is avoided while the traffic is weighted between the other parents. The improved algorithm is represented by the red marks in Figure 1, and shows that the residual-energy of the most depleted node is always distinguishable higher than the general weighting algorithms.

The round-robin approach, represented by the light blue marks, contributes less to balance the energy than the weighting algorithms. The reason is that parent with low residual-energy are loaded with the same amount of traffic as the other parents.

The simulated weighted-fair-sharing algorithms exchange energy information through DIO messages as well as through the algorithm-specific energy information exchange technique. Thus, the energy information update intervals between the different algorithms converge if the DIO exchange frequency is high compared to the packet exchange frequency. This phenomenon is demonstrated in Figure 1 as it is difficult to discern between the simulation results for the algorithms that use weighting as balancing technique. To better illustrate the difference between the weighting algorithms we performed simulations where each node generated 300 data packets. The DIO exchange is limited such that each node only generates two DIOs during the whole simulation. The simulation result is shown in Figure 2.

As expected, when the DIO exchange frequency is reduced, the ACK method has an improved balancing performance relative to weighting based on DIO information. The ACK method is presented in subsection III.I and weighting based on DIO information is presented in subsection III.C. The improved balancing performance is seen in Figure 2 where the blue marker of the ACK method is closer to the average values than the yellow DIO information markers. In Figure 1, the blue ACK markers are actually hidden by the yellow markers. Hence, the ACK and the DIO information performed equally well when the number of data packets per DIO packet is low.

The eavesdrop-algorithm described in subsection III.H and represented with the black marks and line in the Figure 2, has the worst balancing capabilities. This applies especially for the one-hop nodes. The reason is increased traffic density in these areas of the network. High traffic density means that nodes often become hidden terminals preventing them from eavesdropping neighbors' traffic.

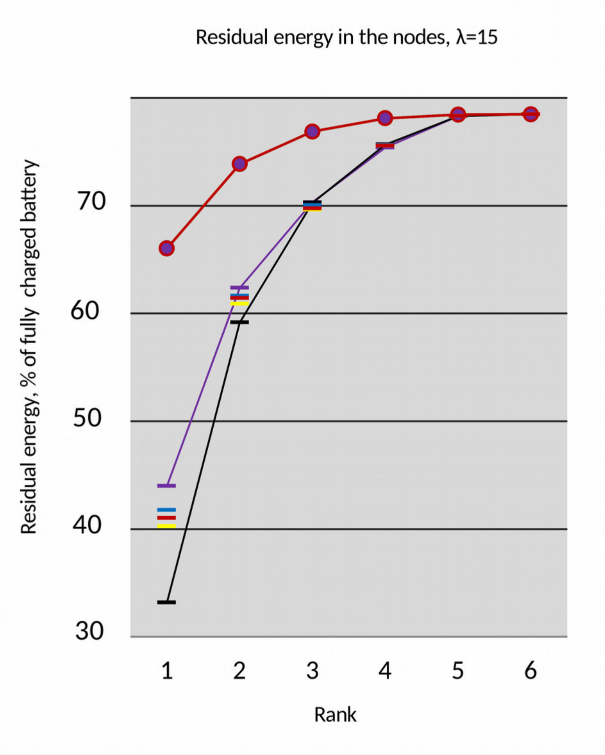


Fig 2. Residual-energy in the nodes after each node has transmitted 300 data packets

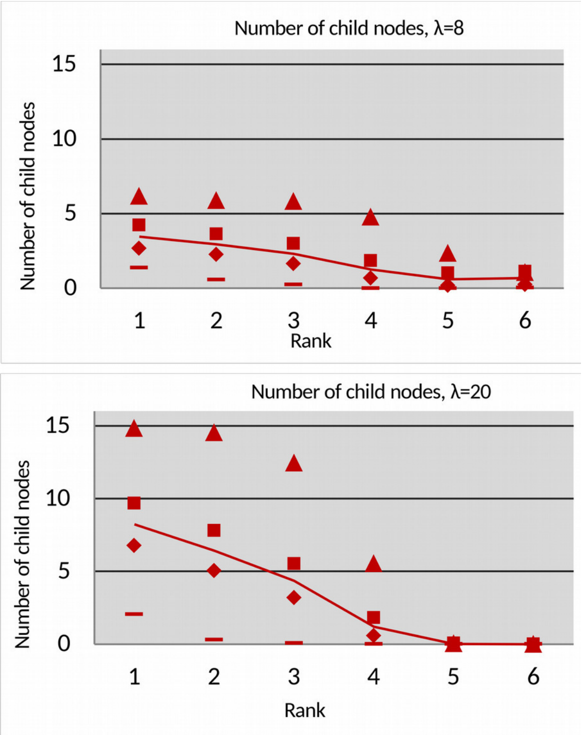


Fig 3. Number of child nodes versus rank for different node densities

Our SPOF-algorithm described in subsection III.D and presented with purple line and marks in Figure 2, gives the best balancing effect. The reason is that higher-rank nodes are encouraged to choose paths that does not include the nodes that act as SPOF. However, SPOF nodes are not completely unloaded from forwarding data since child nodes have to forward all traffic through their SPOF parent nodes.

At higher rank nodes the SPOF-algorithm gives equal or marginally less balance compared to all other weighted-fair-sharing algorithms. However, the most efficient energy-balancing algorithm is the algorithm that focuses on energy balance among the lowest rank nodes, although this may give reduced balancing effect at the higher-rank nodes. The reason is that the lowest rank nodes always consumes the highest average amount of energy.

Increased number of child nodes enlarges energy imbalance between nodes, in particular for algorithms that uses fixed preferred-parent nodes. The number of child nodes increases with increased node density and reduced rank. This is demonstrated in Figure 3 which shows the number of child nodes versus rank for two different node densities, λ=8 and λ=20. This figure supports the findings in Figure 1 related to the rapidly increase in energy imbalance for increased node density. The circled shaped markers in Figure 3 with the corresponding lines show the average values. The square shaped and the diamond shaped markers shows the 95% confidence interval. In addition, the highest values, averaged over all different-seeds simulations are shown as triangular markers. The lowest values, averaged over all different-seeds simulations are shown as short horizontal lines.

## V. CONCLUSION

Simulations presented in this paper show that the energy imbalance is substantial in network running the native RPL routing protocol. Thus, some kind of energy balancing algorithm should be used to prevent premature node depletion.

A total of nine energy balancing algorithms applicable for RPL running networks are analyzed in this paper. Six of the algorithms are based on various approaches suggested in literature. In addition, we suggest three new energy balancing approaches to complement the selection of algorithms. A common context is used to simulate and compare the performance of all the algorithms.

The simplest of our suggested approach is just a tiny adjustment of RPL's parent selection algorithm. Instead of using a preordain tiebreaker parameter, the preferred-parent (next-hop node) is randomly selected among the nodes in the parent-list. The adjustment gives a significant balancing effect. Especially in high density networks, where the residual-energy of the most depleted node is increased more than 10% compared to running native RPL.

The second and the third suggested algorithm use the nodes' residual-energy to weight-balance the transmitted traffic between all available parent nodes. In the second algorithm, a node with a single point of failure (SPOF) parent advertises a residual-energy level equal to the lowest of its own and its parent energy level. The algorithm requires residual-energy information to be exchanged between the nodes during RPL management packet exchange.

In the third algorithm, the nodes eavesdrop on the traffic in their vicinity to estimate neighboring nodes residual-energy level. Increased traffic density degrades the eavesdropping algorithm in the proximity of the sink.

Simulations shows that the SPOF algorithm performs best of all the weighting algorithms. Compared to native RPL is the SPOF algorithm increasing the residual-energy of the most depleted node with over 20%.

However, the best energy balancing is achieved when nodes choose the preferred-parent as the member of the parents list that has the highest residual-energy level. Simulations of the algorithm show that, compared to native RPL, the residual-energy of the most depleted nodes increases by 25%. The merit of this algorithm is that the lowest residual-energy paths are always avoided. On the contrary, weighting the traffic between all potential parent nodes means that also the lowest energy nodes are used, although rarely. However, to select the node with highest residual-energy, the nodes must exchange energy information. Randomly selecting the preferred-parent requires no extra information exchange.

Although the energy consumption is balanced, it is always the nodes closest to the sink that consumed the highest amount of energy. However, balancing the energy consumption of equal rank nodes can give reduced network management cost.

## REFERENCES

[1] C. Buratti, A. Conti, D. Dardari and R. Verdone, «An Overview onWireless Sensor Networks Technology and Evolution,» Sensors, Volume 9 Issue 9, pp. 6869-6896, 2009. http://dx.doi.org/ 10.3390/s90906869

[2] Y. Liu, Y. He, M. Li, J. Wang, K. Liu and X. Li, «Does Wireless Sensor Network Scale? Measurement Study on GreenOrbs,» IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, pp. 1983-1993, 2013. http://dx.doi.org/ 10.1109/TPDS.2012.216

[3] K. Heurtefeux, H. Menouar and N. AbuAli, «Experimental Evaluation of a Routing Protocol for WSNs: RPL robustness under study,» IEEE 9th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 491-498, 2013. http://dx.doi.org/ 10.1109/WiMOB.2013.6673404

[4] T. W. e. al., «RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks,» Request for Comments: 6550, 2012.

[5] P. L. e.al, «The Trickle Algorithm,» Request for Comments: 6206, 2011.

[6] P. O. Kamgueu, E. Nataf, T. D. Ndie and O. Festor, «Energy-based routing metric for RPL,» RR-8208, pp. 1-14, 2013.

[7] G. Xu and G. Lu, «Multipath Routing for DAG-based WSN with Mobile Sinks,» Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering ICCSEE, pp. 1678-1682, 2013.

[8] C. Abreu, M. Ricardo and P.M.Mendes, «Energy-aware routing for biomedical wireless sensor networks,» Journal of Network and Computer Applications, p. 270–278, 2014. http://dx.doi.org/ 10.1016/j.jnca.2013.09.015

[9] L. Chang, T. Lee, S. Chen and C. Liao, «Energy-Efficient Oriented Routing Algorithm in Wireless Sensor Networks,» International Conference on Systems, Man, and Cybernetics (SMC), pp. 3813 - 3818, 2013. http://dx.doi.org/ 10.1109/SMC.2013.651

[10] K. S. Shivaprakasha and M. Kulkarni, «Energy Efficient Shortest Path Routing Protocol for Wireless Sensor Networks,» International Conference on Computational Intelligence and Communication Networks CICN, pp. 333 - 337, 2011. http://dx.doi.org/ 10.1109/CICN.2011.70

[11] O.Iova, F. Theoleyre and T. Noel, «Improving the network lifetime with energy-balancing routing: Application to RPL,» Wireless and Mobile Networking Conference (WMNC), pp. 1 - 8 , 2014. http://dx.doi.org/ 10.1109/WMNC.2014.6878864

[12] S. Chiang, C. Huang and K. C. Chang, «A Minimum Hop Routing Protocol for Home Security Systems Using Wireless Sensor Networks,» Transactions on Consumer Electronics, pp. 1483 - 1489, 2007. http://dx.doi.org/ 10.1109/TCE.2007.4429241

[13] P. T. A. Quang and D. Kim, «Enhancing Real-Time Delivery of Gradient Routing for Industrial Wireless Sensor Networks,» TRANSACTIONS ON INDUSTRIAL INFORMATICS, pp. 61-68, 2012. http://dx.doi.org/ 10.1109/TII.2011.2174249

[14] M. N. Moghadam, H. Taheri and M. Karrari, «Minimum cost load balanced multipath routing protocol for low power and lossy networks,» Wireless Networks,Volume 20, Issue 8, pp. 2469-2479, 2014. http://dx.doi.org/ 10.1007/s11276-014-0753-7

[15] R. Vidhyapriya and P. T. Vanathi, «Energy Efficient Adaptive Multipath Routing forWireless Sensor Networks,» IAENG International Journal of Computer Science, pp. 56-64, 2007.

[16] K. Sha, J. Gehlot and R. Greve, «Multipath Routing Techniques in Wireless Sensor Networks: A Survey,» Wireless Personal Communications, pp. 807-829, 2013. http://dx.doi.org/ 10.1007/s11277-012-0723-2

[17] T. Rault, A. Bouabdallah and Y. Challal, «Energy efficiency in wireless sensor networks: A top-down survey,» Computer Networks, vol. Volume 67, p. 104–122, 2014. http://dx.doi.org/ 10.1016/j.comnet.2014.03.027

[18] Y. M. Lu and V. W. S. Wong, «An energy-efficient multipath routing protocol for wireless sensor networks,» International Journal of Communication Systems, p. 747–766, 2007. doi : 10.1002/dac.843

[19] A. Liu, J. Ren, X. Li, Z. Chen and X.Shen, «Design principles and improvement of cost function based energy aware routing algorithms for wireless sensor networks,» Computer Networks , p. 1951–1967, 2012. http://dx.doi.org/10.1016/j.comnet.2012.01.023

[20] N. A. Pantazis, S. A. Nikolidakis and D. D. Vergados, «Energy-Efficient Routing Protocols in Wireless Sensor Networks: A Survey,» Communications Surveys &amp; Tutorials, pp. 551-591, 2013. http://dx.doi.org/ 10.1109/SURV.2012.062612.00084

[21] C. Efthymiou, S. Nikoletseas and J. Rolim, «Energy balanced data propagation in wireless sensor networks,» Journal Wireless Networks , p. 691–707, 2006. http://dx.doi.org/ 10.1007/s11276-006-6529-y

[22] C. M. Angelopoulos, G. Filios, S. Nikoletseas, D. Patroumpa, T. P. Raptis and K. Veroutis, «A Holistic IPv6 Test-Bed for Smart, Green Buildings,» International Conference on Communications (ICC), pp. 6050-6054, 2013. http://dx.doi.org/ 10.1109/ICC.2013.6655569

[23] «OMNET++ : http://www.omnetpp.org/. (Mars 2015)».

[24] A.-L. Kampen, K. Øvsthus, L. Landmark and Ø. Kure, «Energy Reduction in Wireless Sensor Networks by Switching Nodes to Sleep During Packet Forwarding,» The Sixth International Conference on Sensor Technologies and Applications, SENSORCOMM, pp. 189-195, 2012.

[25] C. R. F. H. IEEE P802.15 Working Group, «Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Specifications for Low-Rate Wireless,» IEEE Std 802.15.4™-2006, 2006.

[26] K. K. Gagneja and K. E. Nygard «A QoS based Heuristics for Clustering inTwo-Tier Sensor Networks,» Proceedings of the Federated Conference on Computer Science and Information Systems, FedCSIS, pp. 779–784, 2012.

DOI: 10.15439/2015F167
ACSIS, Vol. 5

# Using of compressed sensing in energy sensitive WSN applications

Ondrej Karpiš, Juraj Miček, Veronika Olešnaníková
University of Zilina
Univerzitna 8215/1
010 26 Zilina, Slovakia
Email: {Ondrej.Karpis, Juraj.Micek, Veronika.Olesnanikova}@fri.uniza.sk

*Abstract*—**The paper is focused on the use of methods of compressed sensing (CS) in energy efficient monitoring of signals. CS allows to minimize the number of data that need to be transmitted to the sink node in the WSN environment. As a case study, we use compressed sensing for monitoring of mains voltage deformation. In this case we can assume that the measured signal is sparse in frequency domain and using of methods of compressed sensing is meaningful. Computational complexity imposed on the sensor node is minimized. On the other hand, reconstruction of the original signal in the sink node requires relatively high computing power.**

## I. Introduction

INTERNET is undergoing a third stage of development nowadays. Since 1995, the Internet has evolved from interconnecting desktops and later mobile devices (tablets, smart phones), via networking of all devices (Internet of Things) to the Internet of Everything - IoE (networking of things, people, data and processes). According to estimate of Cisco [1] there are about 200 things for every one person on the Earth that can be connected to the data network. It follows that in the near future we can see the network containing up to 1.5 x $10^{12}$ elements. Meaningful use of such a network is a huge challenge for the visionaries and theorists, but also for programmers, workers in the field of transmission technology and developers and technical resources. According to [2] we expect that in the next five years, the number of connected devices will increase from the current value of 12 x $10^{10}$ to about 35 x $10^{10}$, Fig.1.

In the IoE raises huge space for development and implementation of new applications of Wireless Sensor Netvorks (WSN). Recall that WSN consists of spatially distributed autonomous sensing elements that work together. They are distributed in the monitoring area and continuously evaluate the status of the monitored object. The term object here is understood in its broadest sense and may represent guarded area, production line or living being. Based on this definition the WSN represents a natural part of a global network IoE. The above projections indicate that in the near future we will see the growing number of WSN applications.

It is natural that during development of successful applications of WSN we are facing many constraints. Some of these constraints arise from the relationship of the society to the new information technologies (fear of loss of privacy, inability to effectively utilize the benefits of the system and many others). These limits are not analyzed in this paper. Let us mention technical limitations that determine the extent and success of WSN applications:

- limited energy resources of WSN elements,
- limited processing power, storage capacity, communication speed and range of broadcast modules of network elements,
- limited size of individual elements,
- limited (or even excluded) maintenance of the network elements during their lifetime,
- constraints imposed by the requirements for working conditions of WSN components,
- price limits for a single network element and the like.

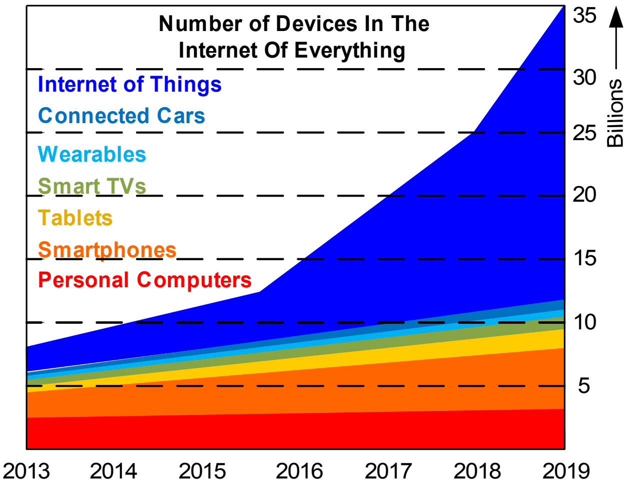


Fig 1. *The growth in the number of IoE devices*
*Source: BI Intelligence Estimates*

Many of these limits are becoming less important with the continuous development of new electronic components and technologies. Others, however, will permanently restrict the

This work was supported by the Research & Development Operational Programme funded by the ERDF.

development of new applications. One of such limitations that must be respected in the development of applications is the limited capacity of energy resources. Note that limited energy consumption of WSN node affects other technical parameters - computing power, communication speed and range, and so on.

The problem of efficient powering of network elements can be addressed in the following ways:

- consistent use of energy harvesting (EH) exploiting resources that are available in the application [3],
- appropriate design of network topology with minimal demands on communication, optimal distribution of network tasks, dynamic reconfiguration of the network with respect to the current state of energy resources of individual elements, etc.,
- reducing consumption of sensor nodes that can be achieved by combining two approaches:
    - using only low-power elements in the process of sensor node development,
    - choosing of such a mode of operation of sensor node that minimizes energy consumption.

Energy harvesting systems used for powering of network elements are usually designed so that they can ensure continued operation of the nodes. The term Zero Power Wireless Sensor [4] indicate precisely those solutions that do not need a power source for their operation, but are able to drain the necessary energy from the environment. The construction of sensor nodes is based on modern circuit with reduced consumption. During the operation the sensor nodes use sleep modes with reduced consumption and the sensor wakes up to the active state only when it is necessary. The wake up event is usually based on external conditions. Such systems are referred to as "event-driven". The node is brought into active mode only under specified conditions (e.g. change of the observed variables by defined amount, the achievement of pre-defined state of the object, etc.). In many cases (e.g. by monitoring of non-stationary processes) is such system more energy efficient when compared to conventional "Time-driven" systems.

However, there is a group of systems that strictly speaking cannot be classified into any of the above mentioned classes. It is a class of systems that use for collection, transmission and processing of information methods known under the name compressed sensing. The next section describes the basic theoretical background of compressed sensing.

## II. Compressed Sensing

The basis of the theory of digital processing of continuous signals is Shannon's theorem, which says that a perfect reconstruction of the sampled signal is only possible if the sampling frequency is at least twice the maximum frequency component contained in the original signal. The theorem is universally applicable, however, in some cases the strict requirement for the sampling rate can be substantially released. Recent research [5], [6] has shown that this is possible particularly in the case if the sampled signal is in some domain sparse. Sparse means that relatively few coefficients describing the signal in the domain are non-zero.

Let $y$ be a one-dimensional discrete signal comprising of $n$ elements. Next, let $x$ be the representation of the signal $y$ in some domain (e.g. Fourier or wavelet). For linear transformation it holds that $x = \Phi y$ and $y = \Psi x$ where $\Phi$ and $\Psi$ and are square $n$ $x$ $n$ matrices representing the direct and inverse transformation and are composed from linearly independent base vectors (usually orthonormal).

We say that the signal $x$ is $s$-sparse (for $1 \leq s \leq n$) if it has at most $s$ non-zero coefficients. Intuitively, if the signal is $s$-sparse, it should have only $s$ degrees of freedom. In that case one needs substantially only $s$ measurements for the reconstruction of the original signal. This is the basic idea of the compressed sensing - the number of measurements that is required for the perfect reconstruction of the original signal is directly proportional to its sparsity.

Let's show how it is possible to reconstruct the original signal $y$, if we have only $m$-dimensional vector of measurements $b$, while $s \leq m < n$. The challenge is to find a solution to the equation $Ax = b$, where $A$ has dimensions $m$ $x$ $n$ and is referred to as a measurement matrix. If $m < n$, the problem is underdetermined and generally it has infinitely many solutions. Provided that any subset of $2s$ columns of the matrix $A$ are linearly independent, the solution of the defined problem is the sparsest vector $x$:

$$min_x \|x\|_0 \quad \text{subject to} \quad Ax = b \qquad (1)$$

where

$$\|x\|_0 = \sum_{i=1}^{n} |x_i|^0 \qquad \{1 \leqslant i \leqslant n, x_i \neq 0\}$$

denotes the sparsity of the vector $x$.

Unfortunately, $l_0$ minimization is, in general, NP hard problem as it requires a search of all $\binom{n}{s}$ possible solutions. The problem is that the $l_0$ minimization is not a convex optimization problem, and therefore one cannot use the appropriate optimization algorithms. Replacing $l_0$ minimization by $l_2$ minimization (least squares problem) yields not satisfactory results. A more accurate estimate of the solution $x^e$ can be obtained by $l_0$ minimization:

$$x^e = min_x \|x\|_1 \quad \text{subject to} \quad Ax = b \qquad (2)$$

where

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$l_1$ norm as a means of finding the most sparse solution has been used already in 1973 in reflection seismology [7]. The problem (2) is a convex optimization problem, which can be effectively solved using methods of linear programming. It is often referred to as basis pursuit (BP). Equivalence of $l_0$ and $l_1$ minimization is guaranteed if either of the following sufficient conditions is met. The first condition is a requirement that the matrix $A$ approximately maintains the Euclidean length of the $s$-sparse signals. This characteristic of the matrix $A$ is called the restricted isometry property (RIP) [9]. RIP is related to the isometric constant $\delta_s$ defined as the smallest number such that

$$(1-d_s)\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1+d_s)\|x\|_2^2 \qquad (3)$$

holds for all $s$-sparse vectors $x$.

We can say that $A$ obeys the RIP of order $s$ if $\delta_S$ is not very close to one. It is met if all the subsets of $s$ columns of the matrix $A$ are approximately orthogonal. However, checking whether the matrix $A$ obeys the RIP is generally NP hard problem. Some matrices are known to obey the RIP with overwhelming probability (e.g. Random Gaussian Matrices, Bernoulli matrices, partial Fourier matrices). There is a prove in [12] that Gaussian and Bernoulli random matrices probably obey the RIP when the number of measurements satisfies the condition $m \geq Cs\ ln(n/s)$, where $C$ is some constant dependent on $\delta_S$.

It is much easier to verify the second sufficient condition based on mutual coherence $\mu(A)$. Mutual coherence is defined as the cosine of the smallest angle between any two columns of matrix $A$. Mutual coherence has to be as small as possible (cosine is close to zero for angles close to 90 °, which means that the columns are independent). Incoherence of the matrix $A$ indicates that if the signal is sparse in one domain, it has to be spread out in the other domain (the one in which it is sampled). In practice, this means that columns of the matrix $A$ are roughly uniform in magnitude.

Both sufficient conditions essentially require independence of columns of the matrix $A$. Interconnection of RIP and mutual coherence has been shown in [11]. For an orthogonal square matrix it holds that both $\delta_S$ and $\mu(A)$ equal to zero. However, for $m\ x\ n$ matrix ($m < n$) it is not possible to ensure the complete independence of columns, we just require them to be roughly orthogonal.

Simulation experiments in [18] have shown that there is rather universal dependence between the ratios $m/n$ and $s/m$. This dependence can be used for selection of appropriate number of samples $m$ if the sparsity $s$ of the signal is known. Practical experiments indicate that most $s$-sparse signals can be perfectly reconstructed if $m$ is in range from $3s$ to $5s$ [13].

We have shown that CS is applicable for the reconstruction of sparse undersampled signals. However, to be useful in practice, it is necessary for CS to cope with both nearly sparse signals and with noise. In other words, CS must be robust. Most real signals are not strictly sparse and measurements are corrupted with at least quantization noise, as the sensors do not have infinite precision. In this case, it is necessary to find solution to the equation:

$$b = Ax + e \qquad (4)$$

where $e$ is a random variable.

If the sensed signal is not strictly $s$-sparse, we require it to be at least $s$-compressible. $S$-compressible signal has at most $s$ significant coefficients and all the other coefficients are close to zero. In other words, if the coefficients are sorted by value, all coefficients, except for the first $s$, are smaller than some small nonzero constant. Thus, the majority of the signal's information content is concentrated in only a few coefficients. Typical examples of the compressible objects are images when expressed in appropriate base (e.g. wavelet). This feature of natural images has been used in compression algorithms such as JPEG2000 for years. We can say that many images are efficiently sparse in wavelet base.

To reconstruct the original signal in case of noisy measurements one can use $l_1$ minimization with relaxed conditions:

$$min_x \|x\|_1 \quad \text{subject to} \quad \|Ax-b\|_2^2 \leqslant \epsilon \qquad (5)$$

This is a linear programming optimization problem with quadratic conditions (Basis Pursuit Denoising - BPDN). The problem can be reformulated as quadratic programming problem with linear conditions (known as LASSO):

$$min_x \|Ax-b\|_2^2 \quad \text{subject to} \quad \|x\|_1 \leqslant \epsilon \qquad (6)$$

For appropriately selected parameter $\lambda$ the problem (5) can be expressed without conditions:

$$min_x \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \qquad (7)$$

Several types of algorithms can be used to reconstruct the original signals. Methods based on a convex optimization are computationally demanding. This category includes Basis Pursuit (BP), Basis Pursuit Denoising (BPDN), interior point methods [17] and projected gradient methods. This category is sometimes supplemented by Iterative Shrinkage algorithms such as Iterative Hard Thresholding (IHT).

The second group of algorithms consists of greedy algorithms that are looking for non-zero coefficients incrementally, starting with the most significant. Among greedy algorithms belong mainly Matching Pursuit (MP) [14] and its variations: Orthogonal Matching Pursuit (OMP) [8], regularized OMP (ROMP), Stagewise OMP (StOMP), OMP with Replacement (OMPR) and also the first sub-linear algorithm OMPR-Hash. This category includes also Subspace Pursuit (SP) [16] and Least Angle Regression(LARS) [15]. Some of greedy algorithms are capable of providing similar guarantees of stability as BPDN while they are faster. Also they have the advantage of being easier to understand.

There are also combinatorial algorithms (e.g. HHS pursuit), which are very fast, but they require a lot of measurements.

Special types of algorithms are Total Variation (TV) algorithms. These are used mainly in the reconstruction of images where one may require sparsity of the gradient of the image. TV algorithms are particularly suitable for images composed from smooth areas separated by curves (objects without complex textures). Such images are common in medicine (MRI, angiogram, and the like).

Compressed sensing can be used if:

- the signal is sparse in any known base,
- measurements or calculations at the sensor are expensive in some sense,
- calculations at the receiver are cheap.

The application area of compressed sensing is very broad: MRI, astronomy, WSN, communication [10] and so on. The use of CS in wireless sensor networks is particularly useful. Compressed sensing allows to substantially simplify sensor nodes. The use of compressed sensing saves limited energy resources at several levels:

- Signal sampling - the number of AD conversions required is a fraction compared to the conventional sampling using Nyquist frequency.
- Preprocessing of the signal - transfer of all acquired samples is usually undesirable. It is necessary to reduce the amount of data to be transmitted by a communication channel. Signal preprocessing algorithms often involve computationally intensive transformations such as Discrete Fourier transformation. Compressed sensing does not require pre-processing at all.
- Data transfer - energy is saved especially if we need to transfer the whole measured signal to the central node. Transmission of data is typically the most energy intensive operation in the processing of distributed data. Transmission of a smaller number of data has a positive effect on energy balance of the node and also on the throughput of the entire network.

From the perspective of information processing the block structure of the node only consist of two parts: sensing and transmission. Simplification of the node and a reduction of its consumption significantly increases its reliability and extends its lifetime.

Note that the sampling process can be done in several ways. We can use random sampling, but then we must transmit pairs of numbers (sample-time). The major advantage of the compressed sensing is then deteriorated. The second possibility is to use a pseudo-random sampling when the signal is sampled according to a predetermined scheme, which, of course, has to be known in the central node, in which the original signal is recovered.

Pseudo-random sampling is advantageous also because it is possible to fit it for a specific application. The sensing matrix can be defined in order to meet all the prerequisites needed for perfect signal reconstruction. The compressed sensing can be realized also in such a way that the signal is captured at a Nyquist rate and then a subset of the samples corresponding to the sensing matrix is picked out. In this case there is no energy saving during sampling but only in blocks of preprocessing and transmission. The development of sensors that allow sensing of physical quantities using the principles of compressed sensing is one of intensively researched areas. Successful penetration of CS into WSN depends mainly on the availability of suitable (cheap and energy-efficient) sensing elements.

## III. Example of using CS in WSN

One of the cases where it is possible to effectively use the compressed sensing is the monitoring of power quality. To provide high quality electric power service, it is essential to monitor number of parameters at different places in the network. Among the monitored parameters belong current and voltage RMS, phase relationship between waveforms of a multi-phase signal, power factor, frequency, total harmonic distortion, different kinds of power and many more. In this example we will focus on measuring of total harmonic distortion (THD).

THD can be calculated using following equation:

$$THD = \frac{\sqrt{\sum_{i=2}^{N} V_i^2}}{V_1} \tag{8}$$

where $V_i$ is the RMS voltage of *i*-th harmonic and $i = 1$ is index of the fundamental frequency component.

As indicated by (8), we need to know the frequency spectrum of the measured signal in order to calculate the THD. The second option is to use filters to obtain the fundamental component and all other components [19]. In both cases it is necessary to make quite a number of calculations on the sensor side.

THD monitoring is now more important than ever, given the increasing use of switching power supplies to power consumer electronics. Switching power supplies are causing clipping of supply voltage. Fig. 2 shows a clipped sine wave and its spectrum. The frequency of the sine wave is 50 Hz and the amplitude is clipped to 0.9.

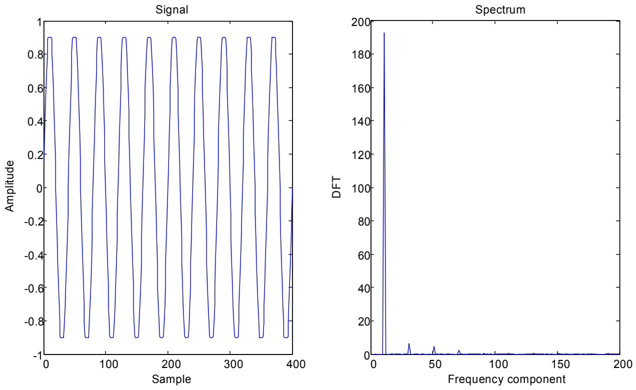


Fig 2. *Clipped sine wave and its spectrum*

The more clipped is the sine wave, the more energy is concentrated in the higher harmonics and the greater is the value of THD. If both half-waves are clipped, non-zero coefficients correspond to odd order harmonics. Clipping only one half of the sinusoid causes a doubling of the number of harmonics (all the harmonics are non-zero). However, this case is not common in practice. The number of non-zero coefficients in the spectrum of clipped sinusoid is theoretically infinite and must be reduced to make calculation feasible. This truncation causes some error, but it is relatively small and it can be neglected. For example, the THD of the example signal is 4.6416 % taking into account only the first ten non-zero coefficients or it is 4.6444 % if the number of non-zero coefficients is one hundred.

Since the representation of the clipped sine wave is sparse in the frequency domain, it is possible to engage principles of the compressed sensing. Compressed sensing allows reducing the cost of the sensor element but, on the other hand, increases the demand for the computing power of the central network node.

To verify the possibility of using compressed sensing in this case we conducted simulation experiments. For the calculation of THD we used first 20 higher harmonics (i.e. first 10 non-zero coefficients). The fundamental frequency equals to line frequency: 50 Hz. The 20-th harmonic then corresponds to 1 000 Hz. According to the Shanon theorem, the signal must be sampled by frequency of at least 2 000 Hz. Suppose that the mains voltage is stationary within an interval of 200 ms (i.e. the statistical properties of the voltage are constant for a short period of time). In this example, we monitor the mains voltage within time windows of a length of 200 ms, which corresponds to the number of samples *n = 400*. Random sampling was realized by random selection of *m*-samples such that the mean period between two samples equals to *n/m*. Input signal is a sine wave with a frequency of 50 Hz and amplitude clipped to 0.9:

$$x(t)=\begin{array}{ll} 0.9, & \sin(2\pi ft)\geqslant 0.9 \\ \sin(2\pi ft), & |\sin(2\pi ft)|<0.9 \\ -0.9, & \sin(2\pi ft)\leqslant -0.9 \end{array} \qquad (9)$$

Using Monte Carlo method, we performed simulations for different values of *m* and different values of signal to noise ratio (SNR). For each pair {m, SNR} we conducted 100 simulations. The simulation consists of a reconstruction of the original signal in the frequency domain and the subsequent calculation of THD using (8). For the reconstruction of signal we used collection of MATLAB routines l1-magic [20]. The resulting THD value is calculated as the arithmetic average of the 100 values. Fig. 3 shows the result of simulations. The THD of the original noise-free signal is 4.32 %.

Reconstruction of the original signal took from 1 to 5 seconds on dual core 2.6 GHz Athlon 64 processor. The duration of simulation depends on the length of the original signal *n,* the number of samples *m* as well as their distribution.

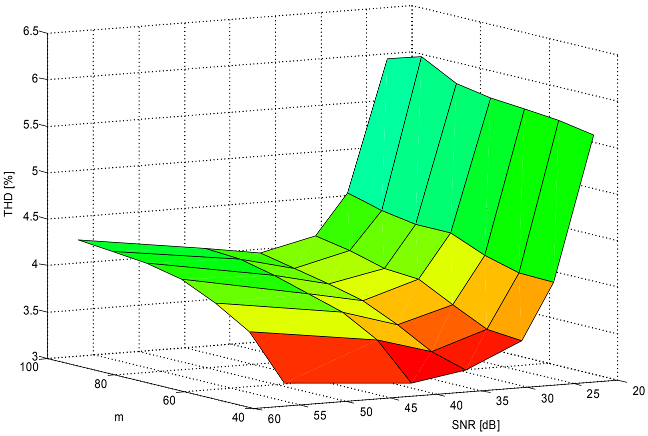


Fig 3. *THD as a function of m and SNR*

The results show that decreasing the number of samples and/or increasing the SNR deteriorate the accuracy of the reconstruction of the signal and thus the accuracy of the calculated THD. The simulations suggest that good results can be obtained when the number of samples *m* is at least 80 and the SNR is greater than 30 dB. In that case the compression ratio would be 1:5. If the mains voltage is stationary over longer periods, it is possible to achieve even higher compression ratios. If we increase the length of the time window to 400 ms (n = 800), it is possible to achieve even better results using the same compression ratio *m/n*. The effect of increasing the number of samples *n* is shown in table 1. The simmulation was done with noise-free sinusoid.

TABLE I.
EFFECT OF DIFFERENT WINDOW LENGTHS ON ACCURACY OF THD ESTIMATE

| **m/n** | **THD [%]** | | |
|---|---|---|---|
| | **n = 400** | **n = 600** | **n = 800** |
| 0.1 | 3.19 | 3.8 | 4.13 |
| 0.125 | 3.7 | 4.12 | 4.25 |
| 0.15 | 3.99 | 4.24 | 4.29 |
| 0.175 | 4.19 | 4.29 | 4.3 |
| 0.2 | 4.28 | 4.3 | 4.3 |

We see that doubling the length of the window allows to reduce the ratio *m/n* from 0.2 to 0.15. Knowledge of the parameters of measured signal is therefore essential in optimal design of compressed sensing parameters.

## IV. CONCLUSION

Simulation experiments demonstrate a possibility to use compressed sensing in appropriately selected applications of WSN. Energy savings due to the use of compressed sampling can be maximized if we know parameters of the measured signals.

In the future, we will focus on design of deterministic sensing matrices suitable for compressed sensing of selected signal classes. It is also necessary to develop fast algorithms for reconstruction of the original signals so they can be used even in less powerful network nodes.

## Acknowledgment

This contribution is the result of the project implementation *Centre of excellence for systems and services of intelligent transport I.* ITMS 26220120028 supported by the Research & Development Operational Programme funded by the ERDF.

Európska únia
Európsky fond regionálneho rozvoja

Operačný program
VÝSKUM a VÝVOJ

Agentúra
Ministerstva školstva, vedy, výskumu a športu SR
pre štrukturálne fondy EÚ

"Podporujeme výskumné activity na Slovensku/
Projekt je spolufinancovaný zo zdrojov EÚ"

## References

[1] J. Bradley, J. Barbier, D. Handler, "Embracing the Internet of Everything to Capture Your Share of $14.4 Trillion", White Paper, Cisco, 2013.

[2] J. Greenough, The Internet of Everything, www.businessinsider.com/internet-of-everything-2015-bi-2014-12

[3] J. M. Gilbert, F. Baluochi, "Comparison of Energy Harvesting Systems for Wireless Sensor Networks", *International Journal of Automation and Computing*, October, 2008, http://dx.doi.org/10.1007/s11633-008-0334-2

[4] S. Grady, "Powering Wearable Technology and Internet of Everything Devices", Cymber Corporation, 2014, www.cymbed.com

[5] E. Candès, J. Romberg, T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006, http://dx.doi.org/10.1109/TIT.2005.862083

[6] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289-1306, Apr. 2006, http://dx.doi.org/10.1109/TIT.2006.871582

[7] J. F. Claerbout, F. Muir, "Robust modeling with erratic data," *Geophys. Mag.*, vol. 38, no. 5, pp. 826-844, Oct. 1973, http://dx.doi.org/10.1190/1.1440378

[8] J. Tropp, A.C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655- 4666, 2006.

[9] E. Candès, T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203-4215, Dec. 2005, http://dx.doi.org/10.1109/TIT.2005.858979

[10] J. Arenas-Garcia, A. R. Figueiras-Vidal, "Adaptive combination of proportionate filters for sparse echo cancellation", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 6, pp. 1087-1098, 2009, http://dx.doi.org/10.1109/TASL.2009.2019925

[11] T. T. Cai, G. Xu, J. Zhang, "On recovery of sparse signals via $l_1$ minimization", *IEEE transactions on Information Theory*, Vol. 55, No. 7, pp. 3388-3397, 2009, http://dx.doi.org/10.1109/TIT.2009.2021377

[12] S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, "Uniform uncertainty principle for Bernoulli and subgaussian ensembles". *Constructive Approximation*, Vol. 28, pp. 277-289, 2008, http://dx.doi.org/10.1007/s00365-007-9005-8

[13] E. J. Candes, J. Romberg, "Practical signal recovery from random projections", In: *Proceedings of the SPIE 17th Annual Symposium on Electronic Imaging*, San Jose, 2005.

[14] S. Mallat, S. Zhang, "Matching Pursuit with time-frequency dictionaries", *IEEE Transactions on Signal Processing*, Vol. 41, No. 12, pp. 3397-3415, 1993, http://dx.doi.org/10.1109/78.258082

[15] B. Efron, T. Hastie, I. M. Johnstone, R. Tibshirani, "Least angle regression", *Annals of Statistics*, Vol. 32, No. 2, pp. 407-499, 2004, http://dx.doi.org/10.1214/009053604000000067

[16] W. Dai, O. Milenkovic, "Subspace Pursuit for Compressive Sensing Signal Reconstruction", *IEEE Transactions on Infor- mation Theory*, Vol. 55, No. 5, pp. 2230-2249, 2009, http://dx.doi.org/10.1109/TIT.2009.2016006

[17] S. J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, "An Interior-Point Method for Large-Scale $l_1$-Regularized Least Squares", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 1, No. 4, pp. 606-617, 2007, http://dx.doi.org/10.1109/JSTSP.2007.910971

[18] D. Donoho, J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, issue 1906, pp. 4273-4293, 2009, http://dx.doi.org/10.1098/rsta.2009.0152

[19] G. E. Mog, E. P. Ribeiro, "Total harmonic distortion calculation by filtering for power quality monitoring," *Transmission and Distribution Conference and Exposition: Latin America*, 2004 IEEE/PES, pp. 629-632, 8-11 Nov. 2004, http://dx.doi.org/10.1109/TDC.2004.1432452

[20] MATLAB routines: http://users.ece.gatech.edu/~justin/l1magic/

DOI: 10.15439/2015F355
ACSIS, Vol. 5

# Control Unit for Power Subsystem of a Wireless Sensor Node

Michal Kochláň, *IEEE Student Member*
Department of Technical Cybernetics
Faculty of Management Science and Informatics
University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: michal.kochlan@fri.uniza.sk

Samuel Žák, Juraj Miček, Jana Milanová
Department of Technical Cybernetics
Faculty of Management Science and Informatics
University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: {samuel.zak, juraj.micek, jana.milanova}@fri.uniza.sk

***Abstract*—This paper addresses analysis of the principle of a bi-directional electric energy charge pump and describes circuit solution for control unit of a wireless sensor node power subsystem. This power and control unit, as the power subsystem of the wireless node, comprises two supercapacitors. Bi-directional energy flow between these two supercapacitors takes place in order to optimally satisfy the sensor node energy requirements. At the same time, the bi-directional energy flow enables optimal energy harvesting from the selected energy scavenging (harvesting) subsystem. The mentioned Power Unit Control System has been successfully verified in the process of energy harvesting from a solar module.**

## I. Introduction

IN RECENT years, we encounter with many new unconventional applications of wireless sensor networks (WSNs). Related to the expected rapid expansion of so called *Internet of Everything* (IoE) applications [1], [2], [3], the problem of WSN is about to stand in the center of scientists community interest as well as of the public.

Basic limiting factors of WSN expansion are price and unattended operation time [4]. Both mentioned limitations are closely related to the energy requirements of essential nodes of a sensor network. The issue of effective sensor node power supply is being solved by two basic methods [5], [6]:

- Reducing a sensor node (SN) energy requirements,
- Consistent utilization of energy harvesting (EH) options from sources that are available in given application.

Both methods are used these days. Modern low power components are used for construction of present SNs, which help to reduce the energy requirements. In running mode, the SN uses modes with reduced power consumption and their combinations [7]. This approach turns MCU into active mode only in case the application or (external) conditions of monitored environment require it. Second option is the utilization of EH systems that are designed so that they would be able to ensure SN operation [8]. Today, there is a popular term *Zero Power Wireless Sensor* [9], which represents just solutions that do not need power supply for continuous operation, but they are able to harvest required energy from the environment. However, until recently, the lifetime of such solutions was limited to finite battery life or the number of charging cycles [10]. These limitations have been overcome by supercapacitors since they are characterized by high number of charging cycles (up to 1 million) [21]. Combination of low-power integrated circuits, effective systems of energy transformation from environment to electric energy and advanced tools for electric energy accumulation enables pushing the boundaries of WSN applications to the areas (mostly applications of information and communication means), where no one would have imagined their effective operation few years ago.

## II. Sensor Node

Sensor networks consist of large number of sensor nodes (SNs). The SNs are deployed in a sensor field, where each of these sensor nodes has the capabilities to collect and route data to the sink and the end users [12]. In order the basic characteristics of a sensor network could be satisfied, each SN has to perform the following functions [13]:

- Data collection,
- Data preprocessing,
- Communication.

It is obvious that except the mentioned basic functions of each network node, there are also other important SN parameters, such as [14]:

- Computing performance of the sensor node,
- Low power consumption/long lifetime,
- Production cost,
- Security,
- Fault tolerance.

Many of mentioned requirements are contradicting each other [12]. Increasing computing performance increases energy requirements as well as sensor price. Enhancing transmission security involves higher computing performance. Higher fault tolerance usually negatively influences production cost. Therefore, it is impossible to develop appropriately universal node, which would be optimal in terms of all mentioned requirements. Moreover, it should be noted that different applications of SNs set different weight for the mentioned requirements. Problem of sensing elements designing is described in

This work was supported by *Centre of excellence for systems and services of intelligent transport I.* and *Research Centre of University of Žilina.*

978-83-60810-66-8/$25.00©2015, IEEE

[15]. The authors propose to use class of universally usable modules for development and application-oriented carriers and, thus, reach high degree of universality with possibilities to use the developed systems in the widest set of various applications. Data collection is performed by the data collection subsystem of a sensor node. This subsystem is extremely dependent on the requirements of the particular application and consist of sensing elements observing parameters that are relevant for solving the desired task. It is clear that different sensing elements are required for road traffic monitoring, other for patient telemonitoring and absolutely another for environment monitoring.

Because the energy requirements for data transmission are very high [5], it is suitable to perform significant part of data processing process directly at place of acquisition - in the sensor node. Then, the other network nodes will transfer only data that contain substantial information for solving the given problem.

It is clear, that in each sensor node data compression is executed. Then, the fundamental processes of data pre-processing relate to extraction methods of information content. Not in all WSN applications is the sensor node's task relegated only to the operations related to signal compression, increasingly there are used *Collaborative Signal Processing Algorithms* [16], that use suitable distribution of partial task among each nodes in order to increase the computing performance of network while minimizing of whole energy consumption. Today there is a wide range of various methods that can be successfully implemented onto WSN platform (Kalman filtering, neural networks, Bayesian inference etc.) [17].

Wireless communication ability is one of the basic features of a sensor node [5]. Communication abilities of WSN nodes are in most cases limited by communication range and transmission speed. It should be reminded that with increasing communication range the transmission power/sensitivity raises. This always leads to increased requirements on energy consumption.

Based on the mentioned functions and parameters of the SN, it is possible to create a generalized schematic block of a sensor node, which is depicted on Fig. 1. Each WSN node contains four basic units [4]:

- Sensing unit,
- Processor unit,
- Communication unit,
- Power unit.

In terms of WSN utilization properties, operating time of the network without need of operator intervention is important. The critical point is just the power supply of the SN. As mentioned earlier, the unit that takes care of power supply for a sensor node is called power unit. This part of sensor node can be simply represented by a power system with use of primary (battery) cells, or it may be constructed of a complex power supply system with energy harvesting (EH) features with effective energy accumulation. The following section describes the power unit developed at our department.

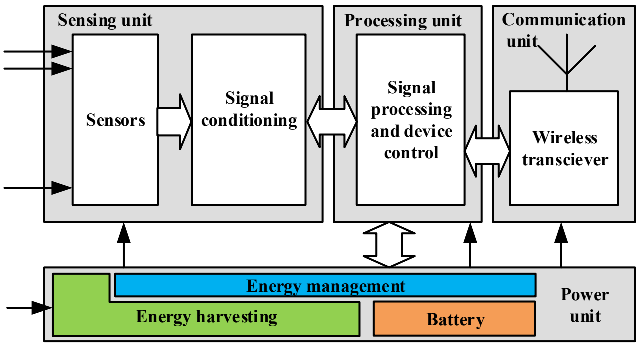


Fig. 1. Wireless node

## III. Energy Management System

Ensuring the continuous effective operation of a sensor node, which has limited energy sources is a difficult problem. Solving this consists of:

- Right choice of EH type,
- Right choice of energy storage capacity,
- Designing such sensor node operation mode that minimizes energy consumption.

Subsystem that covers energy requirements of a sensor node is *power unit*. The power unit can be implemented in several ways. Today we encounter very simple power supply systems using primary (battery) cells that are designed so that the energy requirements and desired SN lifetime are met.

### *A. Power Unit*

Power supply of a SN is realized through the subsystem, which is often referred as power unit. This supply subsystem can be based on one of the following options [18]:

- Supply from primary (battery) cells,
- Supply from rechargeable batteries,
- Energy harvesting (EH) features,
- Near Field Charging (NFC).

Today one can often meet complex solutions for power units that satisfy energy management in such way that all WSN energy requirements are compressed down to zero (zero power systems) [19]. Such unit consists of three main parts, Fig. 2:

- Energy harvesting system,
- Energy processor,
- Energy storage.

Functions of each subsystem of power unit result from their names. EH is used to temporary preserve energy and energy management controls the operation of the whole power unit in the way that it is possible to ensure all requirements of a sensor node with reaching the maximal energy effectiveness.

Unless the application permits performing basic maintenance during sensor node lifetime - battery exchange, then rechargeable batteries are often used. In those simple cases is energy consumption control (transitions to energy saving modes of individual subsystems) often performed by the SN processing unit. In case more complex power units are used

TABLE I
SELECTED PARAMETERS OF ENERGY STORAGE SYSTEMS ( 1KWH=3,6 MJ)

| Type | | Specific energy [MJ/kg] | Number of cycles N | Self discharge rate [%] | Charge/discharge Efficiency [%] |
|---|---|---|---|---|---|
| Primary batteries | Alkaline | 0.6 | - | 3 per year | - |
| | Lithium | 2 | - | 2 per year | - |
| Rechargeable (liquid electrolyte) | Lead Acid | 0.14 | 500-800 | 3-20 per month | 50-90 |
| | NiMH | 0.36 | 500-1800 | 8 per month | 66 |
| | Li-Ion | 0.9 | 400-1200 | 5 per month | 80-90 |
| | Li-Poly | 0.95 | 500-1000 | 8 per month | 80-90 |
| Rechargeable (solid electrolyte) | ss-battery EnerChip | 1.8 | 5000 | 2 per year | 90 |
| Supercapacitors | Twolayers | 0.01-0.02 | 500 000 | 50 per week | 95 |
| Pseudosuper capacitors | Li-Ion | 0.055 | 10 000 | 10 per month | 90 |

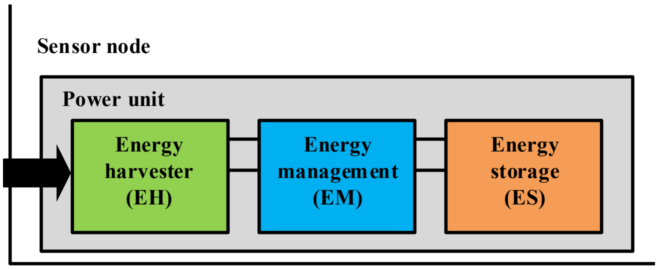


Fig. 2. Power unit

(combined energy storage and energy harvesting, etc.) the control of individual subsystems is often performed by a specialized unit - energy processor. When designing the energy management (EM) system, it is possible to use some of commercially available solutions. As an example, it is possible to mention energy processor CBC915 [20]. Other possibility is to create the Power management unit based on Low power MCU. That solution is universal enough and usable for wide range of applications with low-power energy consumption.

### B. Energy Storages

Energy storage (ES) is one of the basic parts of power unit. It is used for energy storage that is used for SN supply. ES accumulates the electric energy and delivers it to the powered device. In case the EH systems is used and that is not able to generate sufficient amount of energy from its environment, the ES temporarily helps EH subsystem with delivering enough power. It is clear, that either case (using EH or not) there is always need for energy to be accumulated and preserved in some way. In present, there are usually used the following "media" for energy accumulation:

- Primary batteries,
- Rechargeable batteries,
- Solid state batteries,
- Supercapacitors.

Selected parameters of above listed options of energy storages are described in Table I for comparison. Note, that in terms of long-term node operation, particularly the number of charge/discharge cycles is important, partially also a self-discharge rate, but in terms of continuous operation of node with EH system utilization energy efficiency is important. Therefore, it is clear that there is no commonly valid criterion, which recommends specific type of energy storage. Selection is always dependent on specific conditions and nature of the application.

One of the parameters of modern energy storage systems is specific energy (energy density), which moves from 2 $MJ/kg$ with high-end primary lithium batteries up to 0.01 $MJ/kg$ with supercapacitors [18]. Specific energy of primary lithium batteries is up to 100-times higher then current supercapacitors can reach. But specific energy is not the only parameter for choosing suitable type of energy storage. Not the less important parameters are number of charging cycles, self discharge rate, operating conditions, etc.

Because the parameters and limitations of each battery type vary quite much, there are often combinations of various types of elements for energy storage. In many applications it is possible to preferably use the combination of supercapacitor and rechargeable battery. For example, this combination benefits from high number of charge/discharge cycles of a supercapacitor and high specific energy of a battery. Especially, this combination appears a lot in power units based on EH.

### C. Energy Harvesting

EH represents ability to obtain the energy from natural resources in environment and transforms it to the electric energy [6]. Ambient energy, that can be used for power supply of low-power electronic systems is in various forms. It depends on the specific application, which takes the available energy in a suitable form. Common types of ambient energy sources, which is possible to use in EH systems for WSN are [21]:

- Biochemical energy,
- Mechanical energy,
- Acoustic noise energy,
- Wind energy,
- Thermal energy,
- Photovoltaic energy,
- Wireless energy.

It is obvious that it is quite impossible to define generally applicable rule, which type of energy is optimal for utilization in WSN. Choice of power source type depends on specific requirements of the SN and environment possibilities, which the node operates in. Basic comparison of possibilities that can be used when designing EH system are summarized in Table **??**.

TABLE II
POWER DENSITY OF EH TECHNIQUE

| **Energy harvesting technique** | **Power density** |
|---|---|
| Photovoltaic | Outdoors (sunny day) $15\mathrm{mW/cm^2}$<br>Indoors $10\mu\mathrm{W/cm^2}$ |
| Thermal energy - Thermoelectric | Human $30\mu\mathrm{W/cm^2}$<br>Industrial $110\mathrm{mW/cm^2}$ |
| Mechanical energy - Piezoelectric | $250\mu\mathrm{W/cm^3}$ |
| Mechanical energy - Electromagnetic | Human $1-4\mu\mathrm{W/cm^3}$<br>Industrial $800\mu\mathrm{W/cm^3}$ |
| Mechanical energy - Electrostatic | $50100\mu\mathrm{W/cm^3}$ |
| Wireless energy - RF radiation | GSM $0.1\mu\mathrm{W/cm^2}$<br>WiFi $0.01\mu\mathrm{W/cm^2}$ |
| Wind energy | $380\mu\mathrm{W/cm^3@5m/s}$ |
| Acoustic noise | $0.96\mu\mathrm{W/cm^3@100dB}$<br>$0.003\mu\mathrm{W/cm^3@75dB}$ |

In WSN applications with EH systems photovoltaic energy, thermal energy and mechanical energy are most frequently used. In present, there are appearing always new systems of effective energy conversion from environment to electric energy, which encourages the development of new, unconventional solutions.

## IV. Principled Design of Proposed Power Unit Management System

In the context of increasing of capacity elements for storing electric energy (supercapacitors) new applications arise, in which traditional solutions that use rechargeable or primary sources are replaced by supercapacitors. A solution for efficient energy flow between supercapacitors $C1$ and $C2$ or rechargeable cells is shown in schematic Figure 5. First, let's have a look at the principle of the energy flow and then the technical solution will be presented.

### *A. Energy flow from C1 to C2 (direct mode)*

The description of the basic function comes out from the assumption that the voltage on the capacitor $C1$ is within the range of permissible working voltage of the used control circuit (CC).

CC is capable of $C1$ capacitor voltage monitoring via $ADC1$ and $ADC2$, and similarly on the secondary capacitor $C2$. In case it is necessary to transfer energy from the capacitor $C1$ to the capacitor $C2$, CC closes the switch $S1$ through $DO1$ for time $T11$. Immediately after the time $T11$ switch $S1$ closes and opens switch $S2$ for time $T21$. The energy stored in the inductance $L$ at the time $T21$ is transferred to the capacitor $C2$. Waveforms of control signals of switches $S1$ and $S2$ are shown in Figure 3.

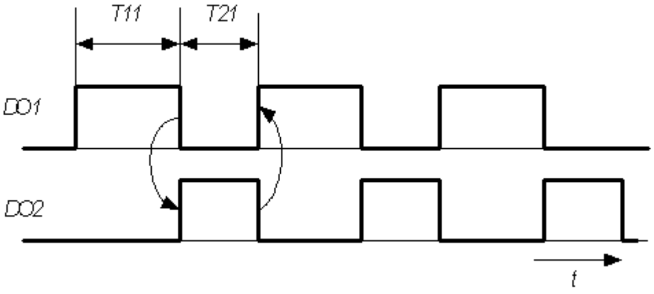


Fig. 3. Control signals in direct mode

Time $T11$ can be calculated with the following assumptions:

- Assume that the inductance coil $L$ is used. Let the $I_{L_{max}}$ be the maximum current that is the inductance coil able to handle;
- The voltage $U1$ denotes voltage at capacitor $C1$ and the voltage on the capacitor $C2$ is labeled $U2$.

Then for time $T11$ applies:

$$T11 \leq \frac{L \cdot I_{L_{max}}}{U_1}. \tag{1}$$

For time $T21$ applies the following:

$$T21 \leq \frac{U_1}{U_2} \cdot T11. \tag{2}$$

Please note that from (1) arises the fact that for times higher than $T11$ there is a current overload of inductance $L$. At higher times than the switch-on time $T21$ there is a discharge of supercapacitor $C2$, which results from (2), and thereby reduces the efficiency of the device.

Further, the CC has up-to-date information about the $U1$ and $U2$ voltages through an integrated ADC peripheral. On the basis of equations (1) and (2) switching times $T11$ and $T21$ can be optimally calculated by CC.

Essential amount of energy transferred in one cycle can be expressed by the following formula:

$$\Delta E = \frac{1}{2} C1 \cdot U_1^2 \left(1 - cos\left(\frac{T11}{\sqrt{L \cdot C1}}\right)\right). \tag{3}$$

Eventually, in simplified form where $T11 << \sqrt{LC1}$ applies, equation (3) turns into:

$$\Delta E = \frac{1}{4} \frac{U_1^2 \cdot T11^2}{L}. \tag{4}$$

*Remark 4.1:* Simultaneous opening of switches $S1$ and $S2$ leads to the destruction of the circuit, therefore, it should be strictly taken to comply with equations 1 and 2. These equations are approximate relations, while assuming that the opening time $T11 << \sqrt{LC1}$ and $T21 << \sqrt{LC2}$ with sufficient accuracy to satisfy the requirements of practical implementation.

### *B. Energy flow from C2 to C1 (inverse mode)*

In case it is necessary to transfer the energy in the opposite direction than described in the previous subsection - from capacitor $C2$ to the capacitor $C1$ - the switching mode of switches $S1$ and $S2$ is similar to the previous mode, except that the order of switching is reversed. In this case, CC through $DO2$ closes at first switch $S2$ for time $T22$ and after that the switch $S2$ is closed and switch $S1$ opens for time $T12$. The energy stored in the inductance $L$ is during time $T12$ transfered to capacitor $C1$. Waveforms of control signals of switches $S1$ and $S2$ are shown in Figure 4.

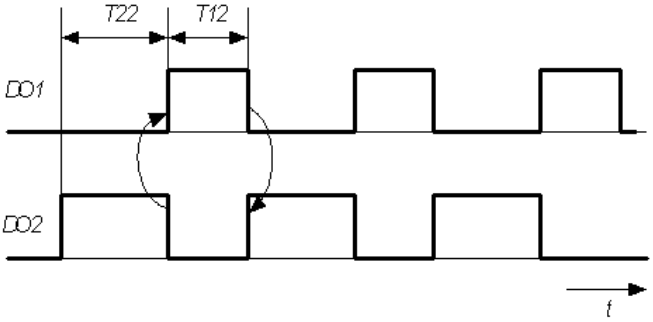


Fig. 4. Control signals in inverse mode

Similarly to the previous case, the conditions for calculating time $T22$ are as follows:

- Assume that the inductance coil $L$ is used. Let the $I_{L_{max}}$ be the maximum current that is the inductance coil able to handle;
- The voltage $U1$ denotes voltage at capacitor $C1$ and the voltage on the capacitor $C2$ is labeled $U2$.

Then for time $T22$ applies:

$$T22 \leq \frac{L \cdot I_{L_{max}}}{U_2}. \tag{5}$$

For time $T12$ applies the following:

$$T12 \leq \frac{U_2}{U_1} \cdot T22. \tag{6}$$

Even in inverse mode, the similar restrictions. Therefore, please note that from (5) arises the fact that for times higher than $T22$ there is a current overload of inductance $L$. At higher times than the switch-on time $T12$ there is a discharge of supercapacitor $C1$, which results from (6), and thus reduces the efficiency of the device.

Further, the CC has up-to-date information about the $U1$ and $U2$ voltages through an integrated ADC peripheral. On the basis of equations (5) and (6) switching times $T22$ and $T12$ can be optimally calculated by CC.

*Remark 4.2:* Simultaneous opening of switches $S1$ and $S2$ leads to the destruction of the circuit, therefore, it should be strictly taken to comply with figures 5 and 6. These equations are approximate relations, while assuming that the opening time $T22 << \sqrt{LC2}$ and $T12 << \sqrt{LC1}$ with sufficient accuracy to satisfy the requirements of practical implementation.

## V. Technical Solution of Proposed Power Unit Management System

The previous section describes the proposed power unit principle of operation - energy flow. This device controls power transmission with EH system, supercapacitor and rechargeable battery or supercapacitor, its principle schematic is depicted in Fig. 5. Please note that some parts that are not substantial in terms of function are not mentioned in the figure.

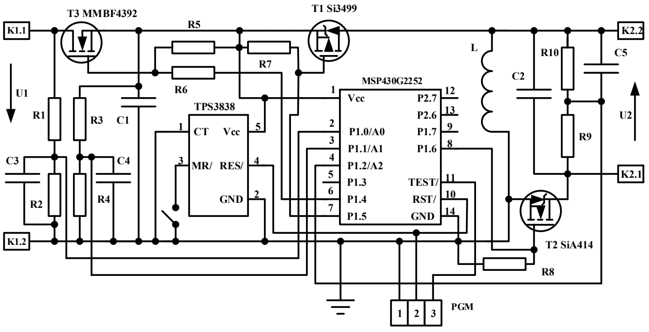


Fig. 5. Energy management system schematic

Output from EH system (photovoltaic cell, thermo generator, vibration generator) is connected to connector $K1$, which has voltage $U1$. The supplied device - a sensor node - is connected to connector $K2$. The energy is stored in supercapacitor $C1$, alternatively in rechargeable battery. Control unit, which is based on MCU MSP430G2252 enables by energy flow between primary supercapacitor $C1$ and secondary supercapacitor $C2$ in both directions by properly switched transistors $T1$ and $T2$. The control unit (the MCU) is able to measure:

- Voltage on generator output with no load or with load according to state $T3$ (input $A0$),
- Voltage $U1$ on supercapacitor $C1$ (input $A1$),
- Voltage $U2$ on supercapacitor $C2$ (input $A2$).

Based on measured voltages and character of EH energy processor, the energy can be moved between energy storages $C1$ and $C2$ to obtain maximal efficiency of the whole system and to operate at maximal point of efficiency. It is clear that in such simple electronic circuit solution (Fig. 5) the system is able to work correctly only if voltage on supercapacitor $C1$ is in range $2.5V - 3.6V$ (supply voltage of the MCU) that imposes relatively serious restriction. Though, this problem is able to be solved quite easily - control MCU power supply is separated, alternatively has its own supercapacitor with charge pump from $C2$.

The Fig. 6 shows voltage behavior at the output of a capacitor $C1$ ($1000uF/16V$). The start of the energy pumping is colored white, the voltage at the gate of the upper transistor is colored in cyan and the voltage at lower transistor is colored in purple. The upper transistor is a P-channel so that it is closed at high voltage and open at low. Lower transistor is N-channel so it opens at high voltage and closes at low. Input source is a constant voltage of $3.6V$, which is able to deliver up to $500mA$ average current. The process of energy flow starts with opening of the upper transistor for $800us$. Immediately, lower transistor opens the same time. During testing, the interval between such "open-closes" was $9.6ms$ and the whole period took $10.4ms$.

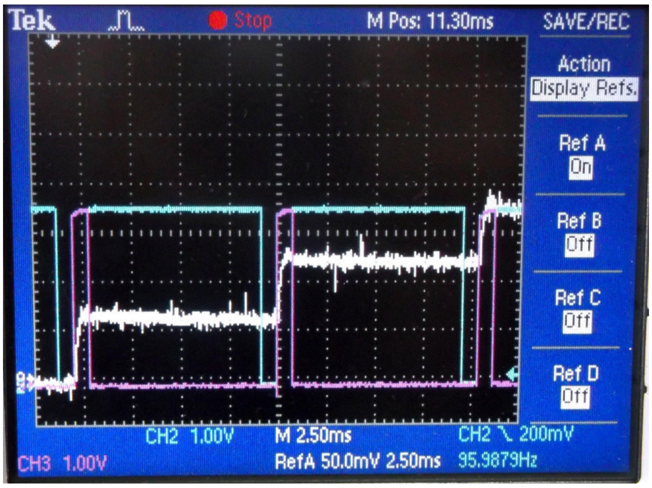


Fig. 6. Voltage behavior of the capacitor $C1$ on the testbed

The Fig. 7 shows a scenario with $C2$ capacitor and its alternating charges and discharges. Charging starts with opening of the upper output transistor for $800us$ following with opening of the lower transistor for $8us$. After a short pause $4.1ms$ inverse energy flow takes place ($C2$ to $C1$).

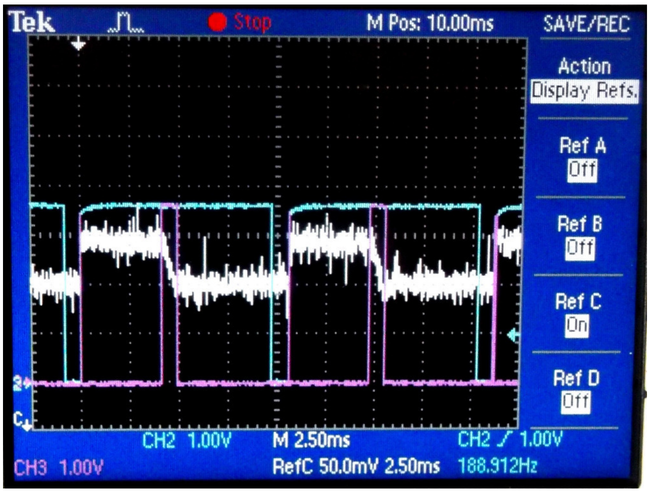


Fig. 7. Voltage behavior at $C2$ with alternating charges and discharges

The whole cycle takes $10.412ms$. The figure shows steady state of the oscillation and its lower voltage level is $360mV$ and its top level reaches $406mV$.

It is possible to achieve interesting solutions by connecting this systems from Fig. 2 in series. Alternatively, modifying the mentioned scheme for various EH systems that load power to a joint primary supercapacitor $C1$. In the latter case the voltage value has to be kept in order to ensure each system to operate in the point of maximal power. As an example, we can speak about more solar panels with various characteristics, alternatively with different exposure, whose energy is stored in a single accumulator cell.

## VI. Case Study

In this section we propose one simple example of the proposed system utilization. Let's imagine the proposed device to be used for energy harvesting purposes from a photovoltaic (PV) module. PV generators have two basic drawbacks - i) conversion efficiency of the incident radiation to electric energy is low (10 - 20%), especially at low light intensities; ii) and the amount of the obtained energy depends on the weather. Other disadvantage of solar cells is nonlinear V-I characteristic that is changing with the radiation intensity and the temperature. At particular, having the value of light exposure and the temperature on V-I characteristic, it is possible to find just one point, in which PV system works with maximal efficiency. This point is called Maximum Power Point (MPP). For the illustration, the V-I characteristic of a solar module is depicted in Fig. 8.

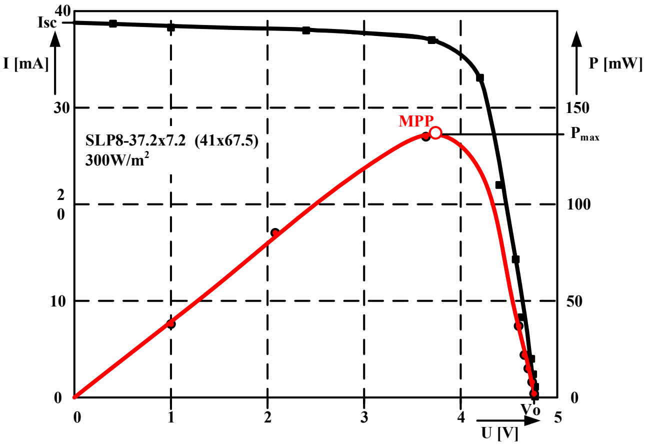


Fig. 8. V-I characterisitcs of solar module

It is clear that during active photovoltaic EH system operation there are various outer conditions. The intensity and angle of incident radiation changes as well as the temperature of solar cells, etc. These conditions considerably influence the parameters of the solar panel as well as the electric energy generator. The aim of the user is a system capable of utilize/accumulate solar energy at the MPP, while changing conditions apply. For this purpose, there are techniques called Maximum Power Point Tracking (MPPT) ensuring that in changes of working conditions, the photovoltaic system (PVS) is able to monitor Maximum Power Point (MPP) and then to ensure the effective operation of the whole system. Problem of designing and evaluation of MPPT methods was dedicated in many papers in specialized journals [22], [23], [24]. At present, there is a large number of used methods. In Table **??** there are most often used methods listed and at the same time there is listed its effectiveness as percentage of theoretical reachable value to obtained energy. It is obvious, that listed values are dependent on character of changes in outer conditions, so the table gives average values obtained in 14 different modes. Further information can be found in [22]. Note, that several methods was in [22] tested with various modifications, so Table **??** summarizes the best achieved results in that cases.

TABLE III
MPPT MEHODS AND THEIR EFFECTIVENESS

| **MPPT technique** | **Effectiveness [%]** |
|---|---|
| Constant Voltage Method (CV) | 79.5 |
| Short Current Pulse Method (SC) | 90.7 |
| Open Voltage Method (OV) | 94.6 |
| Perturb and Observe Method (P&O) | 99.3 |
| Incremental Conductance Method (IC) | 99.5 |
| Temperature Method (TP) | 97.1 |

For the implementation purposes into the proposed system, OV method is used. This method is easy to implement without other demands on the technical equipment. Average effectiveness of solar system with use of this method is sufficient enough. Note, that it is possible to use also other of the above methods, but implementation of MPPT techniques is not the matter of this contribution.

Algorithm of PV system control is the following:

1) Switch off the $T3$.
2) Delay $10\mu$s
3) Voltage with no load measurement $U1$.
4) Switch on $T3$.
5) Voltage calculation $U_p = 0.76 * U1$.
6) Voltage measurement $U_{C1}$.
7) Comparison with $U_p$, if $\mathrm{U_{C1} > U_p} + \varepsilon$ is true, then $\mathtt{pumpU_{C1}}()$ if $U_{C1} < U_p$ is true, then $\mathtt{pumpU_{C2}}()$
8) Delay $100ms$
9) Go to 1

$\mathtt{pumpU_{C1}}()$ is a function that pumps the energy from supercapacitor $C1$ to $C2$ unless the condition is true $\mathrm{U_{C1} > U_p} + \varepsilon$ (opens and closes $T1$ for the time according to (1) and opens and closes $T2$ for the time in accordance with (2)), then measures $U_{C1}$ and compares it with $\mathrm{U_p} + \varepsilon$ and if the condition is met then ends, otherwise repeats the whole "main loop".

$\mathtt{pumpU_{C2}}()$ is a function that pumps the energy from supercapacitor $C2$ to $C1$ unless the condition is true $\mathrm{U_{C1} > U_p}$ (opens and closes $T2$ for the time in accordance with (5) and opens and closes $T1$ for the time in accordance with (6), then measures $U_{C1}$ and compares it with $U_p$ and if the condition is met then ends, otherwise repeats the whole "main loop".

$\varepsilon$ is the constant, which value depends on values of supercapacitor $C1$, it constrains oscillation and often pumping of the energy.

In described algorithm, as the source of the energy was used the solar module SLP8-37,2x7,2 with dimensions 41x67,5 mm. Solar module consists of 8 solar cells that are connected in series. In good working conditions it can generate voltage $U1$ (with no load) that exceeds allowed operating voltage of the MCU. For this reason it is necessary to ensure that $\mathrm{U_{C1} < 3.6V}$.

The described example demonstrates applicability of the proposed control unit in EH WSN applications with utilization of solar energy. It is obvious that effectiveness of the solution, as the proportion of obtained energy to theoretical maximum is dependent on specific conditions. Because self-energy consumption of the control unit is constant, it does not depend upon amount of generated energy by the solar module. But this effectiveness also depends upon the size of the solar module and the light exposure.

## VII. Conclusion

The control unit of EH system that was developed at our department is designed for controlling the power unit with low power. That fact predetermines it for use in WSN applications, where we expect, that individual nodes are supplied through EH systems and contain energy storages based on supercapacitors, alternatively rechargeable batteries. Control unit, thanks to its versatility and simple parametrization with use of programmable resources, can find its enforcement in all cases, where the realization of effective energetic management with low cost is needed. Despite the fact, that experiments were made only with solar module, it is possible to use the proposed system also in many another applications, where it is needed to implement the bi-directional flow of electric energy.

## Acknowledgment

This contribution/publication is the result of the project implementation *Centre of excellence for systems and services of intelligent transport I.* ITMS 26220120028 and *Research Centre of University of Žilina* ITMS 26220220183 supported by the Research & Development Operational Programme funded by the ERDF.

Európska únia
Európsky fond regionálneho rozvoja

Operačný program
VÝSKUM a VÝVOJ

Agentúra
Ministerstva školstva, vedy, výskumu a športu SR
pre štrukturálne fondy EÚ

"Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ"

## References

[1] S. Abdelwahab, et. al. "Enabling smart cloud services through remote sensing: An internet of everything enabler", Internet of Things Journal, IEEE 1.3 (2014): 276–288. DOI: 10.1109/JIOT.2014.2325071

[2] J. Bradley, J. Barbier and D. Handler, "Embracing the Internet of Everything to Capture Your Share of $14.4 Trillion," White Paper, Cisco, 2013.

[3] J. Greenough, "The Internet of Everything," www.businessinsider/internet-of-everising-2015-bi-2014-12?op=1

[4] M. Chovanec, M. Hodon and L. Cechovic, "Tiny low-power WSN node for the vehicle detection", Informatica : an international journal of omputing and informatics. ISSN 0350-5596. Vol. 38, no. 3 (2014), pp. 223–227.

[5] J. Micek and M. Kochlan, "Energy-efficient communication systems of wireless sensor networks", Studia informatica universalis. ISSN 1621-7545. Vol. 11, no. 1 (2013), pp. 69–86.

[6] J. M.Gilbert and F. Baluochi, "Comparison of Energy Harvesting Systems for Wireless Sensor Networks", International Journal of Automation and Computing, October, 2008. DOI: 10.1007/s11633-008-0334-2

[7] B. Al-Ghamdi, M. Ayaida and H. Fouchal, "A dynamic slot scheduling for wireless sensors networks", 2014 IEEE Global Communications Conference, GLOBECOM 2014, art. no. 7036834, pp. 357–361. DOI: 10.1109/GLOCOM.2014.7036834

[8] W. K. G. Seah, et. al., "Wireless sensor networks powered by ambient energy harvesting (WSN-HEAP) - Survey and challenges", Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference, pp.1–5, 17-20 May 2009. DOI: 10.1109/WIRELESSVITAE.2009.5172411

[9] S. Grady, "Powering Wearable Technology and Internet of Everything Devices", Cymbet Corporation, www.cymbed.com, 2014.

[10] H. Fouchal, O. Zytoune and D. Aboutajdne, "A battery recovery aware routing protocol for Wireless Sensor Networks", International Symposium on Computers and Communications, Workshops, art. no. 6912636, DOI: 10.1109/ISCC.2014.6912636

[11] P. Sevcik and O. Kovar, "Power unit based on supercapacitors and solar cell module", SCIECONF 2013 : the 1st international virtual scientific conference, 10.-14. June. ISSN 1339-3561. 2013. ISBN 978-80-554-0726-5. pp. 468–471.

[12] T. Bernard, et. al., "Impact of routing protocols on packet retransmission over wireless networks", IEEE International Conference on Communications, art. no. 6654996, pp. 2979–2983. DOI: 10.1109/ICC.2013.6654996

[13] O. Karpis, "Wireless sensor networks in intelligent transportation systems", International journal of modern engineering research (IJMER 2012), ISSN 2249-6645. Vol. 3, iss. 2 (2013).

[14] J. Micek and O. Karpis, "Wireless sensor networks - design of smart sensor node", ICMT'11 : proceedings of the international conference on military technologies 2011. Brno, Czech Republic, 10.-11. May 2011. ISBN 978-80-7231-787-5. pp. 1109–1116.

[15] P. Dutta, et al., "A Building Block Approach to Sensornet Systems," In Proc. SenSys 08, North Carolina, USA, 2008.

[16] Z. Feng, et al., "Collaborative signal and information processing: an information-directed approach", Proceedings of the IEEE , vol.91, no.8, pp.1199–1209, Aug. 2003. DOI: 10.1109/JPROC.2003.814921

[17] O. Karpis, "FFT on ARM-based low-power microcontrollers", International journal of engineering research and development (IJERD). - ISSN 2278-800X. - Vol. 6, no. 9 (2013), pp. 22–26.

[18] M. Kochlan and P. Sevcik, "Supercapacitor power unit for an event-driven wireless sensor node", Computer Science and Information Systems (FedCSIS), 2012 Federated Conference, pp.791–796, 9.-12. Sept. 2012

[19] S. G. Burrow and P. D. Mitcheson, "Power Conditioning for Energy HarvestingâĂŞTheory and Architecture", Micro Energy Harvesting 5 (2015).

[20] EnerChip EP Energy Processor, www.cymbed.com/pdfs/DS-7215.pdf

[21] P. Sevcik and O. Kovar, "Alternative energy sources for WSN node power supply", ITS 2013 - Intelligent transportation systems 2013. August 26-30, 2013. ISSN 1339-4118. ISBN 978-80-554-0763-0. pp. 146–149.

[22] R. Faranda and S. Leva, "Energy comparison of MPPT techniques for PV Systems," WSEAS Transactions on Power Systems, 6. 2008, ISSN 1790-5060.

[23] D.Freeman, "Introduction to Photovoltaic Systems Maximum Power Point Tracking," Application Report, Texas Instrument, November 2010.

[24] N.Femia, G. Petrone, G. Spagnuolo and M. Vitelli, "Optimization of Pertrub and Observe Maximum Power Point Tracking Method," IEEE Transactions on Power Electronics, vol.20 No.4, July 2005.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1247–1255
DOI: 10.15439/2015F241
ACSIS, Vol. 5

# Effects of Temperature and Humidity on Radio Signal Strength in Outdoor Wireless Sensor Networks

Jari Luomala and Ismo Hakala
University of Jyvaskyla, Kokkola University Consortium Chydenius
P.O. Box 567, FI-67701 Kokkola, Finland
Email: {jari.luomala, ismo.hakala}@chydenius.fi

***Abstract*—Many wireless sensor networks operating outdoors are exposed to changing weather conditions, which may cause severe degradation in system performance. Therefore, it is essential to explore the factors affecting radio link quality in order to mitigate their impact and to adapt to varying conditions. In this paper, we study the effects of temperature and humidity on radio signal strength in outdoor wireless sensor networks. Experimental measurements were performed using Atmel ZigBit 2.4GHz wireless modules, both in summer and wintertime. We employed all the radio channels specified by IEEE 802.15.4 for 2.4GHz ISM frequency band with two transmit power levels. The results show that changes in weather conditions affect received signal strength. Of the studied weather variables, variation in signal strength can be best explained by the variation in temperature. We also show that frequency diversity can reduce the effects of channel-specific variation, and the difference between the transmit power levels.**

## I. Introduction

MANY wireless sensor networks (WSNs) and their applications are in use outdoors exposed to changing environmental conditions. Weather conditions particularly can have a significant impact on the performance of WSNs and therefore cannot be ignored [1], [2], [3]. While the location of nodes may be fixed and their surroundings might remain almost static, the weather will not remain stable. The ambient temperature and humidity will change and fluctuate temporally, having both diurnal and seasonal variation. In addition, there can be spatial variation in weather, which affects WSN due to microclimates. While changes in weather conditions are inevitable and may have significant effects, they are usually measurable and could be mitigated based on experimental measurements. Hence, it is essential to explore weather-related factors affecting radio link quality in order to mitigate their impact and to adapt to varying conditions.

The effects of weather conditions on link quality (e.g., signal strength) in WSNs have been explored in quite a few studies (e.g., [2], [4], [3], [1], [5], [6], [7], [8], [9]). However, no clear consensus has been achieved so far. Some studies report that temperature is the dominating factor affecting signal strength while others claim that humidity is the main reason. Some suggest also other reasons. Furthermore, research methods, radios and platforms employed and a number of explored weather variables vary between studies, occasionally resulting in contradictory results and conclusions. Hence, there definitively seems to be a need for further studies.

This paper sets out to find out the effects of temperature and humidity on radio signal strength in outdoor WSNs. Experimental measurements were carried out using Atmel ZigBit 2.4GHz wireless modules [10] with AT86RF230 radios [11] in a university campus area during December 2013 and July 2014. Unlike most previous studies, we employed all the 16 radio channels specified by IEEE 802.15.4 for 2.4GHz ISM frequency band (channels $11-26$) using two different transmit power levels. For the purpose of measuring local weather conditions, each node was integrated with a sensor (SHT75) [12] to measure the ambient temperature and relative humidity. To begin with, we show the temporal variation of signal strength. To find out the role of temperature and humidity on this variation, we study how signal strength correlates with temperature, relative humidity, and absolute humidity. Furthermore, we apply linear regression to explore the magnitude of these effects. We also highlight the differences between channels, the utility of frequency diversity, and the difference between the transmit power levels used. Our findings may be useful for designing algorithms and protocols which are adaptive and robust against the effects of weather. In particular, RSSI-based ranging and localization could benefit from these results.

In summary, our study has the following main contributions:

- We show that changes in weather conditions affect radio signal strength. Temperature seems to be the best explanatory variable for signal strength variation and has a negative, linear effect on signal strength in general, while high relative humidity may have some effect, particularly when temperature is below 0°C.
- We show that correlation between signal strength and the studied weather variables vary depending on radio channel and link. Applying frequency diversity will alleviate these effects.
- We show that smaller transmit power results in smaller unexplained variation in received signal strength (in most cases) and thus stronger correlation with the studied weather variables.

The rest of this paper is organized as follows. In Section II,

978-83-60810-66-8/$25.00©2015, IEEE

some related studies in the field are shortly presented. In Section III, we briefly take a closer look at the central variables used in this paper. Experimental measurements performed are described in Section IV. Thereafter, experimental results are presented and analyzed in Section V, followed by a short discussion in Section VI. Finally, we conclude our work in Section VII.

## II. Related Works

Temperature has been the main focus in many recent studies dealing with the effects of weather conditions on link quality. For example, Bannister et al. [1] found a linear decrease of $8\mathrm{dB}$ in signal strength when temperature rose from $25°\mathrm{C}$ to $65°\mathrm{C}$ when they used TI CC2420 radio on a Tmote Sky node, both in their outdoor experiment in the Sonoran Desert and in the lab experiment. They also showed the implications this has for communication range, network connectivity, multi-hop data collection, and RSS-based localization. Based on both outdoor and indoor experiments, Boano et al. [5] showed that the increase in temperature decreases both RSSI (Received Signal Strength Indicator) and LQI (Link Quality Indicator). In the outdoor experiment in a wheat field in Govone, Italy, Boano et al. used Tmote Sky nodes. In their indoor experiment, they used both Tmote Sky (CC2420 radio) and MSB430 nodes (CC1020 radio). They also found that the noise floor readings of both platforms decrease with the increase in temperature. Based on their over 10-day deployment of TelosB nodes in a forest garden, Luo et al. [7] showed that both temperature and relative humidity correlate with RSSI. However, based on linear regression, only the effect of temperature on RSSI can be regarded as relatively significant. Wennerström et al. [8] in their half-year experiment near Uppsala, Sweden, showed how variations in meteorological conditions affect IEEE 802.15.4 link performance when using TelosB nodes with CC2420 radio. Particularly, they studied how variations in PRR (Packet Reception Ratio) and RSSI correlate with temperature, absolute humidity, precipitation, and sunlight. Their results show that PRR and RSSI correlate mostly with temperature, while the correlation with other factors is not so clear. They also observed both diurnal and seasonal variation in PRR. In their recent work, Boano et al. [9] studied the impact of temperature on various WSN platforms and radios (CC2420, CC2520) and showed the different effects of temperature on transmitter and receiver nodes. They also showed that the relation between temperature and RSSI is similar with different platforms and can be approximated as a linear function when using platform-specific parameters. Also, Lin et al. [13] found a temporal variation of RSSI (with MICAz nodes) during their 3-day outdoor experiment, in which different transmit power levels were used, but they did not analyze the cause.

Humidity-related issues have also received a lot of attention in the research community. Anastasi et al. [2] found that the transmission range of mica2 sensor nodes (using RFM ChipCon radios) decreases substantially during rain or fog. Contrary to this, based on their measurements conducted in a potato field with the help of Mica2Dot nodes equipped with Chipcon CC1000 radios, Thelen et al. [4] showed that radio waves propagate better under high humidity conditions, in the presence of rain and at night for example. They attributed this positive impact to changes in the reflection coefficient of the top of the canopy of the potato field. They showed that RSSI values were positively correlated with RH and negatively with temperature, but they focused only on humidity in their analysis. Capsuto and Frolik [3] demonstrated how rain and snowfall, freezing rain and fog, and humidity can significantly affect RSSI, causing large fades or even complete loss of connectivity. They used Chipcon CC2420DK nodes at three different frequencies of 2.4GHz ISM band in the experiment. Boano et al. [5] also showed that the effect of both thin and thick fog, and rainfall on RSSI is almost negligible. However, the impact of a very heavy rainfall on wireless communication may be significant. Markham et al. [6] conducted a 26-day measurement in the forest of Wytham Woods, near Oxford and a lab experiment, using T-mote Sky nodes with CC2420 radios. They showed that variations in RSSI are due to the presence of water on a node's casing rather than fading caused by rain. They justified their finding by stating that water is capacitively loading the antenna, thereby changing its radiation pattern.

In contrast to previous studies, our experiment was conducted with AT86RF230 radio transceivers. Further, we utilized all the specified 16 radio channels for 2.4GHz ISM frequency band with two different transmit power levels and used sensors to measure the ambient temperature and relative humidity in each node.

## III. Background

In the following, the related variables used to measure radio signal strength, temperature and humidity are defined.

### A. Received Signal Strength Indicator, RSSI

The well-known basic metrics used to link quality estimation in WSNs are RSSI (Received Signal Strength Indicator), PRR (Packet Reception Ratio), SNR (Signal to Noise Ratio), and LQI (Link Quality Indicator) [14]. Received Signal Strength Indicator (RSSI) is a standard feature built in most radio transceivers typically employed in WSN nodes and indicates the received radio signal power in a particular radio channel. As specified in the IEEE 802.15.4 standard [15], RSSI is computed by averaging RSSI values over 8 symbol periods ($128\mu\mathrm{s}$). The IEEE 802.15.4 standard refers to this (average) RSSI as Energy Detection (ED), as also does the AT86RF230 radio transceiver [11].

The ED value of the AT86RF230 radio used [11] is computed by averaging RSSI values over 8 symbol periods ($128\mu\mathrm{s}$). In AT86RF230, RSSI is a 5-bit register value with $3\mathrm{dB}$ resolution, and ED value is an 8-bit register value with $1\mathrm{dB}$ resolution. The ED value has $84\mathrm{dB}$ range and absolute accuracy of $\pm5\mathrm{dB}$. The RF input power in AT86RF230 can be computed as follows:

$$\mathrm{P_{RF}} = \mathrm{RSSI\_BASE\_VAL} + (\mathrm{ED\_LEVEL} - 1)\ \ [\mathrm{dBm}],$$

where $\mathrm{RSSI\_BASE\_VAL} = -91\mathrm{dBm}$ (RSSI sensitivity) and $\mathrm{ED\_LEVEL} = 1..84$. The minimum ED value $(\mathrm{ED\_LEVEL} = 0)$ indicates receiver power less than $\mathrm{RSSI\_BASE\_VAL}$. The formula used to express RSSI as RF input power [dBm] is radio-specific and usually can be found in the data sheet of the particular manufacturer.

Throughout this paper, the terms RSSI and signal strength will be used to refer to RF input power $\mathrm{P_{RF}}$ $[\mathrm{dBm}]$, unless otherwise stated.

### *B. Relation between temperature and humidity*

The relationship between different weather variables, such as temperature and humidity, can be quite confusing. As it is well known, temperature and humidity are connected to each other either directly or indirectly. For the purpose of analyzing the results, it is of the essence to know their mutual dependence. In the following, we try to emphasize the differences between various humidity definitions and how they are related to temperature and to each other. For clarification, their relationship is also illustrated in Fig. 1.

Humidity is the amount of water vapor, the gaseous state of water, in the air, and is usually invisible. The maximum amount of water vapor in the air depends on air temperature. *Absolute humidity (AH)* is the water content in the air, i.e., the mass of water vapor included in a particular volume of air, expressed in $\mathrm{g/m^3}$. *Saturated humidity (SH)* is the maximum amount of water vapor in the air at particular temperature (the blue line in Fig. 1). *Relative humidity (RH)* defines, in a percent, how much water vapor (AH) is in the air relative to the maximum amount of water vapor (SH) at the same temperature and pressure. Relative humidity of saturated air is $100\%$. *Dew point (temperature)* is the temperature to which air must be cooled down in order that water vapor starts to condense into liquid water or ice $(RH = 100\%)$. [16]

Of particular interest is the point wherein water vapor is changing from gaseous state into liquid (condensation) or solid (deposition, $\mathrm{T} < 0°\mathrm{C}$) state. Condensation/deposition starts when humidity increases or temperature falls, reaching the saturation point $(RH = 100\%)$. The condensed water vapor is called either dew (frost when $\mathrm{T} < 0°\mathrm{C}$) or fog (or clouds), depending on whether formed on a solid surface or in the air. [16]

Absolute humidity $AH$ $(\mathrm{g/m^3})$ can be defined, e.g., as a function of temperature and relative humidity as [17]:

$$AH(t, RH) = 216.7 \cdot \left[ \frac{\frac{RH}{100\%} \cdot A \cdot \exp\left(\frac{m \cdot t}{T_n + t}\right)}{273.15 + t} \right], \qquad (1)$$

where $t$ is the actual temperature $(°\mathrm{C})$, $RH$ the actual relative humidity $(\%)$, $m = 17.62$, $T_n = 243.12°\mathrm{C}$, and $A = 6.112\mathrm{hPa}$.

## IV. Experiment Overview

In the experiment, we set up a wireless sensor network operating in 2.4GHz ISM frequency band in a university campus area in western Finland. The WSN measured and collected data related to weather and radio link quality, both in summer and wintertime. This raw data was sent to a server to be further processed and analyzed. In the following, the experiment is described in more detail.

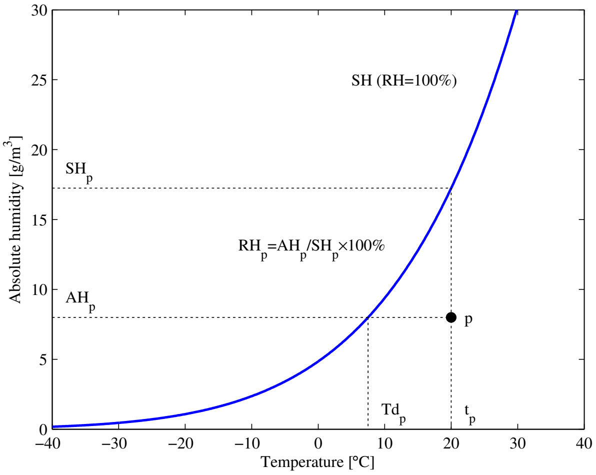


Fig. 1. Relationship between the following weather variables: temperature (t), relative humidity (RH), absolute humidity (AH), saturated humidity (SH), and dew point (Td) (at constant barometric pressure).

### *A. WSN Configuration and Deployment*

We used a wireless sensor network operating in 2.4GHz ISM frequency band in our experiment. The WSN consisted of Atmel ZigBit 2.4GHz wireless modules (ATZB-24-B0) [10] containing Atmel's ATmega1281V microcontroller [18] and AT86RF230 RF transceiver [11]. Furthermore, the sensor nodes were integrated with a Sensirion's humidity and temperature sensor (SHT75) [12] to measure the ambient temperature and relative humidity. The sensor nodes were powered with two AA-size $3.6\mathrm{V}$ primary lithium batteries. The nodes were enclosed with a weatherproof plastic casing, leaving the external antennas and SHT75 sensors outside the casing. A drain valve was added into the bottom of the casing to remove possible moisture or water. The sink node was similar to sensor nodes, without an SHT75 sensor and protective casing, and connected wired to a Raspberry PI. The Raspberry PI had LAN connection for sending raw data from the sink to the server/database.

The equipment used in the experiment consisted of five sensor nodes, one sink node and one Raspberry PI. In addition, there was one server/database. The sensor nodes were attached to five lamp posts with the help of mounting racks. There was one 2.4GHz node in a single rack. The nodes were at the height of approximately $3\mathrm{m}$ (top of the antenna). The sink node and Raspberry PI were inside the university campus, and they were powered by mains current. The network setup can be seen in Fig. 2.

### *B. Data Collection and Processing*

The sink node broadcasted a link-measurement packet twice in every minute using two different transmit power levels,

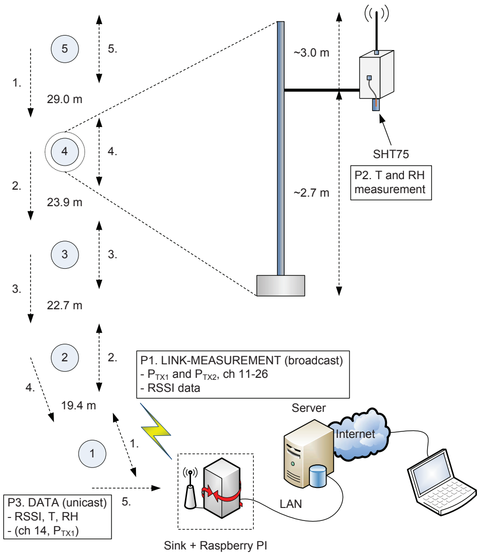


Fig. 2. Measurement network setup (not in scale).

first the maximum transmit power of $+3.0\mathrm{dBm}$ ($\mathrm{P_{TX1}}$) and thereafter $-7.2\mathrm{dBm}$ ($\mathrm{P_{TX2}}$). Radio channel was changed every minute, thus all the 16 channels $(11-26)$ were rotated in 15 minutes. Sensor nodes receiving the link-measurement packets updated their neighbortables for the particular links (RSSI, etc.), and forwarded the packets in their scheduled time frames. After the link-measurement phase, the nodes measured temperature (T) and relative humidity (RH) readings by using the SHT75 sensor. Thereafter, the sensor nodes sent (unicast) neighbortable and sensor data (T, RH) to the sink by using predefined static routing (see Fig. 2). The sink forwarded the collected raw data to Raspberry PI, which in turn sent it to the server/database via LAN. Temperature sensor readings (ADC) were converted to temperature values according to [12] and calibrated with each other by using offset values, before saving to the database. Also, RH sensor readings (ADC) were linearized and temperature-compensated according to [12] before saving to the database.

The collected raw data was downloaded from the server to be further processed and analyzed. We used MATLAB [19] for processing, analyzing and presenting data. Both RSSI data (of each channel) and weather data (T, RH) were averaged over one hour. There were 60 RSSI samples in an hour for each link, i.e., on average $3-4$ samples (for both TX power) for each radio channel. Weather data was measured once in a minute; thus there were 60 samples in an hour for each node. Absolute humidity was also computed; it was based on average temperature and relative humidity values for each node, according to (1).

To mitigate the effects of changes in environmental conditions, effects such as multipath fading, on different radio frequencies, we utilized frequency diversity and computed link RSSI by averaging the RSSI samples collected at different radio channels. The $1\mathrm{h}$ average RSSI for each link $k$, $\overline{RSSI}^{k}_{1\mathrm{h}}$, was computed as

$$\overline{RSSI}^{k}_{1\mathrm{h}} = \frac{1}{n}\sum_{i=1}^{n} RSSI^{k,i}_{1\mathrm{h}} \text{ [dBm]}, \tag{2}$$

where $i = 1..16$ (channels $11-26$ specified by IEEE 802.15.4 for 2.4GHz ISM frequency band). Bardella et al. [20] have shown that exploiting frequency diversity will mitigate the multipath fading effects, which could help us in analyzing the effects of weather conditions.

To calculate statistics (TABLE I), we used the $1\mathrm{h}$ average RSSI change of the analyzed links, $\overline{\Delta RSSI}_{1\mathrm{h}}$, as follows:

$$\overline{\Delta RSSI}_{1\mathrm{h}} = \frac{1}{8}\sum_{k=1}^{8} \underbrace{\overline{RSSI}^{k}_{1\mathrm{h}} - \overline{RSSI}^{k}}_{=\Delta RSSI^{k}_{1\mathrm{h}}} \text{ [dB]}, \tag{3}$$

where $k$ is the link number, $\overline{RSSI}^{k}$ the average RSSI for link $k$ over the measurement period, and $\Delta RSSI^{k}_{1\mathrm{h}}$ the $1\mathrm{h}$ RSSI change for link $k$. The links chosen to be analyzed were the ones between the closest neighboring nodes, i.e., the links $1 \leftarrow 2$, $2 \leftarrow 1$, $2 \leftarrow 3$, $3 \leftarrow 2$, $3 \leftarrow 4$, $4 \leftarrow 3$, $4 \leftarrow 5$, $5 \leftarrow 4$.

As for weather data, we also used $1\mathrm{h}$ average temperature and relative humidity of the analyzed links. Absolute humidity was computed based on these average values, applying (1).

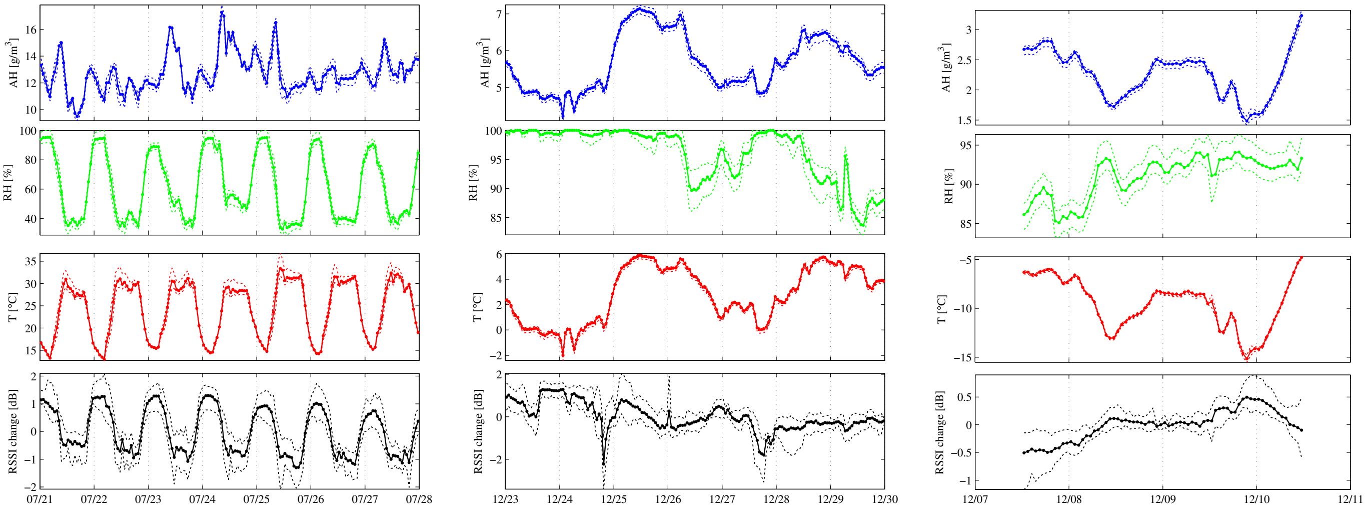
## V. Experimental Results

We performed experiments and gathered data by using our WSN in different seasons. Data from three different periods was chosen, one set in summer and two in winter, to be analyzed here. The representative periods are 1 week in July 2014 (temperature $> 0°\mathrm{C}$), 1 week in December 2013 (temperature $<> 0°\mathrm{C}$), and 3 days in December 2013 (temperature $< 0°\mathrm{C}$). Our aim is to find out how temperature and humidity affect radio signal strength.

### A. *Temporal Variation of Signal Strength*

When exploring the results from our experiments in both summer and winter, it is evident that signal strength has both short-term (diurnal) and long-term (seasonal/weekly) variation. Interestingly, the variation is notably different in each period. Diurnal variation (day/night) is clearly apparent in summer, whereas seasonal variation is easier to detect in winter and between different seasons (summer/winter). These variations are not random but mainly cyclic, following a certain distinct pattern. For example, in summer, as can be seen in Fig. 3 (a), signal strength falls in the daytime and rises in the nighttime.

Comparing the variation of signal strength with the variation of weather variables in Fig. 3, it is easy to find similarity between them, especially in summer. This suggests that there is a relation between signal strength and particular weather variables. However, the relation in summer seems to be different from that in winter. For example, in summer there is hardly any relation between absolute humidity and signal strength, but in winter below $0°\mathrm{C}$ the correlation is quite clear. In contrast, when temperature fluctuates near $0°\mathrm{C}$, the relation between weather variables and signal strength is unclear.



(a) July 2014 (above 0°C).

(b) December 2013 (around 0°C).

(c) December 2013 (below 0°C).

Fig. 3. RSSI change, temperature (T), relative humidity (RH), and absolute humidity (AH) of the analyzed links (mean, min, max) during three different periods ($\mathrm{P_{TX}} = 3.0\mathrm{dBm}$). RH data of node 3 is excluded from all computations for December 2013 (around 0°C) due to humidity sensor malfunction.

The behavior of individual links is quite similar, as shown in Fig. 3. This indicates that RSSI variation is mainly caused by changes in weather affecting all the links, rather than by some site-specific reason. However, between individual channels there can be quite large variations in RSSI behavior, both within and between individual links. Probably this results from other factors, such as multipath propagation. By averaging the RSSI samples of different radio channels, we can thus smooth the random and channel-specific variation to better discover the effects of weather conditions.

Nevertheless, our observations confirm the findings reported in the literature (e.g., [1], [8]), which states that signal strength has temporal variation. Probably these variations are mostly due to changes in weather conditions. In the following, we will focus on temperature and humidity to find out their effects on RSSI variation.

### B. *Effects of Temperature on Signal Strength*

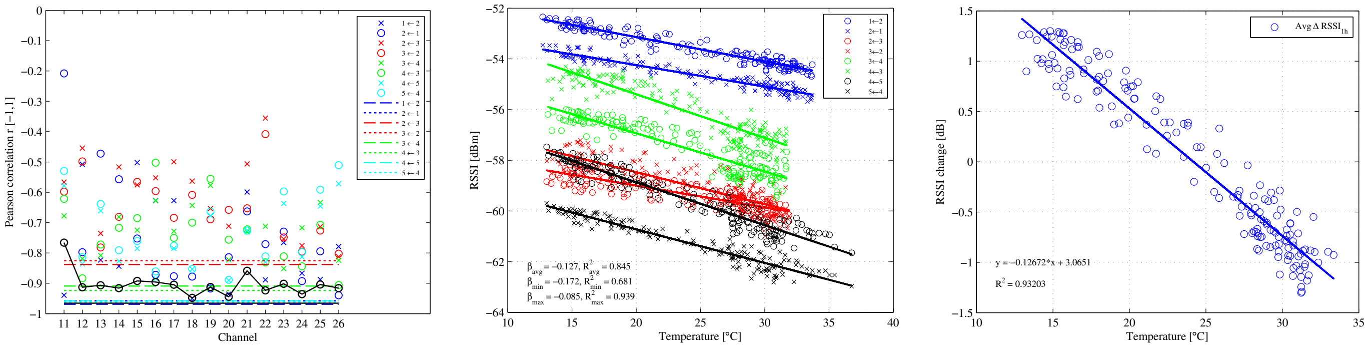
As could already be seen in Fig. 3, there is an obvious relationship between temperature and signal strength. In general, when temperature rises signal strength (RSSI) falls, and vice versa. This indicates negative correlation (dependence) between temperature and signal strength. To get a better understanding of the matter, we computed Pearson correlation coefficient to measure the degree of linear dependence between temperature and RSSI. The results from July are presented in Fig. 4 (a). As can be seen, the correlation varies depending on channel and link. Some link-channel combinations have a very strong correlation, while in some others the correlation is less significant. The difference between the correlation of average RSSI change with temperature (straight black line) and the individual link-channel correlations with temperature is quite clear, indicating the benefits of exploiting frequency diversity. The correlation of the average RSSI change (also channel-specific) is strong, confirming the hypothesis that RSSI correlates negatively with temperature.

The correlation of average RSSI change with temperature is high also while below 0°C in December. However, when temperature fluctuates near 0°C, there is a substantial degradation in correlation. Nevertheless, the negative correlation still holds. The correlation between temperature and RSSI is statistically significant ($p < 0.001$) in each measurement period.

When comparing how RSSI correlates with temperature with two different transmit power levels, we can find some differences. On average, RSSI correlation with temperature is slightly stronger (negatively) when using smaller transmit power ($\mathrm{P_{TX}} = -7.2\mathrm{dBm}$) compared to when using the maximum transmit power ($\mathrm{P_{TX}} = 3.0\mathrm{dBm}$). This holds true for all the three periods, being emphasized in winter when the deviation is bigger.

To quantify the effect of temperature on RSSI, we plotted both RSSI of individual links and average RSSI change versus temperature and applied simple linear regression, where we used temperature as an explanatory variable for RSSI variation. The results from July are presented in Fig. 4 (b) and 4 (c). As can be seen, temperature has quite a considerable effect on signal strength. A linear, negative trend can be observed for all the links, but there is some variation regarding magnitude of the impact (regression coefficient). As for RSSI change, regression coefficient is $-0.127$, i.e., the rise of temperature by $10°\mathrm{C}$ decreases RSSI approximately by $1.3\mathrm{dB}$. The coefficient of determination ($\mathrm{R}^2$) is very high ($0.93$), which implies that in this model RSSI variation can be explained to a high degree by the variation in temperature.

There are no great differences between the different periods regarding the magnitude of the impact. The regression coefficient varies between $-0.09$ and $-0.13$. However, $\mathrm{R}^2$ decreases during frost, and practically plunges while temperature is



(a) (b) (c)

Fig. 4. Relationship between RSSI/RSSI change and temperature (T) during one week in July 2014 ($P_{TX} = 3.0$dBm). (a) Pearson correlation between RSSI/RSSI change and T, (b) Linear regression of RSSI on T (all links), (c) Linear regression of average RSSI change on T.

around 0°C. This indicates that when temperature fluctuates around 0°C, there are some other factors causing large deviations, thus reducing correlation and $R^2$. The regression coefficients are statistically significant ($p < 0.001$) in each measurement period.

As for the magnitude of the effect, the differences between two power levels are not significant. $R^2$ values are somewhat higher with the smaller transmit power ($P_{TX2}$), indicating that the smaller transmit power results in smaller unexplained RSSI variation.

The summary of the relationship between temperature and RSSI is presented in TABLE I.

### *C. Effects of Humidity on Signal Strength*

Another potential factor affecting signal strength is humidity. As can be seen in Fig. 3, there is also a clear relation between both relative (RH) and absolute (AH) humidity and signal strength in particular times. Relative humidity rises and falls together with RSSI (above 0°C), indicating positive correlation, while the trend of absolute humidity indicates negative correlation (below 0°C). As previously, we computed the Pearson correlation coefficients and simple linear regression for both RH and AH. The summary of the relationship between RH/AH and RSSI is presented in TABLE I. The correlation and regression coefficients of RH to be discussed below are statistically significant ($p < 0.01$) in each measurement period, while those of AH are statistically significant only for winter periods.

In July, RH has a very high positive correlation (0.95) with RSSI while AH and RSSI are uncorrelated. $R^2$ in the regression model is also very high for RH, which means that in this model RSSI variation could be explained to a high degree by the variation in RH. As for RSSI change, the regression coefficient of RH is about 0.03, which means that the rise of RH by 10% increases RSSI approximately by 0.3dB.

In December, at the temperature below 0°C, both RH and AH have high, almost equally strong but opposite correlation with RSSI. Interestingly, the regression coefficients of both RH and AH are about $2 - 3$ times higher compared to the other period(s). Also $R^2$ is quite significant ($\approx 0.6$) for both RH and AH, although for RH it is smaller when compared to July.

In the near 0°C period, correlation is quite low for both RH and AH. However, it is still consistent with the other period(s). Due to decreased correlation, it is obvious that $R^2$ in the regression model is also low for both RH and AH. Therefore, the explanatory powers of the linear regression models are not sufficient.

On average, RSSI correlation with AH is slightly stronger (negatively) with smaller transmit power in both winter periods, as are the regression coefficients and $R^2$ values. In the case of RH, the differences between the power levels are relatively minor.

### *D. Temperature vs. Humidity*

While correlation is a good predictor of a potential causal relationship, it does not imply causation and could be caused by some other factor. The high correlation between the studied weather variables and RSSI both in July (T, RH) and below zero (T, RH, AH) could be partly explained by the high mutual dependence of temperature and humidity, as illustrated in Fig. 5. As can be seen, RH correlates strongly with temperature in July, while AH does so in December. This close relationship between the studied weather variables complicated our attempts to distinguish between the actual impact of temperature and humidity on RSSI. Therefore, we applied multiple linear regression with two explanatory variables to find out both the combined effect and the effect of a particular weather variable while the other variable is taken into account. As previously, to analyze the effects of temperature and humidity on RSSI, we used the average RSSI change, T, RH and AH of the analyzed links.

As expected, some of the results show high collinearity between temperature and humidity. In July, temperature and RH are highly collinear (*variance inflation factor*, *VIF* is high), while in December below 0°C, temperature and AH are collinear. Therefore, it is questionable to use them together in the regression model. Contrary to this, temperature and AH are not collinear (VIF = 1.01) in July, nor are temperature and RH (VIF = 1.38) in December when temperature is below 0°C.

TABLE I
AVERAGE RSSI CHANGE VS. TEMPERATURE (T) / RELATIVE HUMIDITY (RH) / ABSOLUTE HUMIDITY (AH).

| Statistical significance: | **RSSI vs. T** | | | **RSSI vs. RH** | | | **RSSI vs. AH** | | |
|---|---|---|---|---|---|---|---|---|---|
| $^{*}$ $p < 0.01$ | July | December | December | July | December | December | July | December | December |
| $^{**}$ $p < 0.001$ | $(T > 0°C)$ | $(T <> 0°C)$ | $(T < 0°C)$ | $(T > 0°C)$ | $(T <> 0°C)$ | $(T < 0°C)$ | $(T > 0°C)$ | $(T <> 0°C)$ | $(T < 0°C)$ |
| otherwise none | $(n = 168)$ | $(n = 168)$ | $(n = 72)$ | $(n = 168)$ | $(n = 168)$ | $(n = 72)$ | $(n = 168)$ | $(n = 168)$ | $(n = 72)$ |
| **Pearson correlation ($r$)** | | | | | | | | | |
| $P_{TX1} = +3.0dBm$ | -0.965$^{**}$ | -0.336$^{**}$ | -0.851$^{**}$ | 0.946$^{**}$ | 0.229$^{*}$ | 0.776$^{**}$ | -0.022 | -0.264$^{**}$ | -0.779$^{**}$ |
| $P_{TX2} = -7.2dBm$ | -0.973$^{**}$ | -0.446$^{**}$ | -0.891$^{**}$ | 0.948$^{**}$ | 0.316$^{**}$ | 0.768$^{**}$ | -0.050 | -0.353$^{**}$ | -0.825$^{**}$ |
| **Regression coef.** | | | | | | | | | |
| $P_{TX1} = +3.0dBm$ | -0.127$^{**}$ | -0.101$^{**}$ | -0.090$^{**}$ | 0.035$^{**}$ | 0.032$^{*}$ | 0.081$^{**}$ | -0.013 | -0.227$^{**}$ | -0.539$^{**}$ |
| $P_{TX2} = -7.2dBm$ | -0.113$^{**}$ | -0.121$^{**}$ | -0.115$^{**}$ | 0.031$^{**}$ | 0.040$^{**}$ | 0.098$^{**}$ | -0.026 | -0.276$^{**}$ | -0.701$^{**}$ |
| $\mathbf{R^2}$ | | | | | | | | | |
| $P_{TX1} = +3.0dBm$ | 0.932 | 0.113 | 0.724 | 0.895 | 0.053 | 0.602 | 0.000 | 0.069 | 0.607 |
| $P_{TX2} = -7.2dBm$ | 0.947 | 0.199 | 0.795 | 0.899 | 0.100 | 0.590 | 0.002 | 0.124 | 0.680 |

The results from July show that temperature is the dominating factor affecting RSSI. Using AH as the other explanatory variable together with temperature in the model does not improve the adjusted $R^2$ ($\bar{R}^2$) in practice. Further, the weight of temperature in the model is significantly higher than that of AH. Therefore, it is sufficient to include only temperature in the regression model.

The situation is different in December below 0°C. While temperature is the most significant variable ($\bar{R}^2 = 0.72$ for $P_{TX1}$), using of RH as the other explanatory variable does improve the $\bar{R}^2$ to a relatively large extent ($\bar{R}^2 = 0.87$). This means that both temperature and RH seem to have effect on RSSI. However, the proportion of temperature in RSSI variation is somewhat higher than that of RH. It is thus reasonable to consider the inclusion of both temperature and RH in the regression model.

When temperature is around 0°C in December, there are other factors apart from temperature or humidity (although probably related to them) causing sudden RSSI variations, as can be seen in Fig. 3. RSSI can experience large variations when temperature fluctuates near 0°C and RH is close to 100%, resulting in low $R^2$ for all the variables. RSSI variation thus cannot be explained with the help of any linear model in this case.

In conclusion, temperature generally seems to be the most significant variable affecting RSSI and could be used in a linear model to explain RSSI change, except in the above-mentioned special case. Relative humidity may have some effect on RSSI, particularly in high humidity conditions below 0°C, where it may be useful to include both temperature and RH in the model.

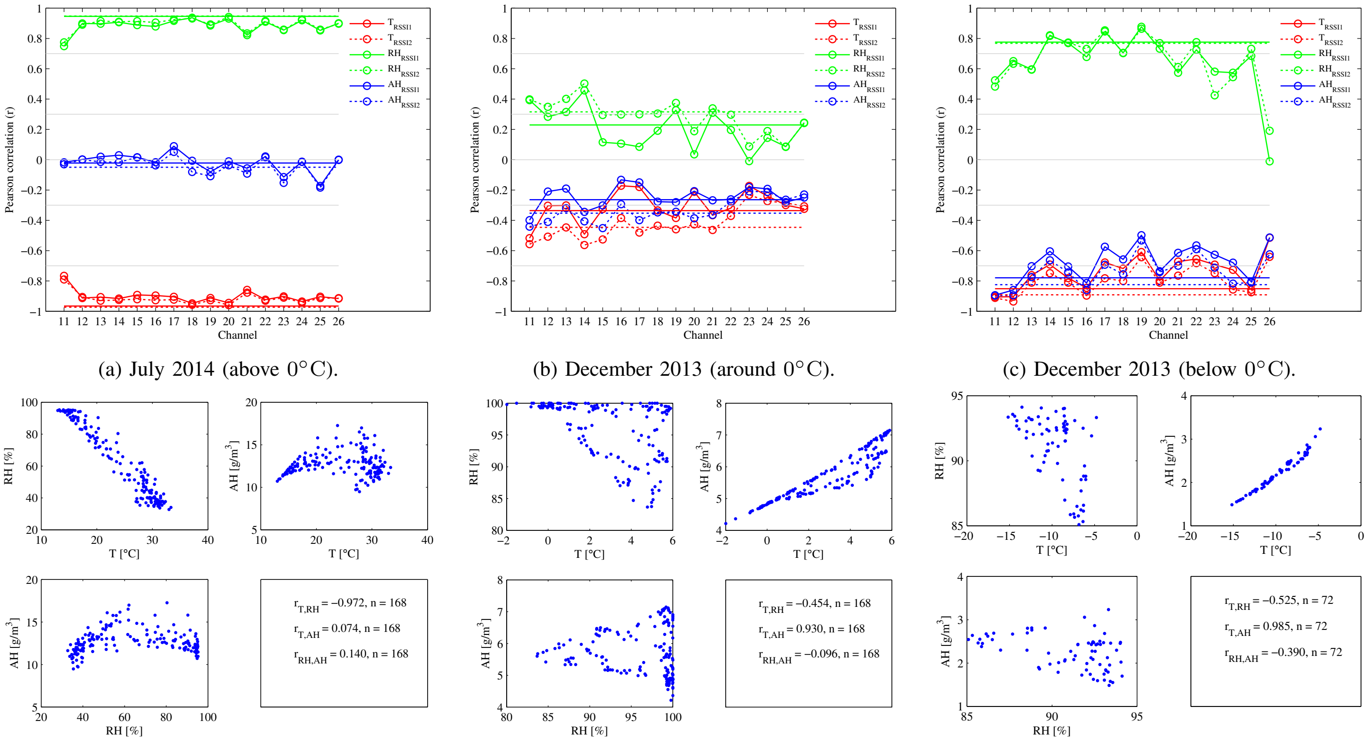
## VI. DISCUSSION

Our findings confirm the previous results in the literature of the effects of temperature on link quality. Moreover, some RF transceiver manufacturers mention the temperature dependency in their data sheets [21], [22], [11], thus supporting the results. It has been reported that both output power and receiver sensitivity [21], as well as crystal frequency accuracy (drift) and characteristics of the VCO (Voltage Controlled Oscillator) [21], [22] vary with temperature.

As it was shown, the effect of humidity is more complicated. Additionally, the accuracy of SHT75 sensor decreases substantially during high humidity conditions (RH $> 90\%$). It decreases linearly from typical $\pm 1.8\%$RH to $\pm 4.0\%$RH when RH increases from $90\%$ to $100\%$ [12]. Furthermore, [12] states that long term exposure to conditions outside the sensor's normal operating range may temporally offset the RH signal ($+3\%$RH after 60h). During both our winter periods, relative humidity is high and mostly over $90\%$. According to [12], RH can drop drastically due to heavy condensation of water on the sensor surface, which was also observed in our experiment. The foregoing aspects may affect the accuracy of the results in winter.

Particularly problematic is the case when near-zero temperature is combined with high humidity (RH close to $100\%$). RSSI fluctuation in that period could be due to ice/snow on top of the nodes and antennas melting into liquid water and/or liquid water (due to condensation or rain) freezing into ice. The same kinds of effects of water/ice on link quality are reported, e.g., in [6] and [3]. This indicates that water in liquid or solid state on top of the nodes or antennas may cause unpredictable changes in signal strength. Therefore, temperature and humidity may have indirect effects on RSSI variation through condensation (RH $\approx 100\%$) and freezing of water or melting of ice/snow (T $\approx$ 0°C).

Regardless, the effects of temperature and humidity on signal strength can have severe implications on different sensor network protocols. Particularly, the accuracy of RSSI-based ranging and localization decreases significantly if the effects of temperature and humidity are ignored. Therefore, temperature and humidity conditions should be taken into account in RSSI-based ranging in order to adapt to prevailing weather conditions. Our results could be used to compensate RSSI variation caused by temperature and humidity, and thus to improve ranging accuracy. Based on the experiment, it could be advisable to use frequency diversity in RSSI-based ranging and localization instead of a single channel. Further, the effects of temperature and humidity have implications on many other



(a) July 2014 (above 0°C).

(b) December 2013 (around 0°C).

(c) December 2013 (below 0°C).

(d) July 2014 (above 0°C).

(e) December 2013 (around 0°C).

(f) December 2013 (below 0°C).

Fig. 5. Correlation between average RSSI change and weather variables (T, RH, AH) during three different periods (a-c). Correlation between temperature (T), relative humidity (RH), and absolute humidity (AH) during three different periods (d-f).

protocols besides localization. These include protocols related to network connectivity and management, routing, etc.

Weather conditions are not the only reason for RSSI variation. Also HW-related issues and other environmental conditions, such as changes in surroundings (especially in winter), interference, etc. may affect signal strength. The effect of RSSI resolution on the results is also unknown. Furthermore, the measurement period we used was relatively sparse. By using a more dense measurement period as well as applying some filtering method to RSSI readings, it could be possible to achieve more accurate results.

In our future studies, we intend to carry out more experiments and lab measurements to study the effects of temperature and humidity on radio link quality in a controlled environment.

## VII. Conclusion

In this paper, the effects of ambient temperature and humidity on radio signal strength of Atmel ZigBit 2.4GHz wireless modules in outdoor WSNs were explored. Experimental results show that changes in weather conditions affect received signal strength. Temperature seems to have a significant negative influence on signal strength in general, while high relative humidity may have some effect on it, particularly below 0°C. Further, it was shown that use of frequency diversity can reduce the effects of channel-specific variation, and the difference between the transmit power levels used. Our findings could be useful when designing adaptive, robust algorithms and protocols, such as those related to RSSI-based localization.

## References

[1] K. Bannister, G. Giorgetti, and S. Gupta, "Wireless Sensor Networking for "Hot" Applications: Effects of Temperature on Signal Strength, Data Collection and Localization," in *The Fifth Workshop on Embedded Networked Sensors (HotEmNets'08)*, Charlottesville, Virginia, USA, June 2008.

[2] G. Anastasi, A. Falchi, A. Passarella, M. Conti, and E. Gregori, "Performance Measurements of Motes Sensor Networks," in *7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM'04)*, Venice, Italy, October 2004, pp. 174–181, http://dx.doi.org/10.1145/1023663.1023695.

[3] B. Capsuto and J. Frolik, "A System to Monitor Signal Fade Due to Weather Phenomena for Outdoor Sensor Systems," in *Fifth International Conference on Information Processing in Sensor Networks (IPSN 2006)*, Nashville, TN, USA, April 2006, Demo Abstract.

[4] J. Thelen, D. Goense, and K. Langendoen, "Radio Wave Propagation in Potato Fields," in *1st Workshop on Wireless Network Measurements (WiNMee 2005)*, Riva del Garda, Trentino, Italy, April 2005.

[5] C. Boano, J. Brown, Z. He, U. Roedig, and T. Voigt, "Low-Power Radio Communication in Industrial Outdoor Deployments: The Impact of Weather Conditions and ATEX-Compliance," in *Sensor Applications, Experimentation, and Logistics: First International Conference, SENSAPPEAL 2009*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, N. Komninos, Ed., vol. 29. Athens, Greece: Springer, September 2009, pp. 159–176, revised Selected Papers, http://dx.doi.org/10.1007/978-3-642-11870-8_11.

[6] A. Markham, N. Trigoni, and S. Ellwood, "Effect of Rainfall on Link Quality in an Outdoor Forest Deployment," in *International Conference on Wireless Information Networks and Systems (WINSYS 2010)*, Athens, Greece, July 2010, pp. 1–6.

[7] J. Luo, X. Xu, and Q. Zhang, “Understanding Link Feature of Wireless Sensor Networks in Outdoor Space: a Measurement Study,” in *IEEE Global Telecommunications Conference (GLOBECOM 2011)*, Houston, TX, USA, December 2011, http://dx.doi.org/10.1109/GLOCOM.2011.6134117.

[8] H. Wennerström, F. Hermans, O. Rensfelt, C. Rohner, and L. Nordén, “A Long-Term Study of Correlations between Meteorological Conditions and 802.15.4 Link Performance,” in *2013 IEEE International Conference on Sensing, Communications and Networking (SECON)*, New Orleans, LA, USA, June 2013, pp. 221–229, http://dx.doi.org/10.1109/SAHCN.2013.6644981.

[9] C. Boano *et al.*, “Hot Packets: A Systematic Evaluation of the Effect of Temperature on Low Power Wireless Transceivers,” in *5th Extreme Conference on Communication (ExtremeCom’13)*, Thorsmork, Iceland, August 2013.

[10] Atmel, “ZigBit$^{\text{TM}}$2.4 GHz Wireless Modules - ATZB-24-A2/B0 Datasheet,” 2009, http://www.atmel.com.

[11] ——, “Low Power 2.4 GHz Transceiver for ZigBee, IEEE 802.15.4, 6LoWPAN, RF4CE and ISM Applications - AT86RF230 Datasheet,” 2009, http://www.atmel.com.

[12] Sensirion, “Datasheet SHT7x (SHT71, SHT75) - Humidity and Temperature Sensor IC,” 2011, Version 5, http://www.sensirion.com.

[13] S. Lin *et al.*, “ATPC: Adaptive Transmission Power Control for Wireless Sensor Networks,” in *The 4th ACM Conference on Embedded Networked Sensor Systems (ACM SenSys 2006)*, Boulder, Colorado, USA, November 2006, pp. 223–236, http://dx.doi.org/10.1145/1182807.1182830.

[14] N. Baccour *et al.*, “Radio Link Quality Estimation in Wireless Sensor Networks: A Survey,” *ACM Transactions on Sensor Networks (TOSN)*, vol. 8, no. 4, September 2012, article 34, http://dx.doi.org/10.1145/2240116.2240123.

[15] IEEE Computer Society, *IEEE Std 802.15.4-2003*, IEEE, New York, NY, USA, 2003.

[16] Finnish Meteorological Institute, http://ilmatieteenlaitos.fi/.

[17] Sensirion, “Humidity at a Glance - Most Relevant Equations with Sample Code,” 2008, Application Note (Version 1.0), http://www.sensirion.com.

[18] Atmel, “Atmel 8-bit AVR Microcontroller with 64K/128K/256K Bytes In-System Programmable Flash,” 2006.

[19] The MathWorks, Inc., “MATLAB - The Language of Technical Computing,” http://www.mathworks.se/products/matlab/.

[20] A. Bardella, N. Bui, A. Zanella, and M. Zorzi, “An Experimental Study on IEEE 802.15.4 Multichannel Transmission to Improve RSSI-Based Service Performance,” in *4th International Workshop on Real-World Wireless Sensor Networks (REALWSN 2010)*, ser. LNCS 6511, Colombo, Sri Lanka, December 2010, pp. 154–161, http://dx.doi.org/10.1007/978-3-642-17520-6_15.

[21] Texas Instruments Inc., “CC2400 - 2.4 GHz Low-Power RF Transceiver,” 2008, Data sheet.

[22] ——, “CC2420 - 2.4 GHz IEEE 802.15.4 / ZigBee-ready RF Transceiver,” 2014, Data sheet.

DOI: 10.15439/2015F332
ACSIS, Vol. 5

# Monitoring of $CO_2$ Amount in Closed Objects via WSN

Róbert Žalman, Veronika Olešnaníková, Peter Ševčík and Peter Šarafín
University of Žilina
Faculty of Management Science and Informatics,
Univerzitná 8215/1 Žilina 010 26,
Email: {Robert.Zalman, Veronika.Olesnanikova, Peter.Sevcik, Peter.Sarafin}@fri.uniza.sk

***Abstract*—Global warming is big issue of this time. It is caused by producing emissions, mainly carbon dioxide. Many organizations tries to establish restrictions to limit $CO_2$ emissions. Our aim is to monitor underground parking lots and detect level of carbon dioxide using wireless sensor network. Gained results are drawn in the map of pollution of monitored area.**

***Index Terms*—WSN, $CO_2$ monitoring, global warming**

## I. Introduction

CARBON dioxide is a chemical compound, formed by two atoms of oxide and one atom of carbon connected by two double bonds. Carbon dioxide is a natural part of chemical compounds and atmosphere in certain amount, which is produced by fauna and flora. After the industrialization, carbon dioxide concentration in the air dramatically arose. This fact was caused by gas and oil combustion.

Recently, $CO_2$ concentration is at such level that population health is at risk. Some animal species can even face extinction because of climatic changes. Carbon dioxide is one of the most typical greenhouse gases and it causes that the heat reflected by Earth surface is kept in the atmosphere. This phenomenon results in increase of Earth temperature, which affects our ecosystem.

According to Kyoto protocol, 141 countries agreed to lower their emissions of greenhouse gases, including carbon dioxide. The aim is to reach 5.2% lower level than concentration of greenhouse gases in the 1990. To control $CO_2$ level in underground parking lot, its monitoring has to be performed. Nowadays there is a lot of underground parking lots in cities, which are not sufficiently air ventilated. For this purpose wireless sensor network (WSN) should serve [1]. This paper deals with $CO_2$ detection at the bases of its level monitoring by WSN.

## II. Consequences of $CO_2$ accumulation

In the garages, monitoring of air pollution is provided by measuring the $CO_2$ level using appropriate sensors which are deployed in whole area, based on the zone division. It is important to deploy sensor all over the building because of different gas concentration in given zones.

Measured values fall in range from 0 to 2 000 ppm (particles per million) $CO_2$. Minimal carbon dioxide concentration is around 400 ppm which is adequate to clean air. $CO_2$ is well mixed in the atmosphere, so observations of concentrations from a single site are an adequate indicator of world trends for atmospheric $CO_2$ [2]. For better imagination, concentration of $CO_2$ is shown in relative values, where 0% is equivalent to outside air (cca 400 ppm), 100% is adequate to maximal acceptable concentration, usually 2 000 ppm. The air pollution may be divided into several levels. First level is up to 40% which means 800 + 400 ppm, second level is in range up to 65 % which is 13000 + 400 ppm and the third level can get values up to 85% (1700 + 400 ppm). Level of carbon dioxide has a significant influence to the people:

- 1 % concentration of carbon dioxide in the air can cause sleepiness,
- 2 % causes human to be slightly dopey, increases blood pressure and heartbeat and decreases hearing ability,
- 5 % usually stimulate breath center, causes dizziness and confusion and troubles with breathing accompanied by headache and anhelation. This concentration can also result in access of panic.

There are two approaches to detect $CO_2$ using sensors. The first one is measurement using method of wavelength absorption, which is one of the properties of chemical compound. This method is called NDIR (not dispersive infrared) [3]. The second method is based on changes of electrical charge of chemical reaction measuring. This reaction is a result of air contact ($CO_2$ particles) with particles in the sensor.

If we want to control and decrease the level of $CO_2$, at first we need to know its concentration in give area. The most used detection method is using NDIR sensors [4], but the price of sensors is quite high (between 100 and 1000 euro). For that reason we decided to use sensor with technology of measuring chemical reactions. We selected MQ-135 sensor, which provides monitoring of air quality and with suitable settings it can be used as sensor for $CO_2$ detection.

## III. Sensor description

The sensor is designed for the use in air quality control equipments for buildings / offices. It is suitable for detection of $NH_3$, $NO_x$, alcohol, Benzene, smoke and $CO_2$ [5]. The sensitivity characteristic of the MQ-135 sensor can be seen at Fig. 1. This characteristic is used for the conversion of sensors output to the related *ppm* characteristic for the gas under test.

As the graph shows, the $CO_2$ curve can be described by the equation in the form

978-83-60810-66-8/$25.00©2015, IEEE

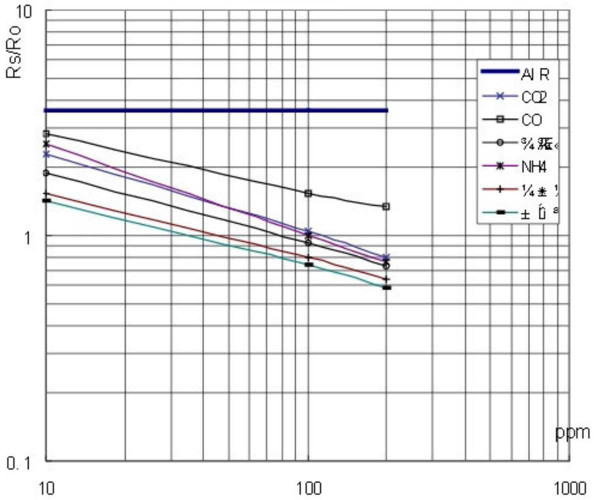


Fig. 1. Characteristics of MQ-135 sensor

$$y = a \cdot x^b, \tag{1}$$

where *y* corresponds to *ppm* and *x* corresponds to $R_s/R_o$.

Since the characteristic is defined graphically, we determine two points that can be described with the minimum deviation. Substituting these points and then solving the set of two equations

$$\begin{aligned} 200 &= a \cdot 0,8^b \\ 10 &= a \cdot 2,1^b, \end{aligned} \tag{2}$$

we obtain coefficients *a* a *b* which represent scaling factor and exponent respectively.

To obtain the calibration constant $R_o$, known current average value of $CO_2$ in the atmosphere and the equation 3 is used, where $R_s$ is the measured value, *a = 100.0482*, *b = -3.1041* and current *ppm = 401.52*.

$$R_o = R_s \cdot \sqrt[b]{\frac{a}{ppm}} \tag{3}$$

On this basis, we can conclude that the ppm value can be determined by the equation

$$ppm = 100.0482(R_s/R_o)^{-3.1041}. \tag{4}$$

Using the above calculation, it is possible to obtain values of $CO_2$ levels in the air of monitored environment. This theory was applied via a wireless network, where each nodes computational part is microcontroller SAM4S.

## IV. Description of printed circuit board - processing and communication

The basis of the board is a microcontroller ATSAM4S. This model contains a Cortex-M4 core and it is equipped with RISC architecture. It works at 120MHz frequency, it is equipped with 12-bit AD Converter with Programmable Gain Amplifier and its consumption is 180uA/1MHz-max [6]. The input of the AD converter is a signal obtained from the $CO_2$ sensor (module with the sensor MQ-135). RF module is also connected to the PCB. Communication between MCU and RFM70 module is provided via SPI communication interface. The sensor node is also equipped with an SD card slot with which the MCU communicates via HSMC interface. The archtecture of WSN node is described at Fig. 2. Sensor, processing and communicational modules are interconnected, so real $CO_2$ monitoring device can be seen at Fig. 3.

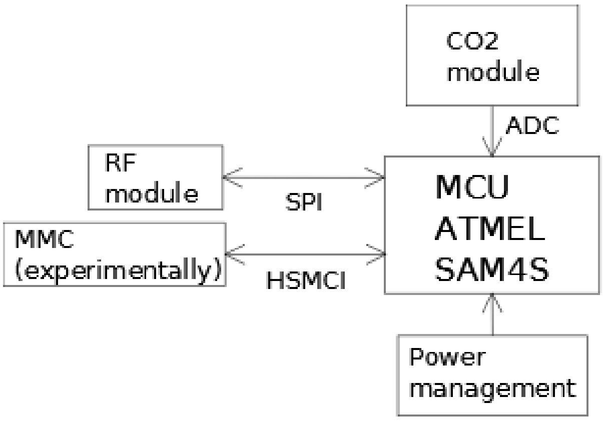


Fig. 2. Architecture of WSN node

### A. Collecting the results (processing)

To be able to determine the level of $CO_2$, it is important to apply mathematical relationship at the output of $CO_2$ sensor. MQ-135 changes the analog output voltage value with respect to the $CO_2$ level.

To define this value in digital domain, we use an integrated 12-bit AD converter. After the timer interrupt is set, method $ADCRead()$ is implemented. In this method, a single conversion is performed. When the end of the transfer is reached, the obtained data are stored to the variable and are ready for further processing.

To work with the proper amount of $CO_2$ level values, it is important to apply the normalization function (4), where we got a correction constant $R_o = 0.869565$, which represents $15\%$ more than measured value. The acquired data are then written to the microSD card (or sent through the RF module) and the MCU waits for the next interrupt from the timer.

WSN node functionality can be divided into several working modes:

- data gathering - sampling, converting signal from analog to digital form (A/D converter), processing, data storage
- data transmission - data encoding, data encapsulation, broadcast
- data receive - filtering, decoding, data extraction, error checking

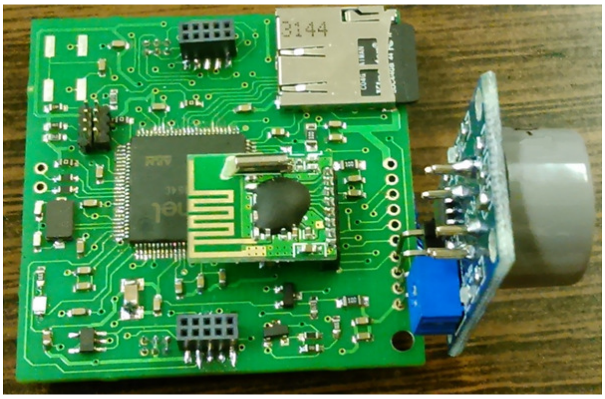


Fig. 3. $CO_2$ detection module

### B. Communication method

When we were creating the network topology, we had to take into consideration some basic information about WSN.

Network designs are based on three types of topologies:

- Bus Topology - This topology consists of a Backbone cable connecting all nodes on a network without intervening connectivity devices.
- Star Topology - In the computer networking world the most commonly used topology in local area networking is the star topology.
- Ring Topology - In ring network, each node is connected to the two nearest nodes so that the entire network forms a circle.

We chose a star topology, which we have expanded to the tree topology. This topology was enough for our purpose of use. Properties of tree topology are:

1) resistance of the network against downtime of the individual stations and lines
2) sensitive to downtime of nodes
3) easy expandability
4) two-point connections

Our next challenge was WSN synchronization. We were deciding between known synchronization methods which are:

- Reference Broadcast Synchronization (RBS), [7]
- Timing-sync Protocol for Sensor Network (TPSN), [8]
- Flooding Time Synchronization Protocol (FTSP), [9]
- Universal synchronization algorithm for wireless sensor networks - "FUSA algorithm", [10]

From these methods, we decided to use the algorithm FUSA, especially because it can be easily implemented and it is not suitable only for hierarchical networks, it is universal and scalable. After resolving these problems, we managed to submit data to the central node and ensure proper network functionality. In the central node, these data are saved on the microSD card. These data are then evaluated in the offline mode, so they can be processed and map of air pollution of the monitored area can be created.

An inexpensive wireless module that meets our requirements is used for communication between nodes. Wireless module RFM70 is a transmitter / receiver that operates at a frequency of 2.4GHz. The module has low power consumption, which is 23 mA when it is used as a transmitter and 18 mA as a receiver. It can be powered by supply from 1.9V to 3.6V, but its inputs withstand the voltage up to 5V. Consumption in standby mode of the module is only 50$\mu$A. The speed of wireless transmission of this module is 1Mbps or 2Mbps. The module supports 126 channels. It communicates with the microcontroller via SPI serial interface [11].

Individual nodes are synchronized by FUSA algorithm [10]. Nodes send data via RF module to the central node, where these data are saved on the SD card. These data are then evaluated in the offline mode, so they can be processed and map of air pollution of the monitored area can be created.

## V. The deployment of WSN and visualization of results

For testing purposes, WSN was deployed in the basement of the shopping centre [12]. Deployment of the nodes can be seen at Fig. 5, nodes are drawn as black bullets. There was an effort to place the sensors effectively for purposes of measurement. Some of the sensors are placed nearby the doors. Entrances and exits are marked in the figure. Other sensors were placed far from entrance / exit and in the corners where we expected the pollution to reach the highest values. One floor of the underground garage has been diagnosed for the period of one week. The collected data are graphically represented in the Fig. 4, where we can observe the dependence of emissions from the time for the sensor S1, so it is possible to monitor the time trend of the pollution.

Occurrence of emissions in the object during the night and early morning is insignificant (very small) [13]. The rate of pollution increases with active use of the parking lot [14]. This fact is the most visible in the early loading and afternoon emptying of these areas. We noted that the level of emissions for various days is very similar. Friday differs from other days significantly, because parking services are used by most drivers.

Fig. 5 shows the map of the environment with an adequate level of emissions at given time. To create such map, measured emissions from all the sensors that are valid for a given time should be collected. We expected the worst air in the corner areas but on the contrary the most polluted spots were nearby southern exit. We analysed the situation and found out that the cars entering and leaving the parking lot use more often this exit than the eastern exit. It means that on this way, there is often traffic jam while waiting for opening the barrier because each car has to stop and check the parking ticket in the machine and all cars have its engine turned on. Reason for choosing this exit is logistical - this entrance / exit has better connection towards the city centre, other shopping centres and also it is a direction to some apartment settlements. The corner parts are quite good ventilated, what outreached our expectation. That is thanks to effective ventilation system. Thus we can recommend to increase ventilation only in the area of southern exit.

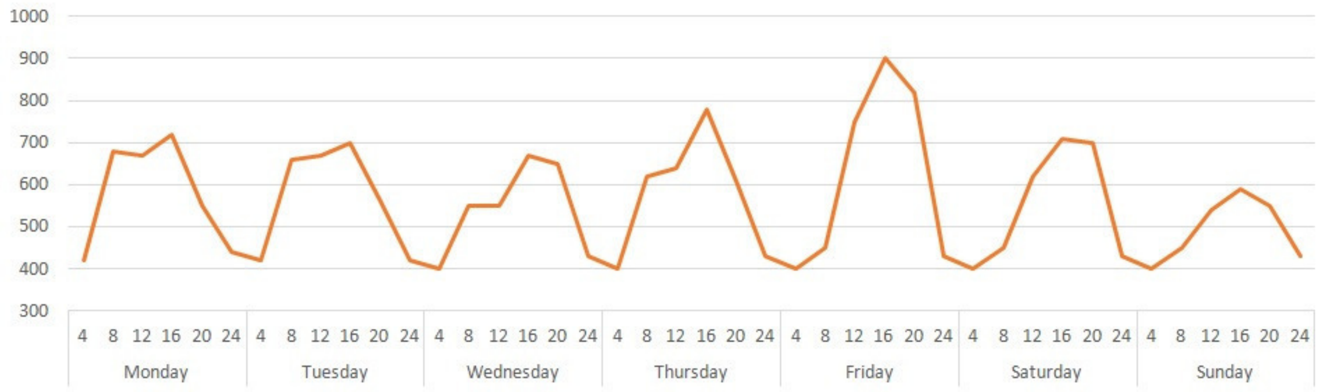


Fig. 4. $CO_2$ concentration recorded at sensor S1 during the week

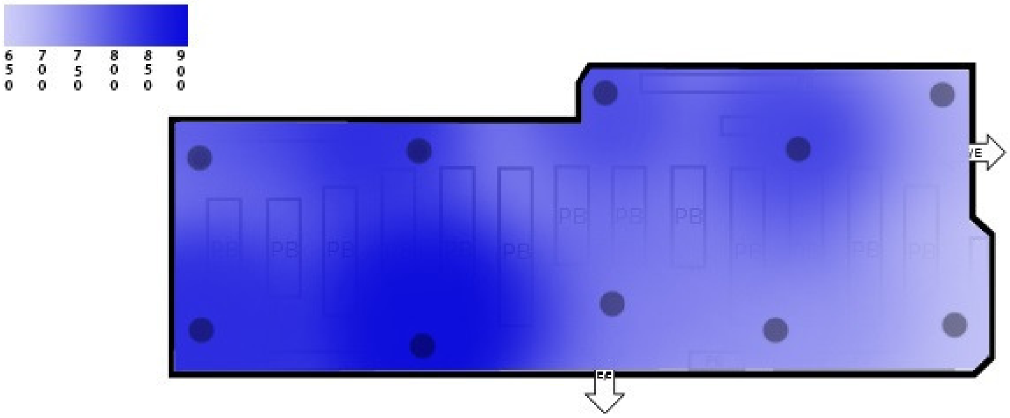


Fig. 5. Map of $CO_2$ concentration

## VI. Conclusion

Underground parking lot monitored during the testing week did not reach the first level of air pollution. This fact is caused by a good ventilation system (air conditioning) deployed in these areas. Nevertheless, it can be said that in the time of rapid usage of these spaces by drivers, the pollution level of parking lot has increased from an average to almost double.

System for the detection of $CO_2$ proposed and designed by us may be used not only in the underground parking lots but also in industrial zones. In the future, the system could be extended by the microphone and thus it might also serve for measuring acoustic emission eg. at road administration (control of the traffic flow, noise and $CO_2$ emissions).

## References

[1] Kapitulík J., Jurečka M., Miček J., Hodoň M., Wireless sensor network - value added subsystem of ITS communication platform, FedCSIS : September 7-10, 2014, Warsaw, Poland: IEEE. ISBN 978-83-60810-61-3. - p. 1017-1023.

[2] $CO_2$ level monitoring http://co2now.org/

[3] Garcia-Romeo D., Fuentes H., Medrano N., Calvo B., Martinez P. A., Azcona C., A NDIR-based $CO_2$ monitor system for wireless sensor networks, Circuits and Systems (LASCAS), 2012 IEEE Third Latin American Symposium on , vol., no., pp.1,4, Feb. 29 2012-March 2 2012

[4] Jonqwon Kwon, Gwanghoon Ahn, Gyusik Kim, Chun Kim, Hiesik Kim, A study on NDIR-based $CO_2$ Sensor to apply Remote Air Quality Monitoring System, ICROS-SICE International Joint Conference 2009

[5] MQ-135 datasheet, https://www.futurlec.com/Datasheet/Sensor/MQ-135.pdf

[6] Microcontrollers Atmel SAM4S datasheet, http://www.atmel.com/Images/Atmel_11100_32-bit-Cortex-M4-Microcontroller_SAM4S_Datasheet.pdf

[7] Jeremy E., Lewis G. and Deborah E., Fine-grained network time synchronization using reference broadcasts, in Fifth Symposium on Operating Systems Design and Implementation OSDI, 2002.

[8] Ganeriwal S. ,Ram K. and Srivastava M. B. , Timing-sync protocol for sensor networks, in First ACM Conference on Embedded Networked Sensor Systems 2003.

[9] Roche M., January 2011[online], Time Synchronization in Wireless Networks http://www.cse.wustl.edu/~jain/cse574-06/time_sync.htm

[10] Chovanec M., Púchyová J., Húdik M., Kochláň M., Universal synchronization algorithm for wireless sensor networks - "FUSA algorithm", FedCSIS, 2014, Warsaw, Poland: IEEE. - ISSN 2300-5963

[11] RFM70 data transciever module datasheet, http://www.futurlec.com/RFM70.shtml

[12] Spachos P., Liang S., Hatzinakos D., Gas leak detection and localization system through Wireless Sensor Networks, Consumer Communications and Networking Conference (CCNC), 2014 IEEE 11th , vol., no., pp.1130,1131, 10-13 Jan. 2014

[13] Karpiš O., Juríček J., Miček J., Application of wireless sensor networks for road monitoring, 10th IFAC workshop on programmable devices and embedded systems, October 6th - 7th, 2010, p. 207-212

[14] Zhang H., Liang Y., Zhou Q., Fan H., Dai J., A self-adaptive greenhouse $CO_2$ concentration monitoring system based on ZigBee, Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on , vol.03, no., pp.1137,1140, Oct. 30 2012-Nov. 1 2012

DOI: 10.15439/2015F322
ACSIS, Vol. 5

# Measuring the performance and energy consumption of AES in wireless sensor networks

Cristina Panait
Faculty of Automatic Control and Computers
University POLITEHNICA of Bucharest
Email: cristina.panait19@gmail.com

Dan Dragomir
Faculty of Automatic Control and Computers
University POLITEHNICA of Bucharest
Email: dan.dragomir@cs.pub.ro

***Abstract*—With WSN deployments increasing in popularity, securing those deployments becomes a necessity. This can be achieved by encrypting inter-node communications and/or using message authentication codes. AES is a well studied symmetric cipher, with no known practical vulnerabilities, that can be used to solve both problems. We provide an optimized implementation of AES, with four modes of operation (ECB, CBC, CFB and CTR), that uses the hardware accelerator available on the ATmega128RFA1 microcontroller, and compare it with the best known software implementation. We show that our hardware AES implementation is both faster and more energy efficient than a software implementation. This is not the case for previous sensor nodes and implementations, which show an improved execution speed, but with a higher energy consumption.**

## I. Introduction

As a general definition, a wireless sensor network is composed of a set of nodes which communicate through a wireless medium in order to perform certain tasks. A couple of examples where WSNs can be deployed, as stated in [1], are: fire extension detection, earthquake detection, environment surveillance for pollution tracking, intelligent building management, access restriction, detection of free spaces in parking lots and so on. Advantages brought by WSNs are enhanced flexibility and mobility, mainly because nodes are generally powered from an on-board battery and thus do not depend on their surroundings. This, however, is also their biggest weakness. The lifetime expectancy of a node depends on its usage. The constraints mainly come from the limited energy source, as data processing and transmission can be energy intensive.

The particular characteristics of these types of networks make the direct implementation of conventional security mechanisms difficult. The imposed limitations on minimizing data processing and storage space and reducing bandwidth need to be addressed. The major constraints for WSNs, as presented in [2], [3] and [4], are: energy consumption (which can lead to premature exhaustion of the energy source and to the denial of service), memory limitations (flash, where the application source code is stored, and RAM, where sensed data and intermediary computing results are stored), unreliable communication (the routing protocols used, collisions), latency (which can lead to synchronization issues and algorithms that cannot act correctly) and unattended nodes (an attacker could have physical access to the nodes).

The concept behind WSNs and their applications presents an increased risk to a series of attacks which can affect the network's functionality. In this paper we analyze algorithms that provide confidentiality for WSNs. We focus our analysis on AES-128, as it is a well studied cipher with no known practical vulnerabilities, has a speed comparable with other symmetrical encryption algorithms and is supported on multiple WSN platforms through a hardware acceleration module.

In section II we discuss some of the related work. Sections III and IV present the algorithm design and modes of operation and the implementation with two methods, software and hardware. Then, in section V, we make a comparative analysis of the solutions, based on execution time and energy consumption, and select the encryption methods suitable for ATmega128RFA1-based platforms, taking also into consideration the provided security. Finally, we present the conclusions of our work.

## II. Related Work

The problem of measuring the cost of encryption on wireless sensor node hardware has been addressed previously. In [5] Lee et al. analyze a range of symmetric-key algorithms and message authentication algorithms in the context of WSNs. They use the MicaZ and TelosB sensor nodes and measure the execution time and energy consumption of different algorithms. For AES they provide measurements for a hardware assisted implementation and conclude that it is the cheapest when either time or energy is considered. They do not however study this implementation on different plaintext lengths and instead rely on datasheets to extend to lengths longer than one block. However, this conclusion is not backed by Zhang et al. in [6] which compares different AES implementations on the MicaZ nodes. They conclude that hardware assisted encryption is faster, but also consumes more energy due to the external chip which handles the computation in hardware.

Compared to their work, we study only AES-128 which is a well known cipher also adopted by the National Institute of Standards and Technology (NIST) and which has been proposed as a viable alternative ([7]) to other less studied ciphers in WSN applications. This choice is also supported by the fact that multiple 802.15.4 transceivers offer a hardware accelerator for AES operations. We study the newer Sparrow v3.2 sensor nodes based on the ATmega128RFA1, which integrates the microcontroller with the radio transceiver and hardware encryption module, and show that AES-128 can be efficiently implemented reducing both execution time and energy consumption. We also provide hybrid implementations for modes of operation that are not natively supported by the

978-83-60810-66-8/$25.00©2015, IEEE

hardware and show that they can still be efficiently implemented with the available primitives.

In [7] Law et al. conduct a thorough survey of the costs of different block ciphers, when implemented on sensor node hardware. They conclude that Rijndael (AES) is the second most efficient cipher, being surpassed only by Skipjack. However, their analysis is based on older hardware and does not consider any hardware accelerated implementations.

In [8] de Meulenaer et al. study the problem of key exchange and measure the cost of two key agreement protocols: Kerberos and Elliptic Curve Diffie-Hellman. They measure the energy consumption of the two protocols on MicaZ and TelosB sensor nodes and conclude that the listening mode is the principal factor in the energy efficiency of key exchange protocols, with Kerberos being the more efficient protocol. Compared to their work, we concentrate on encryption algorithms, and more specifically on AES, with key distribution left for future work.

## III. Design

AES is a block cipher encryption algorithm that uses symmetrical keys for encrypting a block of plaintext and decrypting a block of ciphertext [9]. The algorithm uses a series of rounds consisting of one or more of the following operations: byte-level substitution, permutation, arithmetical operations on a finite field and XOR-ing with a given or calculated key [10]. As a general rule, the operations are handled bytewise.

AES receives as input a plaintext of 16 bytes and the encryption key, which has a variable dimension of 16, 24 or 32 bytes. The input text is processed into the output text (ciphertext) by using the given key and applying a number of transformations. Encryption and decryption are similar, except for the fact that decryption needs an extra step —it first runs a full encryption in order to obtain the modified key needed for decrypting data.

In [11], Schneier divides symmetrical encryption algorithms in two basic categories: block ciphers and stream ciphers. A block cipher encrypts a block of plaintext producing a block of encrypted data, whilst a stream cipher can encrypt plaintexts of varying sizes. This makes block ciphers prone to security issues, if used to encrypt plaintexts longer than the block size, in a naïve way, mainly because patterns in the plaintext can appear in the ciphertext.

A more secure way to encrypt data with a block cipher can be achieved by combining the encryption algorithm with a few basic operations, in a *mode of operation*. It is worth mentioning that the operations are not directly securing data. This is the responsibility of the block cipher. Still, they should not compromise the security provided by the cipher.

### A. Electronic Code Book (ECB)

The ECB mode of operation receives blocks of plaintext, respectively ciphertext, and a key and produces corresponding blocks of ciphertext, respectively plaintext. One property of this mode of operation is that two blocks of plaintext, encrypted with the same key, will result in two identical blocks of ciphertext. ECB is the most simple mode of operation. However, one major drawback is that it does not hide data patterns, meaning that identical ciphertext blocks imply the existence of identical plaintext blocks.

### B. Cipher Block Chaining (CBC)

The CBC mode of operation takes as input parameters the plaintext, respectively the ciphertext, the key and an initialization vector (IV). One property of CBC is that two encrypted blocks are identical only if their respective plaintexts have been encrypted using the same key and the same IV. Unlike ECB, CBC has link dependencies, as its basic chaining mechanism makes the ciphertext blocks dependent on previously encrypted data. This, coupled with a randomly chosen IV, ensures that identical plaintext blocks will be encrypted to different ciphertext blocks.

### C. Cipher Feedback (CFB)

The CFB mode of operation is very similar to CBC regarding its input parameters and the operations it performs. The main difference between them lies in the fact that CBC works as a block cipher, while CFB can be used as a stream cipher. Unlike CBC, CFB can encrypt variable-length blocks (which are not restricted to 16 bytes). The properties of this mode of operation are similar with the ones of CBC. One key difference between the two can be observed at the implementation level: CFB uses only the encryption primitive of the underlying block cipher, both for encrypting and for decrypting data.

### D. Counter (CTR)

The CTR mode of operation also produces a stream cipher. The IV used in CBC and CFB is now associated with the starting value of a counter, which is incremented and used to encrypt each block. In this mode, the output from a previous block is not used for obtaining the input to the current block. In order for the described system to work, a generator is needed on both sides of the communication. The generators have to remain synchronized in order to produce the same stream of data on both sides. A disadvantage of this mode of operation is the possible desynchronization of the communicating entities. This results in the incorrect decryption of all subsequently received data.

## IV. Implementation

A practical example would be a wireless sensor network, which transmits data gathered from three types of sensors: temperature, humidity and luminosity. Because of privacy and integrity concerns all data must be encrypted during transmission. The working platform for this scenario is based on the Sparrow v3.2 node [12]. Its technical specifications are:

- CPU: ATmega128RFA1, 16MHz
- Memory: 128KB flash, 16KB RAM
- Bandwidth: up to 2Mbps
- Programming: C/C++

The ATmega128RFA1 microcontroller is actually a SoC (System on Chip) which incorporates a radio transceiver compatible with the IEEE 802.15.4 standard [13]. It offers, among other things, a relatively low energy consumption (mostly in sleep states), a FIFO buffer of 128 bytes for receiving and transmitting data, a partial hardware implementation of the MAC protocol and support for AES-128.

This microcontroller facilitates secured data transmissions by incorporating a hardware acceleration module which implements the AES algorithm. The module is capable of encrypting and decrypting data in a fast track way, as most of the functions are implemented directly in hardware. It is compatible with the AES-128 standard (the key is 128 bits long) and supports encryption and decryption for ECB mode, but only encryption for CBC mode. The input to these operations consists of the plaintext/ciphertext block and the encryption key. Note that for decryption, the extra round needed by AES to compute the decryption key is performed automatically. Other modes of operation are not supported by the hardware.

As we already stated in the previous sections, energy consumption is the main issue and challenge for WSNs. In order to obtain the best approach for ensuring confidentiality with a minimum energy use, we implemented and compared AES-128, coupled with the ECB, CBC, CFB and CTR modes of operation. All four modes have both a hardware and a pure software implementation. Since only ECB has a full hardware implementation, for the other modes we used a hybrid approach, combining the hardware part from ECB with software implementations for the remaining operations. We also refer to these hybrids as hardware implementations. For the pure software implementation we used an optimized version of AES, called TableLookupAES [6].

## V. Evaluation

### A. Experimental setup

To measure the energy consumption of our implementation, we perform two kinds of measurements: the time required ($t$) and the current drawn by the node ($I$) while encrypting/decrypting. Using the formula $E = P \cdot t$, where $P = U \cdot I$ is the power required by the node, we can compute the energy consumed by the algorithm, be it implemented in software, in hardware or using a hybrid approach. We ensure a constant voltage $U$ using a voltage regulator, as explained in the next subsection.

In certain applications, the latency of encrypting/decrypting a given payload might be more important than the energy consumed. For this reason, this section also presents the timing results of the different solutions, independent of the energy measurements. As we later show, the current drawn by the node using both software and hardware security approaches is practically the same. Thus, the time taken is a sufficient metric for relative comparisons between the different solutions.

*1) Current measurement:* For the purpose of measuring the energy consumption of the Sparrow sensor node during our experiments, we built a current sensing circuit based on the INA 193 current shunt monitor.

Fig. 1 presents the circuit we designed. Power is provided by a $3.3V$ voltage regulator, which ensures a constant voltage regardless of the current drawn by the circuit. A shunt resistor connected in series with the Sparrow node acts as a current sensor. The voltage drop on the resistor is directly proportional with the current drawn by the circuit. This has two implications. On the one hand, the chosen resistor value must be small enough not to disturb the rest of the circuit (e.g. by incurring a big voltage drop). On the other hand, the same value has to be big enough so that the expected currents register a voltage drop that can be sensed with enough precision. In order to improve the measurement precision and sensitivity, without the drawbacks of a big resistor value, we employ a INA 193 current shunt monitor, which provides a constant gain of $20V/V$ on the input voltage drop, and a $4.99\Omega$ precision resistor with a tolerance of $0.01\%$. The output of the current sensing circuit is connected to a Metrix OX 5042 oscilloscope which we used to monitor the current drawn by the node during the different encryption/decryption operations. Determining the current is as simple as dividing the voltage shown on the oscilloscope by the current shunt monitor gain $(20V/V)$ and the shunt resistor value $(4.99\Omega)$.

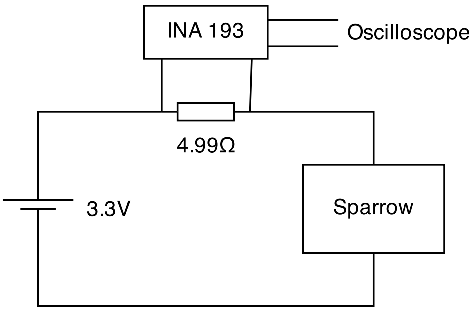


Fig. 1. Current measurement setup

*2) Time measurement:* Using the oscilloscope, we also measure the time required for each encryption/decryption operation. The oscilloscope has a function that accurately measures pulse duration. We create a pulse lasting for the duration of the operation by setting a GPIO pin before the start of the operation and clearing it after it ends. Using this method, we can measure the duration of an operation with minimal overhead: 1 bit set instruction and 1 bit clear instruction, each taking 2 cycles.

Although the proposed measurement scheme is precise, it has the disadvantage of requiring manual intervention. The available oscilloscope cannot be interfaced with a PC, so a measurement point is obtained by uploading a program which encrypts a hardcoded message length in a loop, reading the information from the oscilloscope and repeating the process for all message lengths.

In order to automate the time measurements, we resorted to a software implementation running along side the encryption/decryption operation, that measures the time required. To keep overhead to a minimum, our solution employs the hardware timer module available on the ATmega128RFA1 to count the number of cycles taken by the operation. Each operation is measured by sampling a counter before and after the operation and taking the difference of the two values. The count is then converted to a time value given that the microcontroller operates at $16MHz$.

This time measurement solution allowed us to automate the whole process of evaluating the algorithms for different message sizes. A small overhead can be observed between the software based time measurement and the oscilloscope based one, but the relative difference between the algorithms is unaffected. If absolute numbers are required, the software-based measurements can be corrected by noticing that the overhead increases linearly with the message size when compared with the oscilloscope measurements.

### *B. Results*

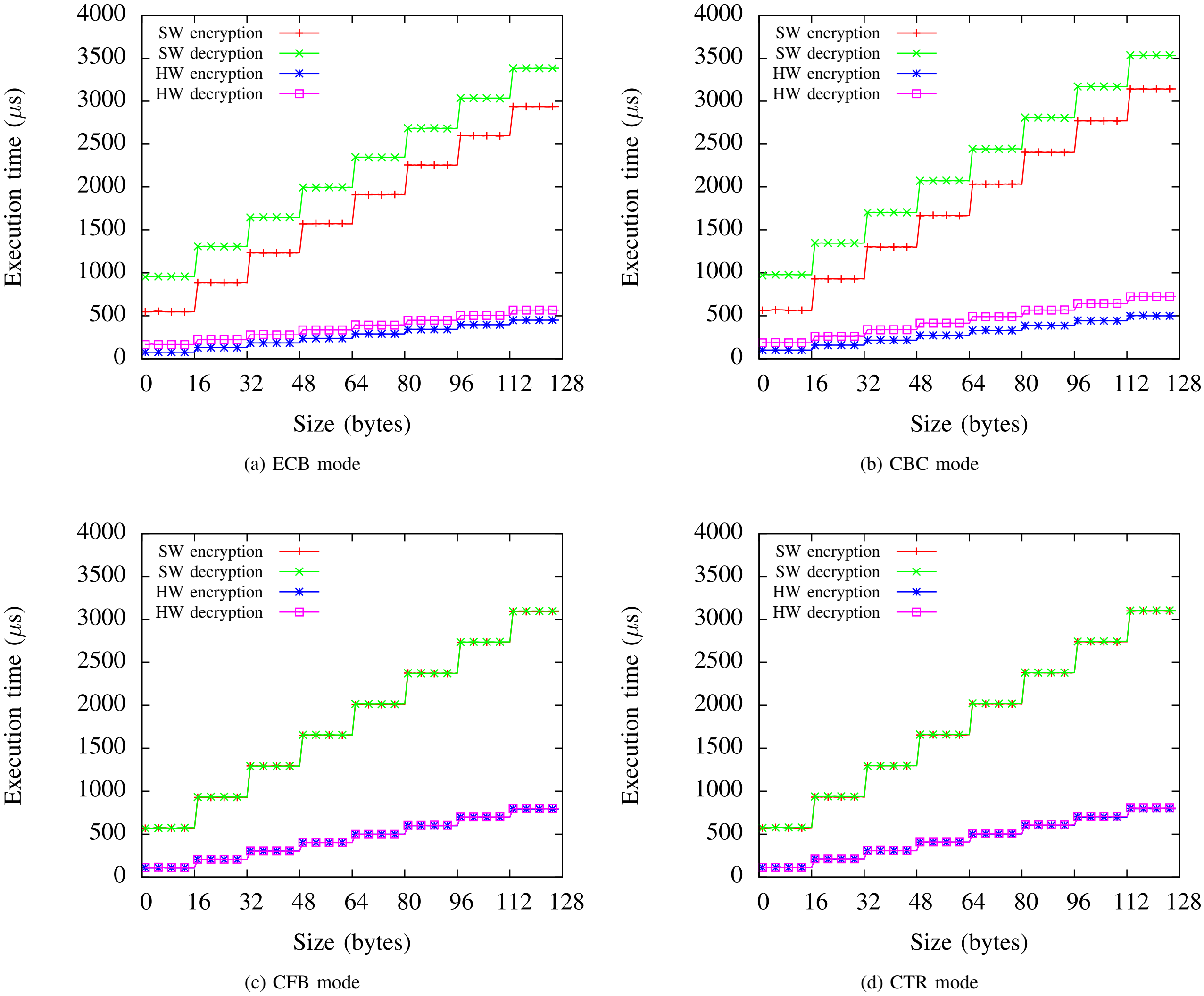
We conducted multiple experiments, to evaluate both the time taken and the energy consumed by AES encryption/decryption operations. We measured our hardware assisted implementation against the pure software implementation based on look-up tables.

*1) Time experiments:* We started of with measuring the difference between the optimized software implementation and our hardware assisted implementation for each of the 4 studied modes of operation. For each type of implementation and operation mode, we measured the time taken by an encryption operation and a decryption operation on varying message lengths. We used message lengths from 1 byte to 127 bytes, which is the maximum packet size allowed by the transceiver and the 802.15.4 standard.

As can be seen in figure 2, the hardware assisted implementation easily outperforms the optimized software implementation. The staircase shape of the graph is easily explained by the requirement of every block cipher, including AES, to operate on multiples of the block size. Plaintext sizes that are not a multiple of the block size need to be padded, thus still incurring the cost of an entire block.

The difference in performance varies between ∼6.5x for the ECB mode, which is fully supported in hardware, down to ∼3.8x for the CFB and CTR modes, which are only partially supported in hardware through the AES single block encryption primitive. The difference in performance between the optimized software implementation and our hardware assisted implementation is summarized in table I.

For the ECB and CBC modes we can also observe (figures 2a and 2b) the extra preparation step needed by the single block decryption primitive, which makes decryption slightly



(a) ECB mode

(b) CBC mode

(c) CFB mode

(d) CTR mode

Fig. 2. Comparison between software and hardware AES implementations

| | encryption | decryption |
|---|---|---|
| ECB | 4.61x - 6.49x | 5.59x - 6.02x |
| CBC | 5.51x - 6.28x | 4.86x - 5.28x |
| CFB | 3.87x - 5.29x | 3.86x - 5.34x |
| CTR | 3.87x - 5.17x | 3.85x - 5.17x |

TABLE I. EXECUTION SPEED-UP HARDWARE VS. SOFTWARE

more time consuming than encryption. No difference can be observed (figures 2c and 2d) between encryption and decryption for the CFB and CTR modes, because they only use the encrypt primitive of AES for both encryption and decryption, albeit with some extra software processing.

Figure 3 compares the hardware assisted implementations of the 4 modes against each other during encryption and decryption. For encryption, ECB has the lowest runtime for all sizes, which was to be expected, as it does no extra operations on the output of the encrypt primitive to mask patterns in the plaintext. CBC is slightly worse, as it adds a XOR operation, which is implemented by the hardware accelerator, but better than CFB and CTR, which have no hardware support, except for the encryption primitive. For decryption, CFB and CTR have a slight advantage over ECB and CBC for small sizes, as they only use the encrypt primitive, which has a smaller setup time than the decrypt primitive. This advantage is lost at around 32 bytes with respect to ECB and at around 64 bytes with respect to CBC. From the plots we can also observe that the extra software processing done on top of the AES encrypt primitive by CFB and CTR is similar in overhead, for both encryption and decryption.

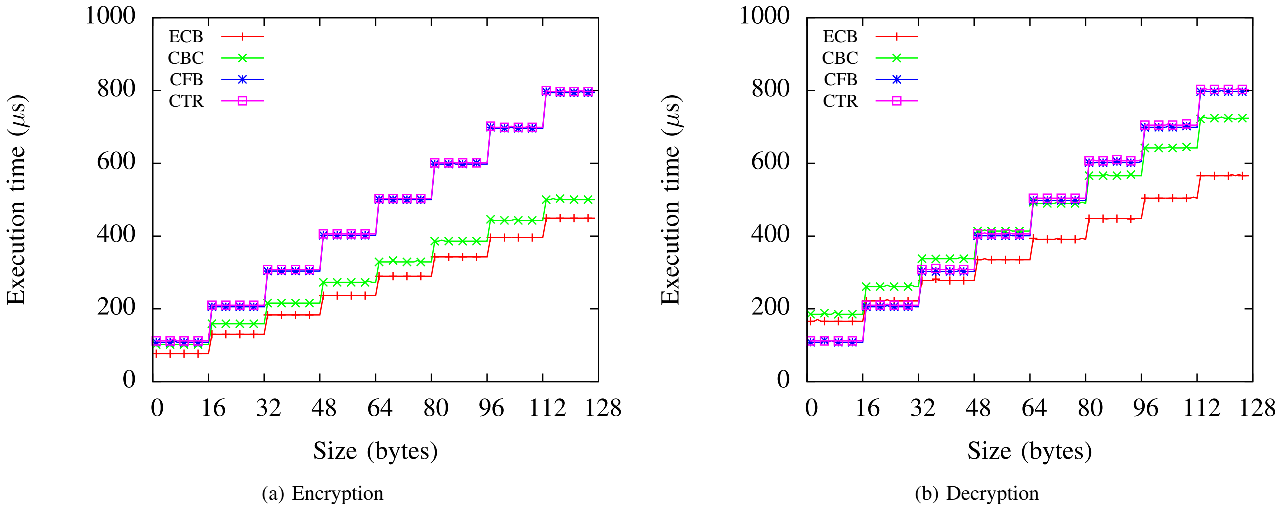
If we look at the cumulated time of both encryption and decryption, CFB and CTR still hold an advantage up to 32 bytes with respect to CBC. Thus, for small message sizes, as it usually happens in WSNs, it might be more efficient to use the CFB or CTR modes even if they are not completely accelerated in hardware.

*2) Energy experiments:* For energy consumption we concentrated our efforts on determining the cost of using AES in CFB mode. We chose this mode based on the fact that the timing measurements showed it to be the best encryption/decryption mode for small message sizes, similar to those that are commonly found in WSNs. We only performed measurements for message encryption, as decryption is identical in terms of the code which is ran. We measured the cost of doing the encryption in software as well as the cost of using our hardware accelerated implementation. For completeness, we also measured the cost of an empty processing loop to compare against the two encryption implementations.

In our experiments, we used the measurement circuit described in subsection V-A1 to measure the base and peak currents during encryption, as well as the voltage and duration of the operation, as reported by the oscilloscope. As with the timing measurements, we performed the experiment for different message size, from 1 byte to 127 bytes. The oscilloscope was configured to report the mean over 16 samples in order to obtain the average energy consumption of the device. An instantaneous energy consumption is hard to obtain and is irrelevant when considering the long time operation of the node.

Using the raw current and voltage measurements, we plot the average power drawn with respect to the encryption size. As can be seen in figure 4a, the software and the hardware solutions draw equal amounts of power. Furthermore, this average power is independent of the plaintext size and is only slightly higher than the average power drawn by the empty processing loop.

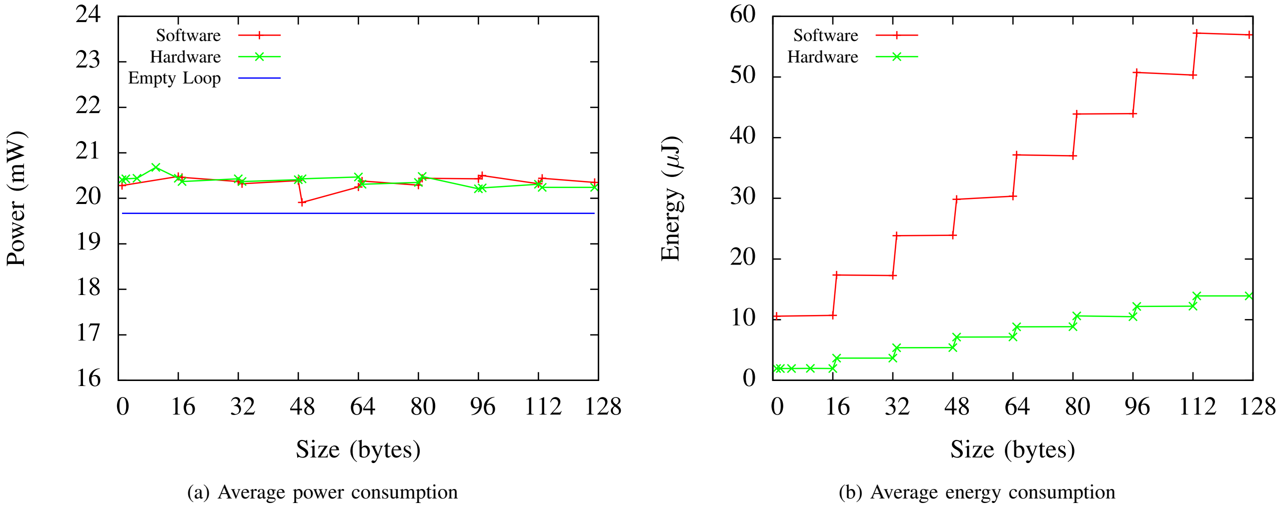
If we plot the average energy consumed by the encryption operation (figure 4b), we see a linear increase in energy consumption with increasing plaintext size. Using the timing measurements performed in the previous subsection and the average power values from figure 4a, we can also derive the average energy consumption for every mode, operation and plaintext size, not just for encryption in CFB mode. This can be done by adjusting for the overhead induced by the software timer, using a correction factor deduced from correlating the oscilloscope timings with the internal timer timings.



(a) Encryption

(b) Decryption

Fig. 3. Comparison between modes of operation with hardware acceleration



(a) Average power consumption

(b) Average energy consumption

Fig. 4. Power and energy consumption of AES encryption in CFB mode

## VI. Conclusion

In this paper an evaluation of the cost of adding AES-128 encryption to WSN communications has been presented. Both the time penalty as well as the more important (from the point of view of a WSN) energy penalty have been analyzed, for multiple modes of operation: ECB, CBC, CFB and CTR and for two implementations: a pure software implementation, based on the optimized table lookup AES and a hardware accelerated implementation, that uses the AES hardware module of the ATmega128RFA1 microcontroller.

We showed how the AES hardware module in the ATmega128RFA1 microcontroller can be used to implement other modes of operation than the ones supported natively. Our solution uses a hybrid approach that runs some operations in hardware and emulates the missing ones in software. Using this approach, we implemented CBC decryption, as well as two full modes of operation for AES, CFB and CTR, which do not have direct hardware support.

We presented a methodology of accurately measuring the power consumption using low cost components and a way of determining the encryption/decryption duration using only the wireless node itself. We compared the different modes of operation and concluded that except for the unsecure ECB mode, CFB and CTR are better overall alternatives for the small message sizes (below 32 bytes) usually exchanged in WSNs. This is true even though the hardware accelerator has native support for the CBC mode and it relates to the way decryption works for CBC.

We also built on the work of Zhan [6] and showed that the newer ATmega128RFA1 microcontroller with an integrated transceiver, used in the Sparrow v3.2 node, can reduce both the duration and the energy consumption of AES operations. This is in contrast to work done on previous sensor nodes, that used a separated microcontroller and transceiver and which had a higher energy cost when running the encryption in hardware as opposed to using a pure software implementation.

## References

[1] H. Karl and A. Willig, *Protocols and architectures for wireless sensor networks*. John Wiley & Sons, 2007. ISBN 978-0-470-09510-2

[2] D. W. Carman, P. S. Kruus, and B. J. Matt, "Constraints and approaches for distributed sensor network security (final)," DARPA project report, NAI Labs, Cryptographic Technologies Group, Trusted Information System, Tech. Rep. 1, 2000.

[3] Y. Wang, G. Attebury, and B. Ramamurthy, "A survey of security issues in wireless sensor networks," 2006. doi: 10.1109/COMST.2006.315852

[4] J. Sen, "Routing security issues in wireless sensor networks: attacks and defenses," in *Sustainable Wireless Sensor Networks*, W. Seah and Y. K. Tan, Eds. InTech, 2010. doi: 10.5772/663 pp. 279–309.

[5] J. Lee, K. Kapitanova, and S. H. Son, "The price of security in wireless sensor networks," *Computer Networks*, vol. 54, no. 17, pp. 2967–2978, 2010. doi: 10.1016/j.comnet.2010.05.011

[6] F. Zhang, R. Dojen, and T. Coffey, "Comparative performance and energy consumption analysis of different aes implementations on a wireless sensor network node," *International Journal of Sensor Networks*, vol. 10, no. 4, pp. 192–201, 2011. doi: 10.1504/IJSNET.2011.042767

[7] Y. W. Law, J. Doumen, and P. Hartel, "Survey and benchmark of block ciphers for wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 2, no. 1, pp. 65–93, 2006. doi: 10.1145/1138127.1138130

[8] G. De Meulenaer, F. Gosset, O.-X. Standaert, and O. Pereira, "On the energy cost of communication and cryptography in wireless sensor networks," in *Networking and Communications, 2008. WIMOB'08. IEEE International Conference on Wireless and Mobile Computing,.* IEEE, 2008. doi: 10.1109/WiMob.2008.16. ISBN 978-0-7695-3393-3 pp. 580–585.

[9] J. Daemen and V. Rijmen, "The block cipher rijndael," in *Smart Card Research and Applications*. Springer, 2000. doi: 10.1007/10721064_26 pp. 277–284.

[10] W. Stallings, *Cryptography and Network Security - Principles and Practice, Fifth Edition*. Pearson Education, 2011. ISBN 978-0-13-609704-4

[11] B. Schneier, *Applied cryptography: protocols, algorithms, and source code in C*. John Wiley & Sons, 1996. ISBN 978-0471117094

[12] A. Voinescu, D. Tudose, and D. Dragomir, "A lightweight, versatile gateway platform for wireless sensor networks," in *Networking in Education and Research, 2013 RoEduNet International Conference 12th Edition*. IEEE, 2013. doi: 10.1109/RoEduNet.2013.6714202 pp. 1–4.

[13] *8-bit AVR Microcontroller with Low Power 2.4GHz Transceiver for ZigBee and IEEE 802.15.4*, ATmega128RFA1, Atmel.

DOI: 10.15439/2015F118
ACSIS, Vol. 5

# Battery Aware Beacon Enabled IEEE 802.15.4: An Adaptive and Cross-Layer Approach

Marwa Salayma, Ahmed Al-Dubai, Imed Romdhani, Muneer Bani Yassein*
School of Computing, Edinburgh Napier University, Edinburgh, UK
{M.Salayma; A.AL-Dubai; I.Romdhani}@napier.ac.uk
*Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan
*masadeh@just.edu.jo

***Abstract*—In Wireless Sensor Networks (WSN$_s$), energy conservation is one of the main concerns challenging the cutting-edge standards and protocols. Most existing studies focus on the design of WSN energy efficient algorithms and standards. The standard IEEE 802.15.4 has emerged for WSN$_s$ in which the legacy operations are based on the principle that the power-operated battery is ideal and linear. However, the diffusion principle in batteries shows the nonlinear process when it releases a charge. Hence, we can prolong the network lifetime by designing optimized algorithms that reflect the battery characteristics. Within this context, this paper proposes a cross-layer algorithm to improve the performance of beacon enabled IEEE 802.15.4 network by allowing a Personal Area Network Coordinator (PANc) to tune its MAC behavior adaptively according to both the current remaining battery capacity and the network status. The performance of the new algorithm has been examined and compared against that of the legacy IEEE 802.15.4 MAC algorithm through extensive simulation experiments. The results show that the new technique reduces significantly the energy consumption and the average end-to-end delay.**

***Index Terms*—WSN; IEEE 802.15.4, MAC; analytical model; battery aware; delay.**

## I. Introduction

RECENTLY, most wired sensors are now being replaced with wireless ones creating the emerging era of Wireless Sensor Networks (WSNs). WSNs consist of sensing devices that can communicate with each other and with the surrounding environment via the wireless communication medium [1] [2]. Yet, a huge number of sensor nodes are often scattered in unreachable areas, and WSNs are often battery powered and cannot be easily recharged. Thus, energy conservation is one of the main concerns in the area of WSN. Many studies that focus on designing WSN energy efficient algorithms and standards, based on IEEE 802.15.4, have emerged recently [3]. The IEEE 802.15.4 standard supports both physical and Media Access Control (MAC) layers. IEEE 802.15.4 MAC supports two types of devices, namely, *Full Functional Devices (FFDs)* and *Reduced Functional Devices (RFDs).* FFDs act as a regular coordinator and/or as a sink node. If both features are taken, the node is typically referred to as Personal Area Network Coordinator (PANc). In contrast, RFDs act as an ordinary end device [4]. Despite these differences, both FFDs and RFDs communicate with each other, forming two types of topologies, *star* and *peer to peer* topologies. All of the supported topologies must have one PANc. In the star topology, all network nodes can only communicate with the PANc in their active period. On the other hand, in mesh network topologies, all devices can talk to each other directly sending broadcast messages.

The IEEE 802.15.4 MAC layer operates either in beacon enabled or beaconless modes. In the beacon enabled mode, the FFD broadcasts regular beacon frames that synchronise nodes when they need to access the channel [4]. The time between two successive beacons is referred to as the *Beacon Interval (BI),* which is divided virtually into 16 equal sized slots. BI duration is specified by the Beacon Order parameter (BO) according to the following formula [4].

$$BI = aBaseSuperframeDuration * 2^{BO} \qquad (1)$$

Nodes can use the channel during the whole BI period or can sleep for some time portions depending on Superframe Order (SO) parameter. This parameter decides the Superframe Duration (SD) active session according to the following formula [4].

$$SD = aBaseSuperframeDuration * 2^{SO} \qquad (2)$$

where $0 \leq SO \leq BO \leq 14$

The aBaseSuperframeDuration value depends on the slot duration according to the following formula.

$$aBaseSuperframeDuration= aBaseslotDuration * total\ slots \quad (3)$$

All these concepts can actually be indicated through one concept: the duty cycle (D). This is the percentage of time the node is awake from the whole time between the two successive beacons. D is mathematically expressed as [5][6]:

$$D=SD/BI * 100\% \qquad (4)$$

When a node needs to access the medium, it has to locate the beginning of the next time slot in order to compete for the channel based on the Carrier Sense Multiple Access/Collision Avoidance algorithm (CSMA/CA). This time portion is referred to as the Contention Access Period (CAP) [4-6].

The lengths of the discussed periods are assigned through the beacon frame, which is transmitted in the first time slot (slot 0) [5][6]. Due to the complicated issues of the inactive period, most beacon enabled IEEE 802.15.4 studies are limited to star one-hop topology. Similarly, this paper considers the same assumption.

BO and SO values controls the performance of beacon enabled IEEE 802.15.4. Small BO values lead to frequent beacon frames and consequently increase beacon overhead, which, in return, drains more power in a short period of time. Small SO values, on the other hand, decrease nodes active time, while increasing the sleep time period. However, while small SO values might save energy, they increase delay and adversely affect throughput. This is because nodes which do not have enough time to send their data frames during the current superframe, will differ in their activity to the next superframe and therefore attempt to send their data packets in one go causing collision. Clearly, this situation becomes worse as the number of nodes increases [5][6]. Beacon overhead, collision and packets retransmission are all reasons for early battery charge depletion. In order to maximize node lifetime, we

978-83-60810-66-8/$25.00©2015, IEEE

need to increase battery lifetime. To achieve that we need to analyse the battery behaviour and study how it cope with IEEE 802.15.4 operations. Most of the studies found in literature are based on the fact that the IEEE 802.15.4 battery operations are ideal and linear. Batteries deliver power based on the electro-chemical reactions that occur between the electrodes and the active material around the electrodes. Continuous electro-chemical reactions deplete active mass near the electrode. Active material is able to diffuse towards the electrodes when the battery is idle allowing battery to heal and gain some of its charge [7]. This is called *battery recovery effect,* which occurs in the idle recovery time. Accordingly, in order to provide energy, real battery behaviour is governed by complex non-linear internal chemical reactions which occur due to the nonlinear behaviour when it gains charge in the recovery time [7].

To conclude, we can prolong network lifetime through adaptive techniques designed according to the battery's behaviour. For instance, we can exploit battery recovery effect by adding a relaxation time artificially between two packets in order to gain more capacity charge. Through this way, the performance of IEEE 802.15.4 could be improved by adopting IEEE 802.15.4 battery-friendly algorithm for the packets transmission.

Nevertheless, improving battery performance should not be at the expense of the standard reliability. Packets average end to end delay can be estimated at higher layers of the protocol stack, while energy consumption and battery behaviour are evaluated at the lower layers with respect to the OSI model. Therefore, the overall IEEE 802.15.4 protocol stack needs to be revisited so that the MAC layer can adaptively tune its parameters according to the actual needs in terms of the available battery capacity and the current delay. In other words, friendly battery management technique should be able to adapt to the actual network operating status; according to average end to end delay for example. This could be achieved for example by exploiting information provided by the different layers of the protocol stack. By following a cross-layer approach we do believe that we can minimize the energy expenditure.

In this paper, we propose an adaptive and cross layer approach that improves the beacon enabled IEEE 802.15.4 performance by allowing the MAC layer to tune its parameters according to the battery behaviour of the coordinator as well as the network status in a star topology. To summarize, the contribution of this paper is fourfold:

- The real behaviour of the battery in a beacon enabled IEEE 802.15.4 MAC is investigated by considering battery nonlinearity by analysing the diffusion of chemical reactions in the battery following *Rakhmatov model*.
- The gain of the battery recovery effect according to what sleep period can increase battery life time of the beacon enabled IEEE 802.15.4 is analysed.
- A cross-layer and adaptive battery aware beacon enabled IEEE 802.15.4 MAC that tunes synchronization time according to current battery status is proposed.
- The network reliability is considered by checking network delay and tune nodes active period accordingly.

This paper is organized as follows. Section II summarizes some of the literature work which is closely related to the paper topic, while Section III illustrates the battery models and our followed methodology. In Section IV we evaluate and discuss the performance of our proposed protocol. Section V concludes the paper and outlines future work.

## II. Related Work

Recently, there has been significant amount of studies that addressed the electro-chemical behaviour of batteries. Li et al. [8] proposed an analytical model that computes the life time of a low duty cycled star sensor network. In their model they considered nonlinearities of lithium-ion battery following Rakhmatov model [15] and they aimed to minimize the total energy consumption of the lithium-ion battery by finding the optimal idle and sleep period while guaranteeing energy efficiency, reliability and reasonable latency. In their proposal, they considered the trade-off between energy that is dissipated in sending frequent preambles and the period thereby sensors stay idle waiting for the preamble. Experimental results show that the proposed method can provide the optimal sleep or channel check intervals that maximize the lifetime of the network while guaranteeing a little latency and high reliability. However, this model target only a simplified work mechanism of MAC protocol, without giving details of battery nonlinearity effects on the proposed protocol on their network.

Li et al. [9] presented three battery aware algorithms that reduce power consumption and extend battery lifetime. Each one of the proposed schemes is targeted towards a specific application type, which are the hard real-time applications, the soft real-time applications, and the periodic applications. For hard real-time applications, Battery-friendly lazy packet algorithm is proposed to minimize battery charge consumption by allowing it drew lower current. A battery-friendly local optimization algorithm with slack time is targeted towards the soft real-time applications. For the periodic applications, a battery-aware task-scheduling algorithm is developed, which performs task rescheduling to achieve the battery friendly discharge profile. Li et al. [9] follows Rakhmatov model to depict both battery and recovery effect and nonlinearity. Simulation results demonstrated that the three battery-friendly algorithms perform better in extending lifetime of battery-operated sensor nodes as they reduce battery charge consumption.

Chau et al. [10] attempted to exploit the battery recovery effect in WSN. They empirically studied the gain at which the battery recovery effect prolongs commercial sensors lifetime. This effect has also been studied analytically corroborated by simulation. The outcome of [10] revealed that there is a saturation threshold at which the battery recovery resulted from idle listening will contribute less in improving the behaviour of the battery. Authors in [10] proposed a distributed battery aware duty cycle protocol and measure the battery runtime under both deterministic and randomized schedules. The authors in [10] studied also the trade-off between both delay and harnessing the recovery effect and suggested that we can perfectly harness battery recovery without increasing delay if we carefully adjust the sleep time period before reaching the saturation threshold. The authors in [11] derived upper bounds of battery lifetime and proposed a more energy-efficient algorithm that is aware of battery recovery effect and this is by extending the pseudo-random duty cycling scheme proposed in [11] by a *forced sleep*. In addition, the authors in [11] achieved analytical results that predict the average delay in sensor networks by setting the sleep duration of the RF transceiver as the saturation threshold of the battery, which can take the maximal advantage of the duration-dependent battery recovery effect. The authors in [11] presented also a useful tool to compromise the trade-off between increasing battery lifetime of sensor networks and the average delay of delivered packets.

Casilari et al. [12] proposed an analytical model that forecasts the minimum, mean and maximum battery lifetime of a WSN by allowing it to work under different traffic load, data rate and probability of packet loss. This is done by an experimental characterization of activity cycles battery consumption in commercial motes that follows the 802.15.4/ZigBee stack and also by measuring the current that is drained from the power source under different 802.15.4 communication operations [12]. The characterization considers the different operations required by 802.15.4 protocol and takes into consideration the delay introduced by the CSMA/CA algorithm applied by the 802.15.4 MAC layer. The model has also been extended to cope with the extra consumption that the node re-association requires when a packet loss occurs. Mario et al. [13] proposed an adaptive and cross-layer energy-aware module for energy-efficient and reliable data collection targeted towards IEEE 802.15.4/ZigBee WSNs. The proposed module captures the packet delivery ratio at the application and configures the MAC layer parameters, which are backoff window size and the number of (re)transmissions, according to the traffic conditions in order to minimize the power consumption.

## III. Battery Model & Protocol Description

Battery is a repository of electrical charges which provides voltage and current for the components attached to it, such as radio transceiver, microprocessor, memory, sensor, etc. A battery losses charge when a load draws current from it, where the loss rate is a function of the load [7]. It is common that the Radio Frequency (RF) transceiver operations are the most energy consumable resources (even in listening mode), as compared to the processing and sensing activities [8]. As it shown in (5), the total energy ($E_{total}$) consumed by RF energy model is the total sum of energies consumed by sensor in performing the four operations, which are: transmitting ($E_{tx}$,), in receiving ($E_{rx}$), being idle ($E_{idle}$) and in sleeping ($E_{sleep}$).

$$E_{total}= E_{tx}+E_{rx}+E_{idle}+E_{sleep}. \quad (5)$$

IEEE 802.15.4 standard, as many of other previous studies, consider ideal behavior of the battery, that is, voltage stays constant over time until the moment it is completely discharged, then the voltage drops to zero, whereas the capacity is the same for all loads that the battery generates [7]. The total energy E in the ideal case can be calculated as follows.

$$E = V \text{ x } C, \text{ and } C= I \text{ x } L \quad (6)$$

where E (Watt-hour) is the provided energy, V is the voltage (volt), C is the total capacity of the battery (Ampere-hour), I is the provided load (Ampere) and L is the lifetime of the battery (hour) [14].

In order to exploit battery characteristics in our protocol, we need to study its electro-chemical behaviour, which can be analysed empirically or through models [7]. Empirical analysis is time consuming and requires expensive prototyping and measurement for each alternative. Therefore, battery behaviour under various conditions of charge/discharge can be predicted through models [7-10].

Models for energy consumption and performance estimation of each system component are described in the following sub-section.

### *A.* Battery Models

There are different models that describe the battery discharge processes. Each model type has a varying degree of accuracy and complexity [7][8]. Those models can be classified as low level electro chemical models and high level mathematical models. Electro chemical models are the least flexible and the most computation intensive, so they are sophisticated models to use for battery modeling and they are the most accurate ones. On the other hand, electrical circuit models, analytical models and the stochastic models can be easily configured for different types of batteries. Electrical circuit models are highly efficient when used for simulation but they ignore the effects of charge recovery during idle periods. The stochastic models are highly efficient for simulation and are capable of modeling rate capacity and recovery effects. Analytical models are computationally efficient, but limited in the discharge effects they model. One of these models is Sarma and Rakhmatov model which is an abstraction of a real battery [15] that we used in this work. Rakhmatov model is chosen for estimating the real residual battery capacity at a specific time, because it is the simplest accurate analytical model. Other models require solving complex Partial Differential Equations (PDEs) which are difficult to optimize [7][15]. For the model to adequately mimic real behaviour of the batteries, one can utilize this formula:

$$\alpha = \mathrm{I}\left[L+2\sum^{\infty}\frac{1-e^{-\beta^{2}m^{2}L}}{\beta^{2}m^{2}}\right] \quad (7)$$

Where I is the applied load and L is battery lifetime, α is the capacity of the battery when it is fully charged, β refers to battery materials diffusion around the electrolyte and measures the nonlinearity of the battery as it tells us how fast the diffusion process can keep up with the rate of discharge. The value of α is a battery related parameter and its value is decided by manufacture of battery designer [15]. Formula (7) indicates that the total capacity of the battery is the sum of two terms, the linear ideal behaviour plus the nonlinear behaviour. As long as β value is large, the battery behaviour becomes closer to battery ideal effect. When β goes to infinity, the battery works in its ideal situation. This means that the higher the value of β, the better the battery performs. The value of β is estimated from the data sheet of the battery. For example, the data sheet of a battery might model rated capacity (in Ahr) vs. discharge current (in hour) [15]. Thus, before one can use the proposed model, the parameters need to be estimated from experimental data for the modeled battery. Simple experiments with constant loads are sufficient for estimation purposes. However, choosing the optimized values for both α and β is beyond the scope of this paper.

It is important to note that the load generated from the battery is discharged according to different transceiver activities (transmission, receive and idle), each has its own time duration. Thus, the load can be depicted in the form of consecutive N constant current values $I_1$, $I_2$, $I_3$, …., $I_N$ , where $I_k$ is the current of activity k which took place at time $t_k$ in the duration of $\Delta k = t_{k+1} - t_k$ [15]. Accordingly, battery capacity when it is fully charged can be depicted as follows:

$$\alpha=\sum_{k=1}^{N} I_K \Delta_k+\sum_{k=1}^{N} 2I_k \sum_{m=1}^{\infty} \frac{e^{-\beta^2 m^2 (L-t_k-\Delta k)} e^{-\beta^2 m^2 (L-t_k)}}{\beta^2 m^2} \tag{8}$$

In order to calculate the remaining capacity at a specific time unit, we need first to calculate the amount of charge consumed after performing M activities (charge lost from the battery) which is denoted by σ as follows:

$$\sigma(t)=\sum_{k=1}^{M} I_k \Delta_k+\sum_{K=1}^{M} 2I_k \sum_{m=1}^{\infty} \frac{e^{-\beta^2 m^2 t}\left(e^{\beta^2 m^2 \Delta_k}-1\right)}{\beta^2 m^2} e^{-\beta^2 m^2 t_k} \tag{9}$$

According to (8) and (9), the residual capacity at a specific time t (the available charge) is presented here:

$$\alpha(t)=\alpha-\sigma(t) \tag{10}$$

### B. *Battery Recovery Threshold*

Duty cycling is a technique adopted to regulate the on/off periods of the RF transceiver, while keeping the rest of sensor module on. It is important to design proper duty cycling and buffering strategies that can maximize the battery recovery effect when a transmitter moves to an inactive state during which the battery load becomes low, allowing the battery to recover. This results in extending the battery lifetime.

Before injecting this factor in order to improve the IEEE 802.15.4 performance, we need to analyse the rate at which sleep period can maximize battery lifetime. To achieve that, we followed Rakhmatov model and studied different active periods with different sleep time portions allowing the duty cycle to decrease gradually and we analysed the residual capacity in a star topology with 7 clients. Simulation parameters are depicted in Table 1 and the achieved results are depicted in Fig. 1. It can be noticed from Fig. 1 that, for all the tested active periods other than 61.44 ms, increasing the sleep time portion by 50%, and thus decreasing the duty cycle, increases the total residual capacity. On the other hand, allowing node sleep more than 50% decreases the total residual capacity. That is because the sleep time portion allows the chemical charge diffuses around the electrode which enables the battery heal and regain some of its charges because of the battery recovery effect principle. Moreover, it can be noticed that the effect of battery recovery increases as the active period increases, because for in a longer active time, node have enough time to do its activities and thus avoid other unnecessary operations such as, retransmission, which will save battery energy due to the increased residual capacity. This explains why for 61.44 ms active period, as the sleep time increases, total battery capacity increases for the three tested duty cycles. For this short period, a node does not have adequate time to perform its activities at all. Instead, it will keep differing its activity to the next superframe. Consequently, as all nodes will try to transmit together, this will cause frequent collision and retransmission which adversely affect network performance. It is therefore better for the node to sleep than to stay active. To conclude, in order to exploit battery recovery effect, BO is needed to be increased only by one as this will allow the node to operate within 50% duty cycle. BO can be incremented according to the current battery status. The algorithm description is presented in the following subsection.

### C. *The Proposed Technique:*

Achieved results motivate us to propose a more energy-efficient duty cycling scheme by setting the sleep duration of the coordinator RF transceiver at the saturation threshold of the battery, which can take the maximal advantage of the duration-dependent battery recovery effect.

Nevertheless, improving battery performance should not be at the expense of other performance metrics. Packets end to end delay can be estimated at higher layers of the protocol stack, while energy consumption and battery behavior are evaluated at the lower layers. Therefore, the overall IEEE 802.15.4 protocol stack is needed to be considered for the MAC layer to adaptively tune its parameters according to the actual needs. In other words, friendly battery technique should be able to adapt to the actual network operating status. Through this approach, both physical and application layers cooperate with MAC layer in order to prolong network lifetime by preserving energy battery charge at the physical layer, while considering average end to end delay delay status at the application layer.

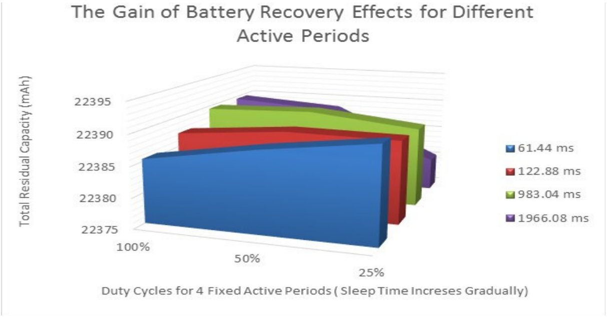


Fig. 1: Residual battery capacity for different duty cycles with different active periods.

This works as follows, before sending a new beacon frame, PANc asks the physical layer for its total residual capacity. If the new residual capacity is worse than the previous one, then it increments the value of BO, otherwise it does nothing. At the same time, PANc also checks the number of received packets at the application layer, for example, if the number is five, and if the new average end-to-end delay is worse than the previous one, then it increments the value of SO, otherwise it does nothing. The new battery aware IEEE 802.15.4 MAC algorithm is summarized in Fig. 2.

## IV. Performance Evaluation

Using QualNet 5.2 Simulator, the performance of the new proposed approach is evaluated by conducting a comparison against the legacy IEEE 802.15.4 performance in terms of total energy consumption, total battery residual capacity, average end-to-end delay and throughput. Evaluation process is applied on a star topology of seven RFDs with 7 Constant Bit Rate (CBR) traffic applications working over 1000 s simulation period.

**Algorithm:** Battery Aware and Reliable Beacon Enabled IEEE 802.15.4 (BARBEI)

**Objective:** Tune MAC superframe structure parameters according to battery nonlinear behavior and network status.
**Input:** FFD node *f*, seven RFD nodes $r_1$-$r_7$
**Output:** New superframe structure with updated BO and SO values

**Phase 1:** Tune BO value according to *f* residual capacity.

```
if f send BEACON
  if f. check RESIDUAL_CAPACITY (t (BEACON)) <
  RESIDUAL_CAPACITY (t (BEACON-1)) = true
    if BEACON. BO! = 8
      BO+=1
    endif
  endif
```

**Phase 2:** Tune SO value according to $r_1$-$r_7$ average end to end delay

```
for r1 to r7
  If r check DATA_PACKETS.num %5=true
    r. calculate (DELAY)
  endif
endfor
if f.check AVERAGE_DELAY.new > AVERAGE_DE-
  LAY.prevouis= true
    if BEACON.SO! = BEACON.BO
      SO+=1
    endif
  endif
```

Fig. 2: Battery Aware and Reliable Beacon Enabled IEEE 802.15.4 (BARBEI)

Table 1**: QualNet 5.2 SIMULATIONPARAMETERS**

| **Parameter** | **Value** |
|---|---|
| Physical and MAC | IEEE 802.15.4 |
| Area | 50 m *50 m |
| Energy Model | MICAZ |
| Number of nodes | 8 |
| Transmission range | 10 m |
| Simulation time | 1000 s |
| Battery type | Duracell AA |
| Battery model | Rhakhmatov |
| Traffic | CBR (1s arrival rate) |
| Payload size | 50 byte |
| BO values | 2,3,6,7 |
| SO Values | 2,3,4,6,7,8,9 |

Data rate is fixed for all nodes and the chosen packet interval is 1s for a 50 bytes packet size. 16 scenarios are tested, each one with different BO: SO combination to cover different duty cycles behaviour. Each time the new algorithm performance is compared against the original IEEE 802.15.4 MAC algorithm. Each case is repeated 10 times. Simulation parameters are the same as those presented in Table 1 but with more BO values considered.

The following subsections illustrate the results achieved for the four metrics:

### *D. Total Energy Consumption (mWh):*

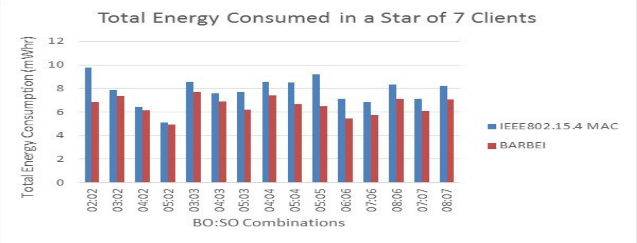


Fig. 3: Total energy consumption for a 7 RFDs in a star topology.

According to Fig. 3, it is apparent that the new algorithm decreases energy consumption regardless of the values in BO:SO combination. This is because the new algorithm tunes the MAC BO parameter according to battery residual charge. BO value is incremented if the current residual capacity is less than the previous one allowing the inactive period to increase. This offers node more time to sleep, which in turn allows PANc battery gain some of charge according to battery recovery effect. Consequently, battery capacity increases providing more energy according to (6). Moreover, increasing BO decreases beacon overhead which will decrease energy consumption, this effect is obviously noticed in BO: SO combination with small BO values, such as BO=2. In addition, energy is saved because the new algorithm avoids packets collision and retransmissions as it increases SO according to the application layer status. This gives nodes more time to transmit packets in the increased active period.

### *E. Residual Battery Capacity (mAh):*

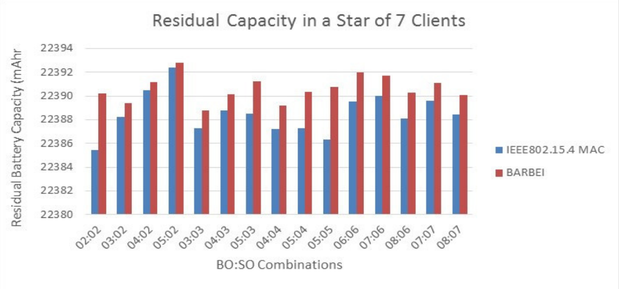


Fig. 4: Residual battery capacity for a 7 RFDs in a star topology

Fig. 4 reveals that despite the duty cycle or BO: SO values, the residual capacity in a network that follows our algorithm is higher than for the network that follows the legacy IEEE 802.15.4 MAC. This is because PANc exploits battery recovery effect by incrementing BO value according to battery status. This allows PANc battery to heal and gain some of its charge which will increase battery residual capacity. Network nodes are also given more time to sleep as BO increases. This saves residual battery capacity for all network nodes. Fig. 5 depicts that the new algorithm increases the throughput at most of BO:SO values. This is mostly obvious in combinations with small BO: SO values, such as 3:2, 4:2, and 5:2. In these scenarios, following the legacy MAC algorithm, the active period is two short causing a node to differ packet transmission to the next superfrme which will cause collision and hence adversely affecting network throughput.

### F. *Throughput (bits/s):*

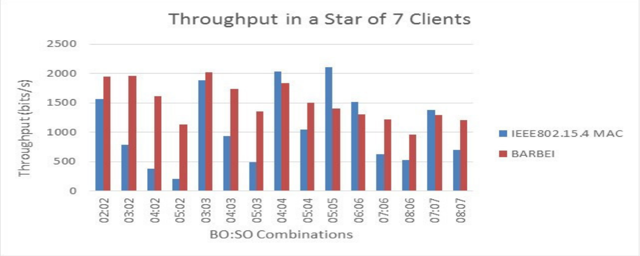


Fig. 5: Throughput for a 7 RFDs in a star topology

However, as the new algorithm allows the SO values to increase according to network performance, this will give more time for RFDs to complete their packet transmissions successfully, and consequently will improve network throughput. For combinations with 100% duty cycle, such as 4:4, 5:5 6:6 and 7:7, the legacy MAC outperforms our algorithm. This is because a node in these situations will have full active period to perform its activities and therefore which increase the throughput. However, increasing the inactive period according to battery status lowers the duty cycle which consequently will decrease the throughput.

### G. *Average End To End Delay (s):*

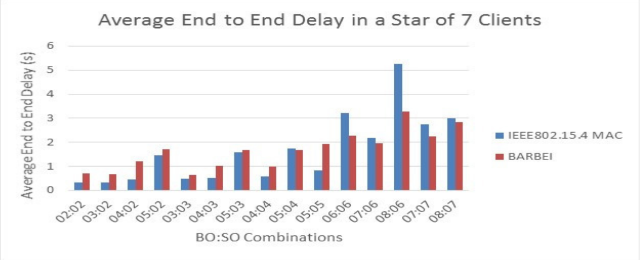


Fig. 6: Average End to End Delay for a 7 RFDs in a star topology

Fig. 6 shows that the average end-to-end delay for our algorithm performs better only for combinations with large BO values, such as BO=6, 7 or 8, because nodes have enough time to do their work and there is no need to increase SO value which avoids the increase of delay. Unfortunately, average end-to-end delay performs worse for the new algorithm for BO: SO combinations with small values such as SO=2, 3, 4 and 5. For small BO:SO values, the delay is always bad, and there will be frequent increments in BO and SO values allowing node to operate in consequent 50% duty cycles which will increase delay.

## V. Conclusion

The IEEE 802.15.4 standard is designed for different types of low-power and low-rate wireless Personal Area Networks. The performance of the standard can be improved by adopting battery-friendly algorithms for packets transmission. This can be achieved by designing battery aware approaches that exploit battery nonlinearity of recovery effects. However, there is a threshold at which battery recovery effect can be exploited. The proposed adaptive cross-layer and battery aware approach improves energy efficiency and power consumption for all possible duty cycle applications that the beacon enabled IEEE 802.15.4 standard offers. As a future work, not only the PANc is allowed to be aware of its battery behaviour, but also all network nodes will tune their activities according to their residual capacity. This can be achieved by taking the priority as criteria for packets transmission. Node priority will be determined according to its residual capacity.

## References

[1] IF. Akyildiz, W. Su, Y. Sankarasubramaniam, and A. Cayirci; "A survey on sensor networks, Communications Magazine ,''Atlanta, GA, USA, vol. 40(8), pp. 102-114, 2002. http://dx.doi.org/ 10.1109/MCOM.2002.1024422

[2] L. Selavo, A. Wood, Q. Cao, T. Sookoor, H. Liu, A. Srinivasan, and J. Porter, "wireless sensor network for environmental research", Proc. The 5th international conference on Embedded networked sensor systems. Sydney, Australia Nov. 2007, pp. 103-116. http://dx.doi.org/10.1145/1322263.1322274

[3] A. Koubaa, "Promoting Quality of Service in Wireless Sensor Networks", Submitted for receiving Habilitation Qualification in Computer Science, National School of Engineering, Sfax, Tunisia, 2011.

[4] SC. Ergen, "ZigBee/IEEE 802.15. 4 (Summary)", [Online][accessed January 2015], Available from URL http://pages.cs.wisc.edu/~suman/courses/838/papers/zigbee.pdf.

[5] M. Salayma, W. Mardini, Y. Khamayseh, and M. Yassein, "Optimal Beacon and Superframe Orders in WSNs," in *Proc. The Fifth International Conference on Future Computational Technologies and Applications (IARIA 2013)*, FUTURECOMPUTING 2013, Valencia, Spain, pp. 49-55, May 2013.

[6] M. Salayma, W. Mardini, Y. Khamayseh, and M. Yassein, "IEEE802. 15.4 Performance in Various WSNs Applications, ''in *Proc. The Seventh International Conference on Sensor Technologies and Applications*, *SENSORCOMM 2013,* conference on Embedded networked sensor systems, Sydney, Australia, pp. 103-116, Nov. 2007. O

[7] M. R. Jongerden and B. R. Haverkort, "Battery modeling,'' Technical report, TR-CTIT-08-01, CTIT, 2008. http://dx.doi.org/ 10.12691/ajmo-3-2-2

[8] Y. Li, , Y. Shouyi, l. Leibo, W. Shaojun and W. Dong, "Battery-Aware MAC Analytical Modeling for Extending Lifetime of Low Duty-Cycled Wireless Sensor Network,'' in *Proc. IEEE 8th Int.Conference Networking, Architecture and Storage (NAS), IEEE,* pp. 297-301, 2013. http://dx.doi.org/ 10.1109/NAS.2013.47

[9] H. Li, Y. Chenfu and L. Ye, "Battery-Friendly Packet Transmission Algorithms for Wireless Sensor Networks, " *Sensors Journal, IEEE 13*, vol. 10, pp. 3548-3557, 2013. http://dx.doi.org/ 10.1109/JSEN.2013.2276617

[10] C. Chau, Q. Fei, S. Sayed, m. Wahab and Y. Yang, "Harnessing battery recovery effect in wireless sensor networks: Experiments and analysis, '' *Selected Areas in Communications, IEEE Journal on 28*, vol. 7, pp. 1222-1232, 2010. http://dx.doi.org/ 10.1109/JSAC.2010.100926

[11] C. Chau, M. Wahab, F. Qin, Y. Wang and Y. Yang, "Battery recovery aware sensor networks", In Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, 2009. WiOPT 2009. 7th International Symposium on, pp. 1-9. IEEE, 2009. Communications, IEEE Journal on 28, vol. 7, pp. 1222-1232, 2010. http://dx.doi.org/ 10.1109/WIOPT.2009.5291623

[12] E. Casilari, J. M. Cano-García and G. Campos-Garrido, "*Modeling of current consumption in 802.15. 4/ZigBee sensor motes,*'' *Sensors*, vol. *10*, pp. 5443-5468, 2010. http://dx.doi.org/ 10.3390/s100605443

[13] M. Di Francesco, G. Anastasi, M. Conti, S. K. Das and V. Neri, "*Reliability and Energy-Efficiency in IEEE 802.15. 4/ZigBee Sensor Networks: An Adaptive and Cross-Layer Approach,* ''IEEE Journal on *Selected Areas in Communications, vol. 29, pp.* 1508-1524, 2011. http://dx.doi.org/ 10.1109/JSAC.2011.110902

[14] D. Linden, and T. B. Reddy, "Handbook of batteries,'' 1985. http://dx.doi.org/*10.1036/0071414754*

[15] D. Rakhmatov, S. Vrudhula and D. A. Wallach, "A model for battery lifetime analysis for organizing applications on a pocket computer. Very Large Scale Integration (VLSI) Systems, '' *IEEE Transactions*, vol. 11, pp. 1019-1030, 2003. http://dx.doi.org/10.1109/TVLSI.2003.819320

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1273–1278

DOI: 10.15439/2015F306
ACSIS, Vol. 5

# Influence of on-Device Measurement Analysis on Energy Efficiency in Machine-to-Machine Systems

Pavle Skocir, Mario Kusek and Gordan Jezic
University of Zagreb
Faculty of Electrical Engineering and Computing
Department of Telecommunications
Unska 3, HR-10000 Zagreb, Croatia
{pavle.skocir, mario.kusek, gordan.jezic}@fer.hr

***Abstract*—Machine-to-Machine Communication (M2M) enables communication between heterogeneous devices without human intervention. It is considered to be a key enabler technology for the concept of Internet of Things (IoT) and Cyber Physical Systems (CPS). With M2M's integration with Wireless Sensor Networks (WSN), information from different kinds sensors can be obtained. In order to discover useful knowledge from sensor data, various data mining techniques need to be applied. Due to the development of microprocessors on end devices in M2M system which collect data from sensors, data processing can also be executed on those devices. However, since end devices are often battery powered, energy consumption when running those algorithms needs to be taken into account. In this paper we implement an algorithm in M2M system, on Libelium Waspmote devices, which detects temperature plummeting in an indoor space. Afterwards, energy consumption of Waspmote devices is analyzed for two cases: when algorithm is executed on-device and when algorithm is executed on gateway or on back-end system.**

## I. Introduction

MACHINE-to-Machine Communication (M2M), a concept which enables connection of heterogeneous devices with limited human intervention, is considered to be one of the enablers for the process of provisioning advanced applications and services, such as smart cities and hospitals, automated vehicular and industrial operation, along with others [1]. Through integration with Wireless Sensor Networks (WSN), M2M systems can obtain wide range of information [2]. By analyzing that information, useful knowledge can be discovered, and appropriate actions can be initiated. M2M is considered to be one of the fundamental technologies for enabling the concept of Internet-of-Things (IoT) and Cyber Physical Systems (CPS) [3]. The IoT concept includes connecting sensors and other devices to the broader Internet by using general Internet technologies [4]. CPS is considered as evolution of M2M which supports more intelligent and interactive operations, under the architecture of IoT [3]. Although interlaced with the aforementioned areas, in this paper we use the term of M2M systems for a sensing systems in which end devices collect measurements from sensors, and send them via gateway to the back-end system.

M2M systems generate massive data sets which are considered to be of high business value [5]. To extract hidden knowledge from data, data mining algorithms can be applied. For instance, by analyzing sensor measurements collected within a smart home, the system can detect actions of the inhabitants and predict their future behavior. Additionally, it can recognize outliers, events which are not within usual patterns, and indicate towards a potential problem [6].

Most of the existing M2M solutions incorporate a central point for collecting and analyzing information [7]. However, with the development of hardware technologies which enabled miniaturizing wireless devices, smart sensors or actuators and micro-controllers, and enhancing their processing power, new schemes for refining software for embedded systems started to evolve [6]. The example of those new schemes is the possibility for certain algorithms to be executed on end devices or on gateways, instead of only on back-end system [8].

In this paper we analyze data mining techniques in M2M networks. Furthermore, we want to disclose how the functionalities of nodes within M2M system architecture influence energy efficiency of the system with limited energy resources. Particularly, we compare energy consumption of the battery-powered end devices when end device has measurement analysis functionality (i.e. analysis is performed on-device) and when measurement analysis functionality is performed elsewhere (i.e. when analysis is performed on gateway or on back-end system).

Section 2 presents research activities within the area of interest of this paper. Section 3 describes the network architecture compliant with existing M2M standards in which we conduct our measurements. In section 4 we introduce an algorithm which was deployed on end devices of our system and which monitors temperature fluctuations in an indoor space. Section 5 presents energy consumption comparison of end devices for two cases - when algorithm was executed on end devices, and when it was executed elsewhere. Section 6 concludes the paper and gives an outline for future work.

## II. Related work

This section presents current research efforts in the area of applying data processing techniques within M2M network. Stojmenovic [7] considers M2M as a key enabling technology for the CPSs. The author identifies the problem that existing work in the area of M2M communication is based on small-scale M2M models and centralized solutions, while a few existing distributed solutions do not scale well. A paradigm

978-83-60810-66-8/$25.00©2015, IEEE

shift is suggested where end nodes should also make decisions based on local knowledge, instead of only forwarding collected messages to back-end system. By using this new paradigm, M2M solutions could scale to a significantly larger number of M2M devices. In the use-case example, the author considers a smart building control application in which sensors and actuators exchange information directly, without communication with servers, and coordinate by using distributed decision making to react to data. This reactions include opening of windows or injecting fresh air when needed.

Chen et al. [5] introduce an overview of data mining techniques in the area of Internet of Things. They present five data mining functionalities: classification, clustering, association analysis, time series analysis, and outlier analysis. The application areas of data mining are also presented: e-commerce, banking, retail, health care, and city governance. The main research issues are identified in the area of finding erroneous data, analysis of data streams and developing a framework to support big data mining. As for the nodes where data processing is performed, the authors suggest servers in the cloud where open source solutions like Hadoop, HDFS, Storm or Oozie can be used. According to their perception, end devices are used only for data gathering and forwarding towards cloud.

Bruns et al. [9] state that in traditional M2M systems, data processing is usually hard-coded and scattered all over source code, which makes it difficult to maintain. Therefore, they propose a complex event processing system (CEP) which separates event processing from source code. In CEP, event processing is capsulated in rules which can be efficiently adapted and maintained. The authors discuss the application of the proposed system in solar power plants and printer supply and maintenance service. CEP is implemented on M2M server, it processes data streams just as new data arrives to the server.

M2M systems where end devices are serving as nodes which collect measurements from sensors are referred to as Wireless Sensor Networks (WSN). Analytics in WSN is also an ongoing research topic closely connected with the concept of M2M systems. Mahmood et al. [10] propose a taxonomy of data mining techniques for WSN. First level of classification is connected with general data mining classes: frequent pattern mining, sequential pattern mining, clustering and classification. Second level of classification is based upon the ability of the approach to process data in a centralized or distributed way. The third level of classification is determined according to the focus on two different aspects - WSN performance issues and application issues. The approaches which focus on WSN performance issues try to take into account resource constraints like energy, memory, and communication bandwidth. On the other hand, approaches which focus on application issues try to satisfy application requirements without much consideration for WSN performance. The authors group existing data mining techniques for WSN according to the presented taxonomy. Moreover, they also discuss what is the focus of this paper - on which nodes to execute different data mining techniques. Sensor nodes, referred to as M2M devices in M2M system architecture, perform single pass algorithms and forward only the required and partially processed data to the network. In network, referred to as M2M gateway, data from various end devices is collected, and activities as network pattern identification are carried out. On sinks, reffered to as M2M servers, computationally demanding tasks are executed.

Alsheikh et al. [11] present an overview of existing machine learning techniques used in WSNs. They group them into the categories of supervised, unsupervised and reinforcement learning. As important aspects which need to be taken into account when deploying machine learning techniques in WSNs, the authors emphasize power and memory constraints of sensor nodes, topology changes, communication link failures and decentralized management. The main functional challenges for which machine learning techniques were adopted include routing in WSN, clustering and data aggregation, event detection and query processing, localization and objects targeting, and medium access control. As for the nodes where processing is taking place, the authors also consider in-network processing of data since it enables the nodes to rapidly adapt their future behavior and predictions in correspondence with current environmental conditions. However, when executing learning algorithms on-device, special attention needs to be payed not to exhaust the nodes with complex and resource demanding computational tasks.

Suryadevara et al. [6] developed a smart home solution which enables identification of the Activities of Daily Living (ADL) in order to determine the wellness of elderly people. This was done by processing time series of data collected by the sensors deployed at users' homes. Among other things, the authors discuss different storage mechanisms for WSN data. They identify two approaches for storing data in WSNs - centralized and decentralized way. In centralized approach, data is stored and can be analyzed on a node which generates it. In decentralized approach, data is stored on different nodes. The most common decentralized storage approach is data centric storage, where data is stored on a node called sink. Centralized approach is identified as not appropriate for a setup with recurrent bursts of activities since it quickly overfills memory resource. Moreover, when sequences gathered from different sensors need to be processed, as is the case in the solution developed by the authors, data-centric storage appears to be a better solution. Therefore, the authors developed a system which stores sensor data in the form of event activities on a central system which then analyses that data and makes assumptions of ADLs.

Research efforts analyzed within this section focus on presenting current trends in the area of data processing in M2M/IoT/WSN domain. However, we have not identified solutions which would take into consideration optimal position in the network to perform data processing with regard to energy efficiency of the system with limited energy resources.

## III. M2M Architecture

The architecture of our M2M system is shown in Figure 1, and is compliant with functional M2M specification from oneM2M[1] [12] and high-level M2M system architecture by ETSI[2] [13].

Application Dedicated Node (ADN) contains sensors which collect measurements from their physical environment. It contains at least one Application Entity (AE), which is an entity in the application layer that implements M2M application service logic. Examples of these application entities are power metering application or remote blood sugar monitoring application. ADN does not contain Common Services Entity (CSE) which is situated on Middle Node (MN) or Infrastructure Node (IN). CSE is a set of common service functions of the M2M environment, like data management, device management, subscription management and location services. Middle node contains CSEs and can contain one or more AEs. When ADN contains CSE, it is called Application Service Node (ASN). ADN/ASN can communicate directly with IN, or can communicate with IN via MN. In our case scenarios, ADN and ASN communicate via MN. ADN/ASN, MN, IN, defined in oneM2M standard [12], can be referred to as M2M Device, M2M Gateway and M2M Server respectively in previous standards by ETSI [13].

In this paper we focus on M2M communication within smart homes which can include following sensors: temperature, humidity, luminosity, presence, hall effect, electricity consumption or water metering. Based on the data collected by these sensors, various events can be detected, for instance if residences are present in the house, if temperature fluctuates unnecessarily (e.g. a window is opened too long during cold winter or hot summer days) and should be controlled, if power is more affordable in certain times of day and when should specific consumers with variable operating times like washing machine be turned on etc. Moreover, the system can detect outliers and report them to the interested users, like water leakage, extreme temperatures or unusual power consumption. Our system is composed of ADNs/ASNs which monitor temperature fluctuations and an MN through which the devices communicate with IN.

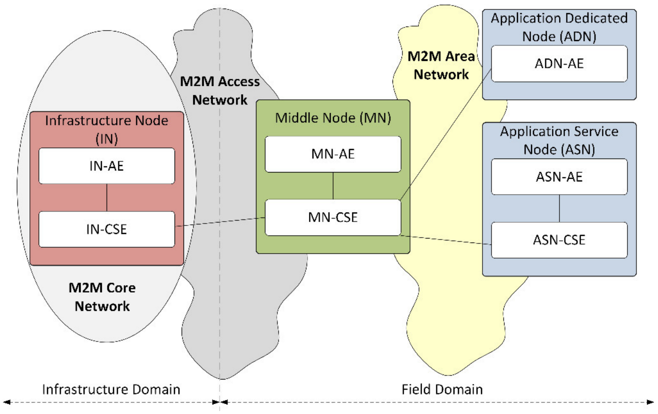


Fig. 1: M2M Architecture

The easiest way to perform the analysis of data collected by sensors is to transfer it to infrastructure domain where servers have plenty of memory and processing power [14]. However, due to rapid development of embedded devices which have more and more processing power and due to application needs for faster response, processing can also take place in field domain, on MNs and ADNs/ASNs [11], [15]. When executing in-network data processing, special attention needs to be paid not to exhaust the nodes with complex computational tasks. By analyzing data in networks, valuable information instead of raw data is delivered to infrastructure domain, from where it can be easily accessed by user applications.

In the first case scenario of our energy consumption measurement process, explained in detail in Section V, end device is according to oneM2M specification an ASN because it contains both AE and partially CSE - data management. In the same scenario, gateway is MN with only CSE functionality. In the second case scenario, end device is an ADN since it contains only AE, while CSE - data management and other functionalities, as well as a part of Application Entities (AE) is executed on MN (gateway).

## IV. Outlier Detection Algorithm

Temperature function shown in Figure 2 recorded temperature values at our laboratory during one week in winter from February $23^{rd}$ until March $1^{st}$. Vertical dotted lines separate measurements belonging to a certain day of the week. The measurements were obtained from sensors and forwarded to back-end system every 10 seconds. Three peaks can be identified, which are marked on Figure 2 and which occurred when the window was opened, and no-one was present in the room. In these cases, temperature fell to around 12°C. The algorithm that we propose has its main goal to raise the alarm when temperature plummets like in those 3 cases. The reason for that is because in such occasions the temperature in the room was too low to reside there. By raising the alarm, window could be closed sooner to prevent those unpleasant conditions.

In order to determine the algorithm which could detect rapid temperature fall, which we define as an outlier from normal temperature fluctuation during the day, we monitored the derivative of the temperature function. Since the temperature function is discrete, its derivative, often referred to as backward difference, is calculated by using the following expression [16]:

$$\Delta_n f(n) = f(n) - f(n-1) \qquad (1)$$

where $f(n)$ is a current temperature value measured by end device, while $f(n-1)$ is the value measured in the last measurement. In our case, the difference between those two consecutive measurements is 10 seconds. By calculating the difference of the function, we wanted to analyze how does the difference of the function behave when the temperature plummets. When analyzing the difference between two consecutive

[1] oneM2M (http://www.onem2m.org/) - organization which develops standards for M2M and the Internet of Things

[2] ETSI (http://www.etsi.org/) - European Telecommunications Standards Institute

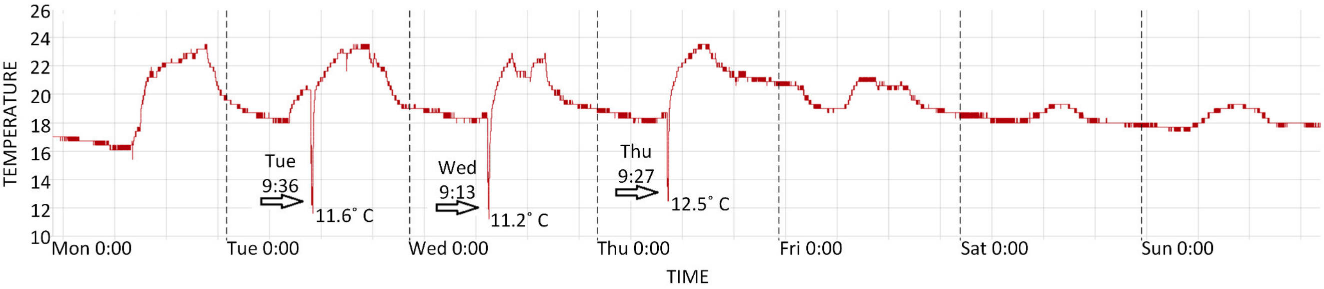


Fig. 2: Measurements from temperature sensor during 1 week in winter

values, extreme descends as the ones seen in Figure 2 cannot be detected since in those events the temperature does not fall in every interval. However, if we monitor the difference in larger time window, which is a good approach when dealing with data that arrives continuously in time [17], rapid decrease can be detected when comparing current measurement with the 5th historic measurement. The equation of such differences calculation is the following:

$$\Delta_n f(n) = f(n) - f(n-5) \qquad (2)$$

where $f(n)$ is current measured value, and $f(n-5)$ is 5th historical measurement, measured 50 seconds ago. We have tried to use more recent historic measurements (1st-4th), but in those cases certain temperature decreases were captured which were not so steep as the ones we wanted to identify. When running an algorithm which analyzes when difference values calculated as in Equation 2 are falling in 10 more than consecutive intervals, extreme descends in temperature like the ones marked in Figure 2 can be detected. The number of falling intervals was set according to empirical evaluation. Table I shows the number of falling intervals for the cases when temperature plummeted. The times of these events were 9:36 on Tuesday, 9:13 on Wednesday, and 9:23 and 9:27 on Thursday. The reason for two captured intervals on Thursday is because temperature was falling for a longer period on that day. The temperature for that case did not fall in every interval, in some intervals it remained unchanged or even slightly grew. Since the algorithm was analyzing constant falls of differences from Equation 2, this particular outlier was captured twice. If the number of intervals in which descend is monitored is between 7 and 10, those rapid falls can always be detected. However, if the number of falling intervals is smaller than 7, then some other events can be captured, which are not interesting to us, like the fall of temperature during night when the heating was off. Since the algorithm in which 10 consecutive falling intervals are detected fits to our needs (it identified rapid temperature fall caused by the opened window), it was implemented on our M2M system in order to monitor energy consumption of end devices in different case scenarios described in detail in the next section.

TABLE I: Captured outliers when looking for negative difference in more than 10 consecutive measurements

| time | number of falling intervals |
|---|---|
| Tue 24.2.2015 9:36:04 | 12 |
| Wed 25.2.2015 9:13:01 | 10 |
| Thu 26.2.2015 9:23:00 | 15 |
| Thu 26.2.2015 9:27:08 | 10 |

## V. Energy consumption analysis

The proposed algorithm described in Section IV was implemented on Libelium Waspmote devices v1.2 [18] with XBee communication modules to monitor energy consumption of end devices for two case scenarios: when the algorithm is executed on end devices and when algorithm is executed on gateway or on back-end system. Since device consumes different amounts of energy by executing different tasks during its operating cycle, it was necessary to identify those tasks, measure its duration and power consumed during the execution. Figure 3 shows the tasks and their order of execution for both scenarios.

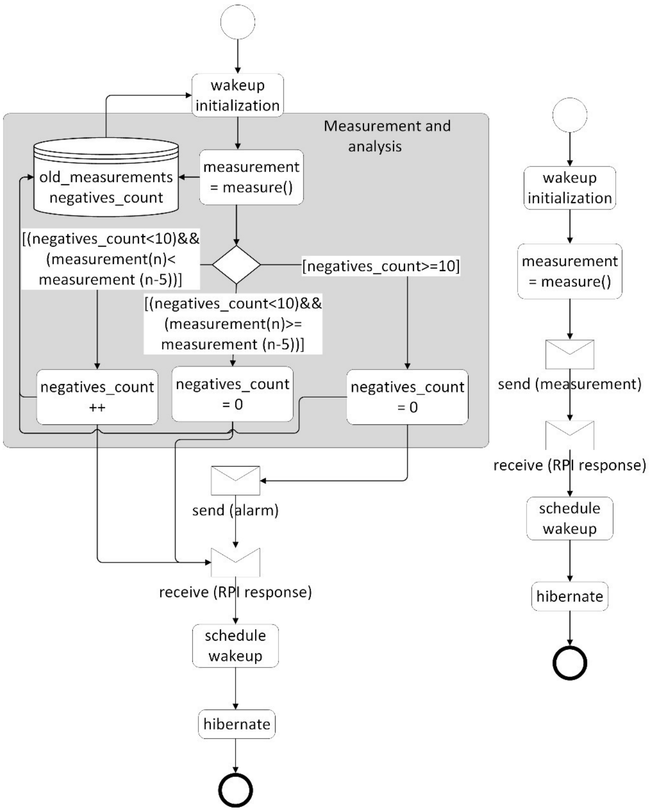


Fig. 3: Activity diagram of end device

In each operating cycle, device wakes up from hibernate mode, which is a state on Waspmote device where lowest amount of energy are consumed. Afterwards, it obtains measurements from sensors. In the first case scenario, shown on the left-hand side of Figure 3, that task is followed by executing the algorithm to detect rapid temperature decrease. When the overseen decrease is identified, the device reports it by sending a message to the gateway. After reporting the alarm message, it waits for a certain amount of time to receive a new message from gateway which contains information for a new task, e.g. if measurements are to be sent in a specified interval instead of forwarding only alarms. If no alarm needs to be raised, the device only waits for a message from the gateway. After the time for receiving message has expired, the device is ready to return to hibernate mode. The process of exchanging messages between end devices and gateway in order to reach an agreement about operating times is described in our previous work [19].

In the second case scenario, shown on the right-hand side of Figure 3, after wake up from hibernate mode and obtaining a measurement from sensor, the end device forwards that data to gateway. Since end device does not execute algorithm for detecting temperature decrease, acquired measurement needs to be sent to back-end system for further analysis. As in the first case scenario, the end device then waits for messages from gateway. After the eventual processing of the received message from gateway, end device goes back to hibernate mode.

Table II shows power consumption and duration of tasks executed on end devices when analysis is performed on-device (ASN functionality), while Table III shows power consumption and duration of tasks executed on end devices when analysis is performed on gateway or back-end system (ADN functionality of devices). Current consumption and voltage levels are measured by using Rigol DS1102D oscilloscope[3].

TABLE II: Power consumption and duration of tasks in the case of on-device analysis

| state/task | wake-up initialization | measure | analyze | send | receive | handle response |
|---|---|---|---|---|---|---|
| power consumption (mW) | 70,13 | 130,83 | | 258,95 | 262,68 | 141,12 |
| duration (ms) | 80 | 70 | | 255,3 | 208,8 | 1 |

TABLE III: Power consumption and duration of tasks when analysis is not performed on-device

| state/task | wake-up initialization | measure | send | receive | handle response |
|---|---|---|---|---|---|
| power consumption (mW) | 70,13 | 77,7 | 258,95 | 262,68 | 141,12 |
| duration (ms) | 80 | 10 | 255,3 | 208,8 | 1 |

Figure 4 shows the energy consumption for one operating cycle. The left-hand and central columns represent energy consumption for the first case scenario when analysis is performed on-device. Since the end device in that scenario needs to raise an alarm, it sends a message to the gateway only when temperature value plummets. The consumption of that operating cycle is shown in the left-hand column. When rapid temperature decrease is not detected, the device does not send any data to gateway. The consumption of such operating cycle is shown in the central column. The consumption of the operating cycle in second case scenario, when end device reads the value from sensor and forwards it to the gateway, is shown in the right-hand column of Figure 4.

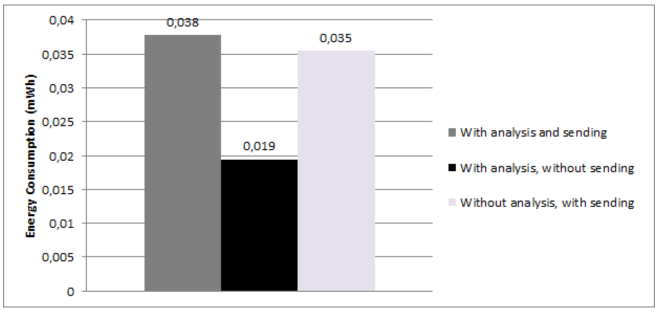


Fig. 4: Energy consumption of end devices during one operating cycle

It can be observed that energy consumption presented in the left-hand and right-hand columns is similar. It appears that on-device analysis does not add much to total energy consumption of the device in one operating cycle. However, when comparing the power consumption for measurement and analysis task in Table II - 130.8 mW and only measurement task in Table III - 77.7 mW, it can be observed that for the scenario when data is analyzed, the power consumption is around 70% higher. But since both actions last for quite a short time (0,07 s and 0,01 s), they do not have a large influence on the overall energy consumption. On the other hand, energy consumption shown in the central column is about 50% lower than in two aforementioned cases. The reason for that lies in the fact that sending data has a high influence on overall energy consumption. Although it consumes only twice as much power compared to measuring and processing, it lasts around 3.5 times longer.

Total energy consumption during one week for the two cases when analyzing data presented in Figure 2 is shown in Figure 5. Left-hand column shows consumption for the first case scenario when analysis is performed on-device. In that scenario, in most of the operating cycles the device was only performing measurements and analysis without sending alarm message to the gateway due to the fact that on the

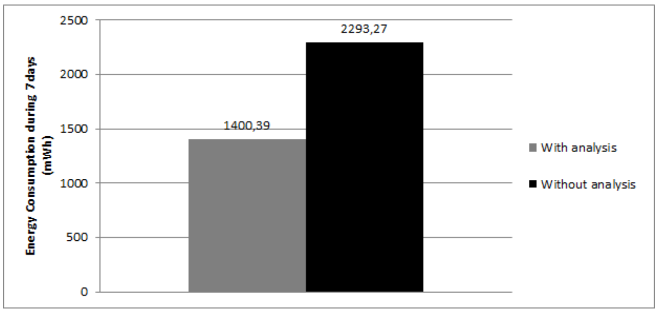


Fig. 5: Energy consumption of end devices during 1 week

[3] http://www.rigolna.com/products/digital-oscilloscopes/ds1000d/ds1102d/

analyzed data set alarms were needed to be raised only on four occasions. Since operating cycle which includes analysis without sending consumes around 45% less energy than the operating cycle in which the measurement data is forwarded to back-end system, the total consumption of the first case scenario is around 40% less than the consumption of the second case scenario which is shown in the right-hand column in Figure 5 and in which data was sent during each operating cycle.

As a power source for end devices used in this experiment we had at our disposal batteries with capacity of 6600 mAh. According to energy consumption shown in Figure 5, and by taking into account that the average voltage of the battery is 3.7 V, by using the first case scenario, device can operate for 17.4 weeks. By using the second case scenario, the device can remain operational for 10.6 weeks. For this particular application and data set, on-device analysis extends end-device lifetime for about 65%.

## VI. Conclusion

This work presented the analysis of the influence of data measurement analysis on energy efficiency in Machine-to-Machine system with Libelium Waspmote v1.2 devices. By implementing a specific outlier detection algorithm on end devices, it was shown that on-device data analysis when using this algorithm does not consume as much energy as communication. More energy can be saved by reducing communication. As a result, it can be advisable to perform local data analysis on end devices in those cases when the duration of the measurement and analysis tasks is shorter than the duration of communication tasks.

If the data set were different and temperature plummeting occurred more often, the difference in total energy consumption during one week would have been smaller. However, since temperature outliers do not occur often, the scenario in which the measurement analysis is performed on-device will usually spend less energy than the scenario where measurement analysis is performed on gateway or on back-end system.

In future work, we plan to compare energy efficiency for more sophisticated algorithms in smart home environment which would enable event detection. We also plan to extend this research to different types of applications, not only the applications which should raise alarms, but applications which require continuous or periodical streaming.

## Acknowledgment

This work was supported by strategic project ”Energy Efficient M2M Device Communication” in cooperation with FTW - Forschungszentrum Telekommunikation Wien GmbH, Austria and by Croatian Science Foundation under the project 8065.

## References

[1] R. Ratasuk, A. Prasad, Z. Li, A. Ghosh, and M. Uusitalo, “Recent advancements in M2M communications in 4G networks and evolution towards 5G,” in *Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on*, 2015, pp. 52–57. doi: http://dx.doi.org/10.1109/ICIN.2015.7073806

[2] J. Zhang, L. Shan, H. Hu, and Y. Yang, “Mobile cellular networks and wireless sensor networks: toward convergence,” *Communications Magazine, IEEE*, vol. 50, no. 3, pp. 164–169, 2012. doi: http://dx.doi.org/10.1109/MCOM.2012.6163597

[3] J. Wan, H. Yan, Q. Liu, K. Zhou, R. Lu, and D. Li, “Enabling cyber-physical systems with machine-to-machine technologies,” *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 13, no. 3/4, pp. 187–196, 2013. doi: http://dx.doi.org/10.1504/IJAHUC.2013.055454

[4] J. Hller, V. Tsiatsis, C. Mulligan, S. Karnouskos, S. Avesand, and D. Boyle, *From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence*. Academic Press, 2014. ISBN 978-0-12-407684-6

[5] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, “Data mining for the internet of things: Literature review and challenges,” *International Journal of Distributed Sensor Networks*, in press.

[6] N. K. Suryadevara and S. C. Mukhopadhyay, *Smart Homes: Design, Implementation and Issues*. Springer, 2015. ISBN 978-3-319-13556-4

[7] I. Stojmenovic, “Machine-to-machine communications with in-network data aggregation, processing, and actuation for large-scale cyber-physical systems,” *Internet of Things Journal, IEEE*, vol. 1, no. 2, pp. 122–128, 2014. doi: http://dx.doi.org/10.1109/JIOT.2014.2311693

[8] I. Stojmenovic and S. Wen, “The fog computing paradigm: Scenarios and security issues,” in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, vol. 2. IEEE, 2014, pp. 1–8. doi: http://dx.doi.org/10.15439/2014F503

[9] R. Bruns, J. Dunkel, H. Masbruch, and S. Stipkovic, “Intelligent M2M: Complex event processing for machine-to-machine communication,” *Expert Systems with Applications*, vol. 42, no. 3, pp. 1235 – 1246, 2015. doi: http://dx.doi.org/10.1016/j.eswa.2014.09.005

[10] A. Mahmood, K. Shi, S. Khatoon, and M. Xiao, “Data mining techniques for wireless sensor networks: A survey,” *International Journal of Distributed Sensor Networks*, vol. 2013, pp. 1–24, 2013. doi: http://dx.doi.org/10.1155/2013/406316

[11] M. Abu Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, “Machine learning in wireless sensor networks: Algorithms, strategies, and applications,” *Communications Surveys Tutorials, IEEE*, vol. 16, no. 4, pp. 1996–2018, 2014. doi: http://dx.doi.org/10.1109/COMST.2014.2320099

[12] oneM2M, “M2M Functional Architecture,” Technical Specification, draft, 2015. [Online]. Available: http://www.onem2m.org/images/files/deliverables/TS-0001-Functional_Architecture-V1_6_1.pdf

[13] ETSI, “Machine-to-Machine communications (M2M); Functional architecture,” Technical Specification ETSI TS 102690 V2.1.1, 2013. [Online]. Available: http://www.etsi.org/deliver/etsi_ts/102600_102699/102690/02.01.01_60/ts_102690v020101p.pdf

[14] S. Kitagami, M. Yamamoto, H. Koizumi, and T. Suganuma, “An M2M Data Analysis Service System Based on Open Source Software Environments,” in *Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on*, 2013, pp. 953–958. doi: http://dx.doi.org/10.1109/WAINA.2013.124

[15] Cisco Systems, “Proposed Computation and Analytics Use Case,” Input contribution, 2013. [Online]. Available: ftp.onem2m.org/Meetings/REQ/2013%20meetings/20130225_REQ-ARC19_REQ20_San%20Francisco/oneM2M-REQ-2012-0102R01-Analytics_for_oneM2M.DOC

[16] B. Hamrick, “Discrete calculus.” [Online]. Available: http://homepages.math.uic.edu/ kauffman/DCalc.pdf

[17] N. C. Krishnan and D. J. Cook, “Activity recognition on streaming sensor data,” *Pervasive Mob. Comput.*, vol. 10, pp. 138–154, 2014. doi: http://dx.doi.org/10.1016/j.pmcj.2012.07.003

[18] Libelium Comunicaciones Distribuidas S.L., “Waspmote,” Technical Guide, 2013. [Online]. Available: http://www.libelium.com/uploads/2013/02/waspmote-technical_guide_eng.pdf

[19] M. Kusek, I. Lovrek, and H. Maracic, “Rich presence information in agent based machine-to-machine communication,” *Procedia Computer Science*, vol. 22, no. 0, pp. 321 – 329, 2013. doi: http://dx.doi.org/10.1016/j.procs.2013.09.109

DOI: 10.15439/2015F172
ACSIS, Vol. 5

# Design of Smart Dust Sensor Node for Combustible Gas Leakage Monitoring.

Denis Spirjakin,
Alexander M. Baranov,
Vladimir Sleptsov
MATI – Russian State Technological
University Moscow, Russia
Email: denis.spirjakin@gmail.com

***Abstract*—In this work we present the results of design of smart dust sensor platform for combustible gas leakage monitoring. During the design process we took into account a lot of combustible gas sensor specific problems such as their huge power consumption, the necessity to work in explosive environment and sensor parameters degradation. To decrease power consumption we designed specific energy efficient algorithms for measurements. The resulting average power consumption of the node is low enough for one year autonomous lifetime. The methods and algorithms which was designed are very promising for catalytic combustible gas sensors.**

## I. Introduction

SMART dust is a term which was introduced in 2001 by Kristopher Pister and it is related to tiny devices for environmental monitoring with self-organization feature. Such devices should be able to measure different physical parameters, to process data and to send it to the end user through wireless data transmission channel. The main functional block of the "smart dust" is a sensor node which consists of microcontroller, memory, wireless transceiver, power source and one or several sensors (Fig. 1).

The power source of the sensor node is typically batteries. However, in the last time several platforms were designed which use supercapacitors, alternative power sources or their combination. Sensor node generally use microcontrollers which combine enough computational capability and performance and low power consumption. All these functions are very important for autonomous "smart dust". Wireless transceivers very often comply IEEE 802.15.4 standard and ZigBee specification. Such devices are low power and are able to transmit data for the distances up to 30-50 meters. The most popular sensors are temperature and humidity sensors and accelerometers. They are widely used for "smart dust" because of their low power consumption.

During design process of wireless sensor network for gas monitoring it is necessary to take into account that combustible gas sensors consume power up to several hundreds milliwatt. At the same time it is necessary for "smart dust" sensor nodes to have as long autonomous lifetime as it is possible. To increase sensor node autonomous lifetime, developers strive to select electronic components with low power consumption, to use energy harvesting technologies [1], [2], to equip sensor nodes with power sources with huge capacity [1], [3] or to use "smart" algorithms and methods [4], [3], [13] to decrease power consumption.

In this work we present the results of design of "smart dust" sensor node for combustible gas leakage monitoring as well as the results of design of energy efficient measurement algorithms for power consumption decrease.

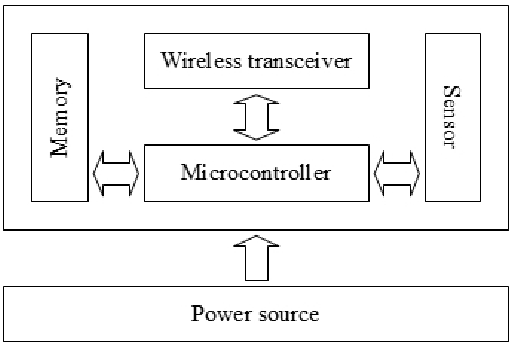


Fig 1. Typical structure of the "smart dust" sensor node.

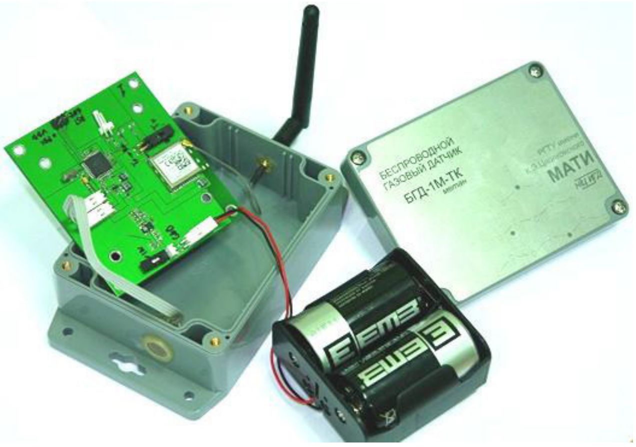


Fig 2. The smart dust gas sensor node prototype.

978-83-60810-66-8/$25.00©2015, IEEE

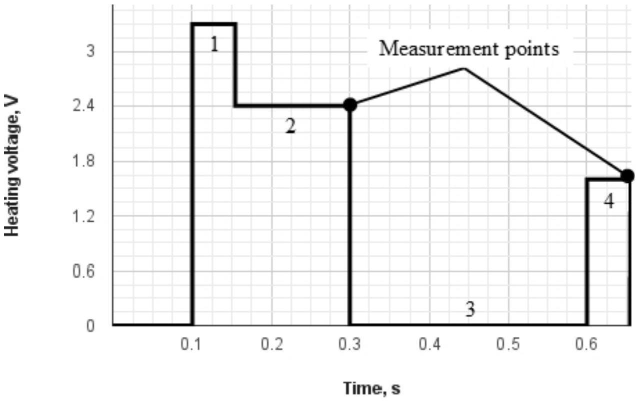


Fig 3. Multistage pulse heating profile.

At present time there are a lot of "smart dust" platforms. Ones of the firsts platforms are TelosB [5] and the family of modules MICA (MICA/MICA2/MICAz) [6], [7]. These platforms used light, temperature and humidity sensors, but end user was be able to add almost every kind of sensors through available connectors and interfaces.

After that commercial modules took their place [8], [9]. With that, typical sensor node became modular. The platform now is a motherboard and it is possible to connect other boards to it (with sensors, wireless transceivers and data processing modules). The main philosophy of these platforms is to provide every possible configuration of the sensor node to the end user, both from software and hardware points of view, using all platform possibilities.

However, all available sensor nodes do not take into account all problems which are related with using wireless sensor networks for combustible gas leakage monitoring. Particularly, such problems as huge power consumption of catalytic, semiconductor and optical sensors, the necessity to work in explosive environment and sensor parameters degradation.

The smart dust sensor node (Fig. 2) which we present in this work is able to work with catalytic combistible gas sensors.

Heating-up of the sensor to working temperature is performed using PWM-controlled voltage. With that, sensor temperature is controlled using complex algorithm. Sensor response measurements are performed in different temperature points. And these sensor response measurements are used to calculate combustible gas concentration.

## II. Experimental Details

In this paper we used the commercial catalytic sensors manufactured by NTC IGD, Russia. The sensor box height is 9.5 mm and diameter 9 mm, power consumption is 200 mW in the continuous measurement mode. Its relatively low power consumption (compare to Figaro, Nemoto and Hanwei) is achieved by applying a heater implemented as 10 µm platinum micro wire in glass insulation (2 µm). The active sensor has a platinum micro wire covered by porous gamma alumina oxide material that is used as catalyst support for catalytically active metals (mixture of Pd and Pt). In order to impregnate the catalyst support by the catalytic metal, salts of palladium chloride ($PdCl_2$) and platinum acid ($H_2PtCl_6$) are used. After annealing at 500 C, noble metal clusters are formed in the catalyst support.

Circuits for gases detection with catalytic sensors are commonly based on the Wheatstone bridge, which includes two resistors and two sensors, one active and one for reference. Most of the power goes into the sensor heating process (about 450 C for methane detection), required to perform the measurement. The power consumption is about 200 mW for Wheatstone bridge and that’s much higher than suitable for wireless application.

Excluding the reference sensor decreases power consumption. But at the same time it’s necessary to compensate atmosphere humidity and temperature which usually performed by reference sensor. This compensation was made by applying the specific multistage heating pulse. This method was offered and discussed in works [4], [3], [14].

Here we use four stages for every multistage pulse (1-4 regions in the diagram, Fig. 3).

The first and second stages provide the sensor heating to the external diffusion region of catalysis and the partial evaporation of surface water (~450 C). These stages are followed by a pause (third stage) during which the surface is

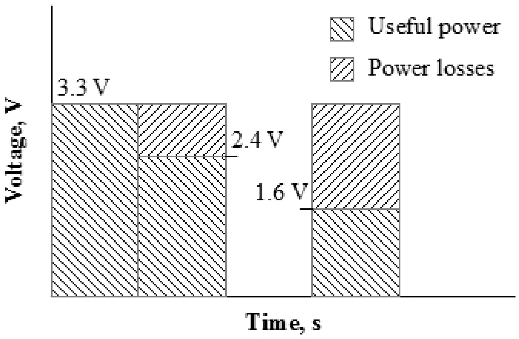


Fig 4. Useful and wasted power during multistage heating pulse in DAC circuit design.

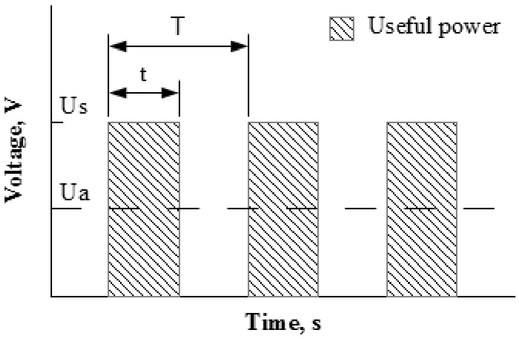


Fig 5. Pulse-width modulation parameters.

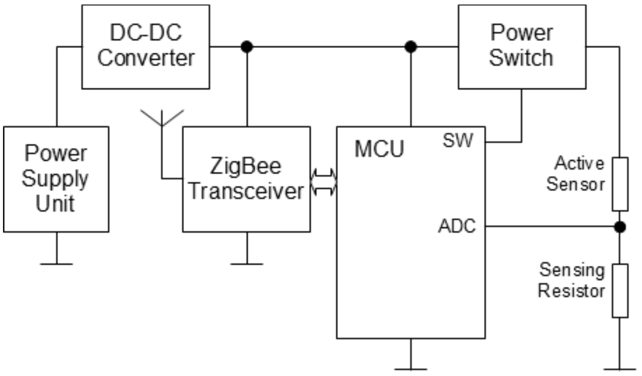


Fig 6. Block diagram of the wireless gas sensor node.

not heated. The final fourth stage heats the sensor to the beginning of the kinetic region of catalysis (~200 C). After this pulse, the element cools down to ambient temperature. Heating voltages for the stages are 3.3 V, 2.4 V, 0 V and 1.6 V respectively.

The measurement result is the difference between sensor voltages at two different temperatures (measurement points in the diagram).

Multistage heating pulse is formed by applying different levels of voltage to sensing circuit. Traditionally it can be done using DAC with buffer amplifier. But in this case relatively huge amount of power is wasted by means of power dissipation on buffer amplifier (Fig. 4). Another way is to generate these levels using pulse-width modulation.

Pulse-width modulation (PWM) is frequently used to heat the sensors for working temperature [10] because it allows one to change the average voltage with fixed voltage from power supply.

Pulse-width modulation is the method of regulation of average voltage value on the load by controlling pulse duty ratio.

The main parameters of PWM are its period (T) or frequency, pulse width (t), supply voltage ($U_S$) and average voltage ($U_A$) on the load. These parameters are illustrated in Fig. 5.

The average voltage of PWM can be solved using following equation

$$U_A=\frac{t}{T}U_S \tag{1}$$

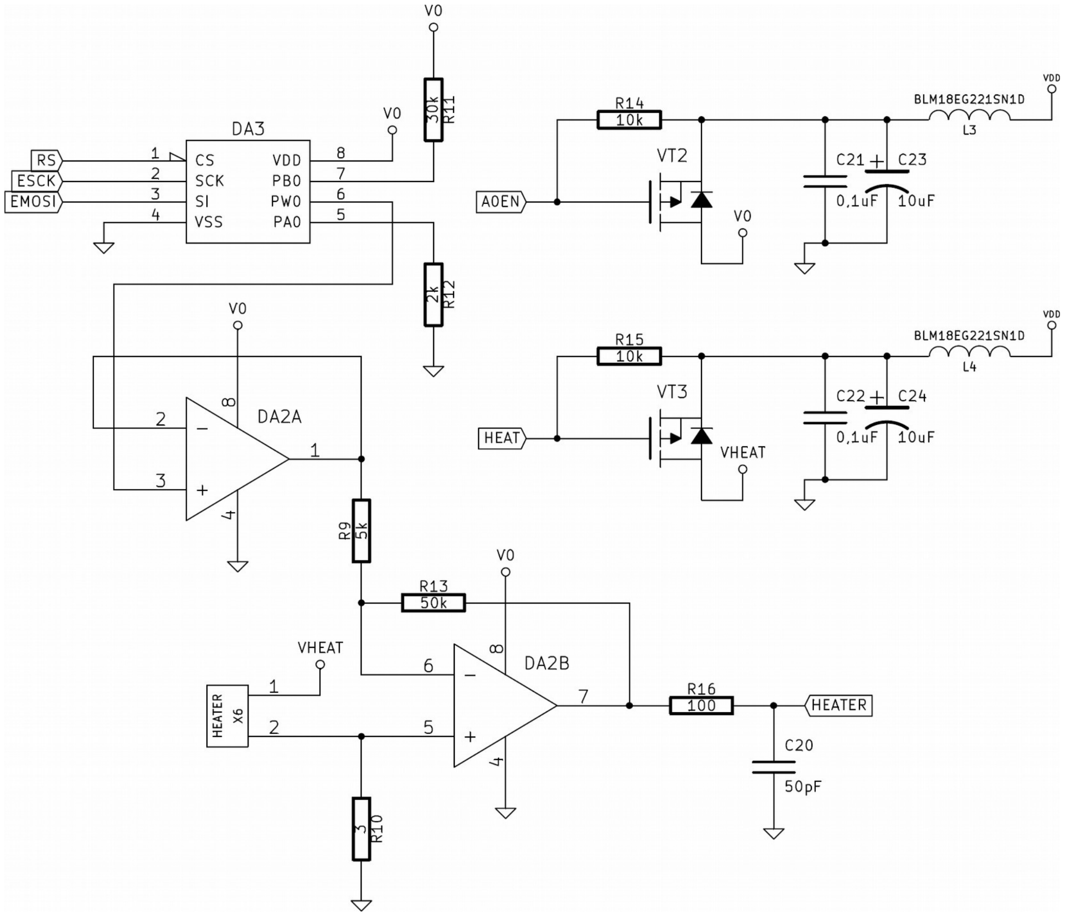


Fig 7. Sensing circuit for multistage pulse with PWM heating.

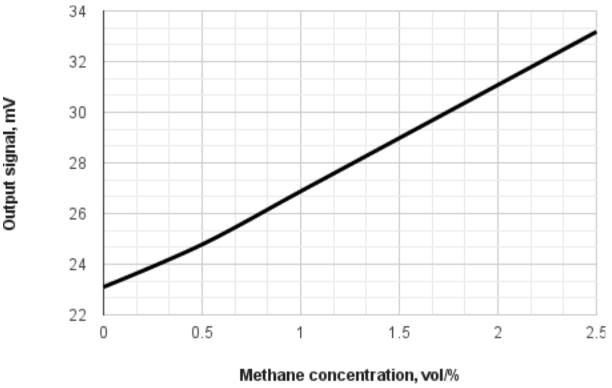


Fig 8. The dependence of the output signal of the sensor on the methane concentration.

As it follows from this equation, for the same PWM frequency and supply voltage changing pulse width it's possible to regulate average voltage on the load.

## III. Sensor Node Design

The full block diagram of the wireless gas sensor node is presented in Fig. 6. The sensor node is based on an AtXmega32A4 microcontroller and use an ETRX3 communication module. The selection of the MCU was mainly driven by the following requirements: low power consumption, on-chip temperature sensor, and good ADC integrated in MCU.

The wireless communication unit employs the low power ETRX3 wireless modem supporting IEEE 802.15.4 standard (ZigBee specification) and transmitting in unlicensed 2.4 GHz ISM band. The modem has an integrated chip antenna used in this design (up to 25 m) and a connector for an external antenna to enable a boost mode allowing data transmission for up to 350 m. Besides that, the modem has a number of self-x features enabling, for instance, WSN self-configuration and self-diagnostics which significantly reduce WSN debugging and deployment time.

Voltage conversion is performed by a DC-DC converter TPS63060 to provide maximum efficiency. The device generates stable output voltage of 3.3 V from 2.5 V to 12 V on its input.

Sensing circuit is presented in Fig. 7.

Since it is necessary to have long autonomous lifetime for wireless sensor node, all measurements are performed periodically. But at the same time, according to safety standards [11], [12], sensor node response time should be less than 20 seconds. Therefore, it is reasonable to perform two measurements every 20 seconds to assure this claim. To maintain low power consumption between measurements, the measurements circuit is turned off using power switch based on VT2 transistor.

As it was said before, sensor heating is performed using signal with pulse-width modulation. Here we use PWM frequency of 10 kHz. The average heating voltage is regulated by changing pulse duty rate with fixed power supply voltage. Therefore, with power supply voltage value of 3.3 V and voltage values for pulse stages of 3.3, 2.4, 0 and 1.6 V, duty rate values are 100, 73, 0 and 48 percent respectively. Voltage switching is performed using VT3 transistor. Since this switch also provides power supply for sensor itself, all measurements are performed with activated heating.

The sensor signal is the voltage from resistive divider which consists of sensor itself and reference resistor. The signal goes to the amplifier with gain value of 10.

Another input of the amplifier is connected to the output of voltage reference circuit. This circuit provides reference voltage for different stages of multistage pulse to exclude constant part of the signal from the output.

The circuit consists of resistive divider which is based on digital potentiometer. This potentiometer controls the reference voltage value. This voltage goes to the buffer amplifier. The output of this amplifier is the output of the circuit.

For all measurements built-in ADC of MCU with internal 1 V reference voltage source is used.

## IV. Sensor Response

As it was said before, the measurement result is the difference between sensor voltages at two different temperatures. The first value is the voltage at the end of the second stage of pulse heating (which heats the sensor up to 450 C and lasts for 190 ms). The second value is the voltage at the end of the fourth stage of pulse heating (which heats the sen-sor up to 200 C and lasts for 550 ms). This is done to compensate the absorption of moisture on the sensor surface that occurs between two cycles, and influences the quality of the measurement.

The dependence of the output signal of the sensor on the methane concentration is shown in Fig. 8. The value of the signal is changed from 23 mV for 0% of CH4 to 33 mV for 2.5% of CH4. Therefore, sensor sensitivity for this method is about 4 mV / %CH4.

The sensor node operates as a two-threshold device. Threshold values are 0.5 and 1 % vol. If the methane concentration is less than 0.5 % vol., there is no reaction from the sensor node. If the concentration is more than 0.5 % vol. methane, sensor node provides the light and sound alarm. This alarm is different for every threshold. At the same time, after every threshold is exceeds, node sends specific information to the data sink node.

## V. Power Consumption

The power consumption of the wireless sensor node is presented in Fig. 9.

As it was shown before, the advantages from PWM heating are in second and fourth stages of multistage pulse. For second stage the resulting power consumption is 150-160 mW. For fourth stage it is less than 80 mW. The average power consumption for overall multistage pulse is 81 mW. This value includes data transmission.

Measurements are performed periodically with period length of 10 seconds. The length of multistage pulse is about 0.6 seconds. The average power consumption, including data transmission, for overall period is about 4.9 mW.

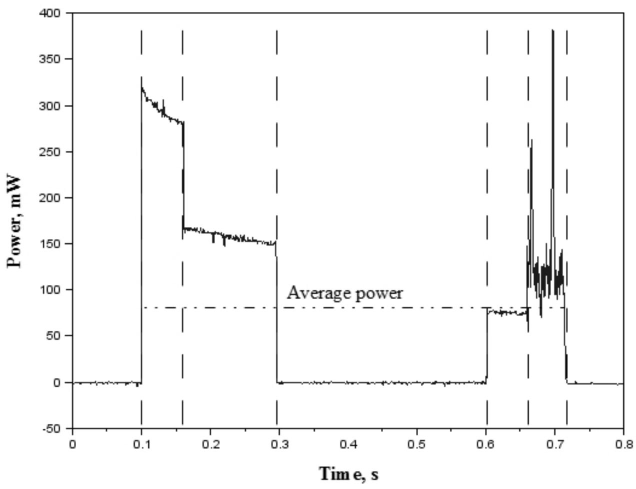


Fig 9. The power consumption of the wireless sensor node.

Since node power consumption is relatively low, the power consumption of data transmission is large enough part in overall power consumption.

The average power consumption of that data transmission cycle is about 125 mW. And its length is about 0.06 seconds. The average power consumption of multistage pulse including data transmission is about 81 mW. And its length is 0.6 seconds.

So the length of data transmission is about 10% of the pulse length. And during this time the power consumption is much higher than average values.

At the same time, usually it doesn't necessary to send data every time than measurements were made. It's much reasonable to transmit it only when gas concentration was changed or in other emergency situations.

The average power consumption excluding data transmission cycle is about 77 mW. The average power consumption for overall period in this case is 4.6 mW. These values are about 5% lower.

These power consumption values are relatively low. Since capacity of a single cell lithium battery of D type is typically 15000 mAh, its voltage is 3.6 V and there are 8760 hours a year, the average power consumption for one year sensor node lifetime is no more than 6 mW. Therefore, the average power consumption of the node is low enough for one year autonomous lifetime. This duration also complies with safety standards which claim to perform gas sensors calibration every year.

## VI. Conclusion

In this work the design of the "smart dust" sensor node for wireless sensor network for combustible gas concentration monitoring was presented.

The analog and digital circuits and energy efficient algorithms for sensor response measurements were designed. The parameters of designed sensor node was investigated.

The calculations of methane concentration were made using analysis of commercial catalytic methane sensor response in multistage pulse heating mode of operation which includes four heating stages. With that, sensor heating in every stage was performed using PWM voltage regulation unlike traditional heating with constant voltage.

It was shown that PWM regulation allows to save about 20% of power which sensor consume. Particularly the average power consumption for multistage pulse is 81 mW. And for 10 seconds measurement period the average power consumption value is about 4.9 mW.

The power consumption values provides the autonomous lifetime of the node more than one year. Therefore, the methods and algorithms which was designed are very promising for catalytic combustible gas sensors.

## Acknowledgment

This work was supported by the grant No. RFMEFI57714X0133 from the Ministry of Education and Science of Russian Federation.

## References

[1] M. Magno, D. Boyle, D. Brunelli, B. O'Flynn, E. Popovici, L. Benini, "Extended wireless monitoring through intelligent hybrid energy supply," IEEE Trans. On Ind. Electron., vol.61, no.4, pp.1871-1881, April 2014. http://dx.doi.org/10.1109/TIE.2013.2267694

[2] S. Akbari "Energy Harvesting for Wireless Sensor Networks Review" in Proc. Federated Conference on Computer Science and Information Systems (FedCSIS), 2014, pp. 987-992. http://dx.doi.Org/10.15439/2014F85

[3] Andrey Somov, Evgeny F. Karpov, Elena Karpova, Alexey Suchkov ,Sergey Mironov, Alexey Karelin, Alexander Baranov, Denis Spirjakin, Compact Low Power Wireless Gas Sensor Node with Thermo Compensation for Ubiquitous Deployment, Industrial Informatics, IEEE Transactions on, 2015, Issue: 99. http://dx.doi.org/10.1109/TII.2015.2423155

[4] A. Somov, A. Baranov, D. Spirjakin, R. Passerone. Circuit design and power consumption analysis of wireless gas sensor nodes: one-sensor versus two-sensor approach. IEEE Sensors Journal 14(6): 2056-2063, 2014. http://dx.doi.org/10.1109/JSEN.2014.2309001

[5] http://www.willow.co.uk/TelosB_Datasheet.pdf

[6] http://www.eol.ucar.edu/isf/facilities/isa/internal/CrossBow/DataSheets/mica2.pdf

[7] http://www.openautomation.net/uploadsproductos/micaz_datasheet..pdf

[8] http://www.libelium.com/products/waspmote/

[9] http://www.openpicus.com

[10] A. Somov, D. Spirjakin, M. Ivanov, I. Khromushin, R. Passerone, A. Baranov, and A. Savkin. Combustible gases and early fire detection: an autonomous system for wireless sensor networks. In Proceedings of the First ACM International Conference on Energy-Efficient Computing and Networking (e-Energy'10), pp. 85-93, Passau, Germany, April 13-15, 2010. http://dx.doi.org/10.1145/1791314. 1791327

[11] Standard GOST R EN 50194-1-2012, Signalizators for the detection of combustible gases in domestic premises, 2000.

[12] Standard EN 50194-2000, Electrical apparatus for the detection of combustible gases in domestic premises. Test methods and performance requirements, 2000.

[13] Denis Spirjakin, Alexander Baranov, Alexey Karelin, Andrey Somov, Wireless Multi-Sensor Gas Platform for Environmental Monitoring, Environmental, Energy and Structural Monitoring Systems (EESMS), 2015 IEEE Workshop on 9-10 July 2015, Italy, Trento, pp. 232 - 237. http://dx.doi.org/10.1109/EESMS.2015.7175883

[14] Alexander Baranov, Denis Spirjakin, Saba Akbari, Andrey Somov, Optimization of power consumption for gas sensor nodes: A survey. Sensors and Actuators A 233 (2015) 279–289. http://dx.doi.org/10.1016/j.sna.2015.07.016

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1285–1290

DOI: 10.15439/2015F266
ACSIS, Vol. 5

# Comparisons between 2D and 3D Uniform Array Antennas

Andy Vesa
POLITEHNICA University
Timişoara, Romania,
Communications Department, V
Parvan No 2, 300223, Timişoara
Email: andy.vesa@upt.ro

Florin Alexa
POLITEHNICA University
Timişoara, Romania,
Communications Department, V
Parvan No 2, 300223, Timişoara
Email: florin.alexa@upt.ro

Horia Baltă
POLITEHNICA University
Timişoara, Romania,
Communications Department, V
Parvan No 2, 300223, Timişoara
Email: horia.balta@upt.ro

□

***Abstract*—For any wireless communications antenna system becomes indispensable. In this paper we analyzed linear array, planar array and three – dimensional (3D) array antennas. The array systems are simulated in Matlab based on uniform linear array antennas. Comparisons between planar array antenna and 3D array antenna are provided take into account different phases of currents injected in antenna elements. Also we propose to use the array antenna in WSN due to the advantages in signal to noise ratio and power consumption.**

## I. Introduction

A radio antenna, transmitting or receiving, is an independent and yet integral component of any wireless communication system. An antenna acts as a transducer that converts the current or voltage generated by the feeding-based circuit, such as a transmission line, a waveguide or coaxial cable, into energy field propagating through space and vice versa. Each radio signal can be represented as an electromagnetic wave that propagates along a given direction. The wave field strength, its polarization and the direction of propagation determine the main characteristics of an antenna operation [1].

Antennas can be divided in different categories, such as wire antennas, aperture antennas, reflector antennas, frequency independent antennas, horn antennas, printed and conformal antennas, and so forth. When applications require radiation characteristics that cannot be met by a single radiating antenna, multiple elements are employed forming “array antennas”. Arrays can produce the desired radiation characteristics by appropriately exciting each individual element with certain amplitudes and phases.

Over the last decade, wireless technology has grown at a formidable rate, thereby creating new and improved services at lower costs. This has resulted in an increase in airtime usage and in the number of subscribers. Recently, smart antenna systems have been widely considered to provide interference reduction and improve the performance of wireless mobile communication. The term “smart antenna” reflects the antenna’s ability to adapt to the communication channel environment in which it operates. Smart antenna arrays with adaptive beamforming capability are very effective in the suppression of interference and multipath signals. One practical solution to this problem is to use spatial processing. Spatial processing is the central idea of adaptive antennas or smart-antenna systems [2].

In this paper we are focused on changing of the phase of currents injected in antenna elements in order to conclude relating to position control of the radiation pattern characteristic and directivity.

## II. ANTENNA ARRAY RADIATION PATTERN

An antenna *radiation pattern* or *antenna pattern* is defined as “mathematical function or a graphical representation of the radiation properties of the antenna as a function of space coordinates. In most cases, the radiation pattern is determined in the far-field region and is represented as a function of the directional coordinates.” [3] The radiation property of most concern is the two- or three-dimensional spatial distribution of radiated energy as a function of the observer’s position along a path or surface of constant radius.

Antennas with a given radiation pattern may be arranged in a structure (line, plane, three – dimensional) to yield a different radiation pattern. Given an antenna array of identical elements, the radiation pattern of the antenna array may be found according to the pattern multiplication theorem [4]:

*Array pattern = Array element pattern x Array factor*

where *Array element pattern* is the pattern of the individual array element and *Array factor* is a function dependent only on the geometry of the array and the excitation (amplitude, phase) of the elements.

### A. Linear Array

The array factor (AF) is independent of the antenna type assuming all of the elements are identical. Thus, isotropic radiators may be utilized in the derivation of the array factor to simplify the algebra. The field of an isotropic radiator

□ Romanian Ministry of Education, CNCS – UEFISCDI

978-83-60810-66-8/$25.00©2015, IEEE

located at the origin may be written as (assuming $\theta$-polarization):

$$E_\theta = I_0 \frac{e^{-jkr}}{4\pi r} \tag{1}$$

where:

$I_0$ = the complex excitation of the isotropic radiator,

$k$ = the free space wave number,

$r$ = distance of the observation point from the origin.

In our approach, we assume that the elements of the array are uniformly – spaced with a separation distance $d$ (see Fig. 1).

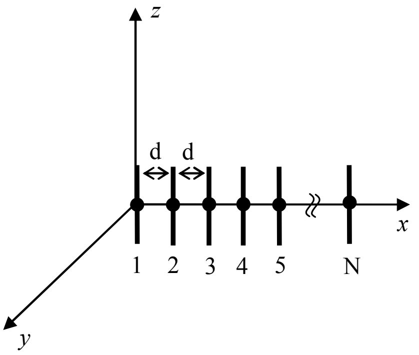


Fig. 1 The linear array antennas

The current magnitudes of the array elements are supposed to be equal and the current on the array element located at the origin is used as the phase reference (zero phase).

$$I_1 = I_0 \quad I_2 = I_0 e^{j\phi_2} \quad \cdots \quad I_N = I_0 e^{j\phi_N} \tag{2}$$

The far electromagnetic fields of the individual array elements are:

$$\begin{aligned} E_{\theta_1} &\approx I_0 \frac{e^{-jkr}}{4\pi r} = E_0 \\ E_{\theta_2} &\approx I_0 e^{j\phi_2} \frac{e^{-jk(r-d\cos\theta)}}{4\pi r} = E_0 e^{j(\phi_2 + kd\cos\theta)} \\ &\vdots \\ E_{\theta_N} &\approx I_0 e^{j\phi_N} \frac{e^{-jk[r-(N-1)d\cos\theta]}}{4\pi r} = E_0 e^{j\left[\phi_N + (N-1)kd\cos\theta\right]} \end{aligned} \tag{3}$$

The overall array far field is found using superposition and could be express as:

$$E_\theta = E_{\theta_1} + E_{\theta_2} + E_{\theta_3} + \ldots + E_{\theta_N} = E_0 * AF \tag{4}$$

The array factor for a uniformly – spaced $N$ – element linear array from (4) is:

$$AF = \left[1 + e^{j(\phi_2 + kd\cos\theta)} + \ldots + e^{j\left[\phi_N + (N-1)kd\cos\theta\right]}\right] \tag{5}$$

A uniform array is defined by uniformly – spaced identical elements of equal magnitude with a linearly progressive phase from element to element:

$$\phi_1 = 0 \quad \phi_2 = \alpha \quad \phi_3 = 2\alpha \cdots \phi_N = (N-1)\alpha \tag{6}$$

Inserting this linear phase progression into the formula for the general $N$ – element array, we obtain:

$$AF = \left[1 + e^{j(\alpha + kd\cos\theta)} + \ldots + e^{j(N-1)[\alpha + kd\cos\theta]}\right] = \sum_{n=1}^{N} e^{j(n-1)\psi} \tag{7}$$

The function $\psi$ ($\psi = \alpha + kd\cos\theta$) is defined as the array phase function and is a function of the element spacing, phase shift, frequency and elevation angle. If the array factor from (5) is multiplied by $e^{j\psi}$, the result is:

$$AF * e^{j\psi} = \left[e^{j\psi} + e^{2j\psi} + e^{3j\psi} + \ldots + e^{jN\psi}\right] \tag{8}$$

Substracting the array factor from the equation above, we obtain:

$$AF = e^{j(N-1)\frac{\psi}{2}} \frac{\sin\left(\frac{N\psi}{2}\right)}{\sin\left(\frac{\psi}{2}\right)} \tag{9}$$

The complex exponential term in (9) represents the phase shift of the array phase center relative to the origin. If the position of the array is shifted so that the center of the array is located at the origin, this phase term goes away. The array factor after phase shift and normalization becomes:

$$AF = \frac{1}{N} \frac{\sin\left(\frac{N\psi}{2}\right)}{\sin\left(\frac{\psi}{2}\right)} \tag{10}$$

### *B. Planar Array*

Unlike linear arrays that can only scan the main beam in one polar plane ($\theta$ - the elevation plane or $\phi$ - the azimuth plane), planar arrays scan the main beam along both $\theta$ and $\phi$. Planar arrays offer more gain and lower sidelobes than linear arrays, at the expense of using more elements. The design principles for planar arrays are similar to those presented earlier for the linear arrays. Since the elements are placed in two dimensions (see Fig. 2), the array factor of a planar array can be expressed as the multiplication of the array factors of two linear arrays [5]: one along the x-axis and one along the y-axis:

$$AF_{planar} = AF_x * AF_y = \frac{\sin\left(\frac{N_1\psi_x}{2}\right)}{N_1 \sin\left(\frac{\psi_x}{2}\right)} \frac{\sin\left(\frac{N_2\psi_y}{2}\right)}{N_2 \sin\left(\frac{\psi_y}{2}\right)} \tag{11}$$

where:

$$\begin{aligned} \psi_x &= kd_x \,\mathbf{sin}\,\theta\,\mathbf{cos}\,\phi + \alpha_x \\ \psi_y &= kd_y \,\mathbf{sin}\,\theta\,\mathbf{sin}\,\phi + \alpha_y \end{aligned} \tag{12}$$

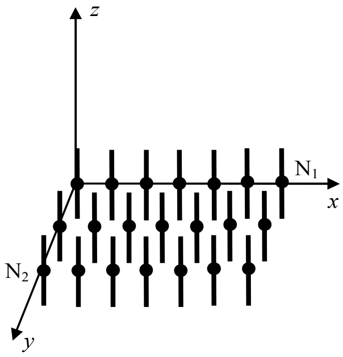


Fig. 2 The planar array antennas

### C. Three – dimensional Array (3 – D Array)

Starting from the planar array, where it is considered that the system has a rectangular configuration of elements, the three – dimensional array antenna (Fig. 3) is achieved by introducing a number of planar arrays on the *z* axis. In this case, the array factor is:

$$AF_{3D} = AF_x * AF_y * AF_z =$$
$$= \frac{\sin\left(\frac{N_1\psi_x}{2}\right)}{N_1\sin\left(\frac{\psi_x}{2}\right)} \frac{\sin\left(\frac{N_2\psi_y}{2}\right)}{N_2\sin\left(\frac{\psi_y}{2}\right)} \frac{\sin\left(\frac{N_3\psi_z}{2}\right)}{N_3\sin\left(\frac{\psi_z}{2}\right)} \quad (13)$$

where:

$$\begin{aligned}\psi_x &= kd_x \sin\theta\cos\phi + \alpha_x \\ \psi_y &= kd_y \sin\theta\sin\phi + \alpha_y \\ \psi_z &= kd_z \cos\theta + \alpha_z\end{aligned} \quad (14)$$

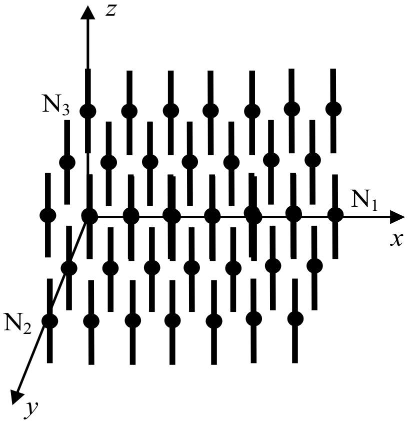


Fig. 3 Three – dimensional array antennas

In case of three – dimensional array antennas there are restrictions regarding the distance between elements on *z* axis which could not be less than elements dimensions.

## III. Simulation Results

Our simulations are made for linear, planar and three – dimensional array antennas. Starting from a linear array antennas with logarithmic directivity pattern, composed between 2 – 9 elements (dipole antenna) that are evenly spaced with the distance of *λ/2*, we built a planar array and, respectively three – dimensional array antennas.

The simulations are made in Matlab and all results are displayed normalized. The radiation patterns obtained in different situations are generated in two planes, *xOy* plane (*H* plane) and *xOz* plane (*E* plane). We use the opening angle of the main lobe to compare the results obtained in these situations [7].

Because for a linear array antenna the increases in a phase shift between currents injected in two consecutive elements leads to decreases in directivity, we are focused on planar and *3D* array antennas.

The radiation patterns obtained in *xOy* plane (*H* plane) for planar array antennas with different number of elements and phases are represented in Fig. 4. In Fig. 4.c a radiation pattern for 45° phase shift of currents injected in array elements is presented.

The radiation pattern obtained in *xOz* plane (*E* plane) is presented in Fig. 5.

For planar array antenna with 7 × 7 elements, and three – dimensional array antenna with 7 × 7 × 7 elements, and 0° phase shift, the radiation patterns are presented in Fig. 6.

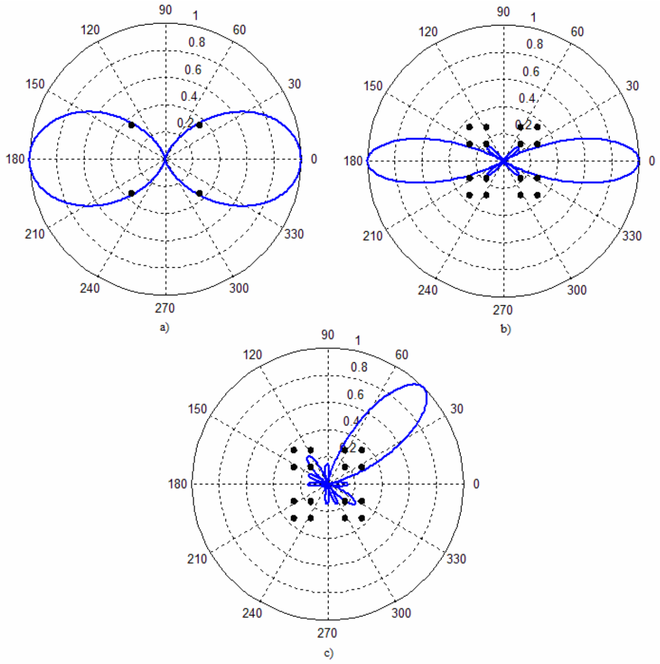


Fig. 4 Radiation patterns obtained in *H* plane for planar array antennas: a) with 2x2 elements and 0° phase shift; b) with 4x4 elements and 0° phase shift; c) with 4x4 elements and 45° phase shift

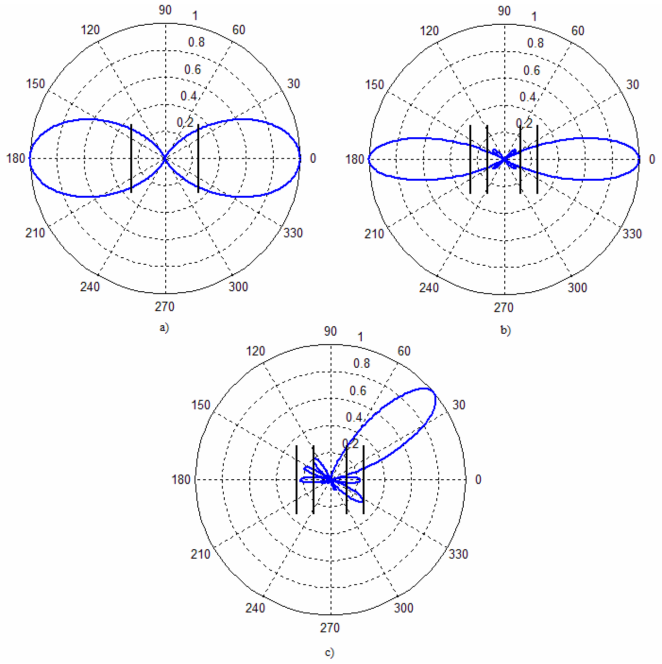


Fig. 5 Radiation patterns obtained in *E* plane for planar array antennas: a) with 2x2 elements and 0° phase shift; b) with 4x4 elements and 0° phase shift; c) with 4x4 elements and 45° phase shift

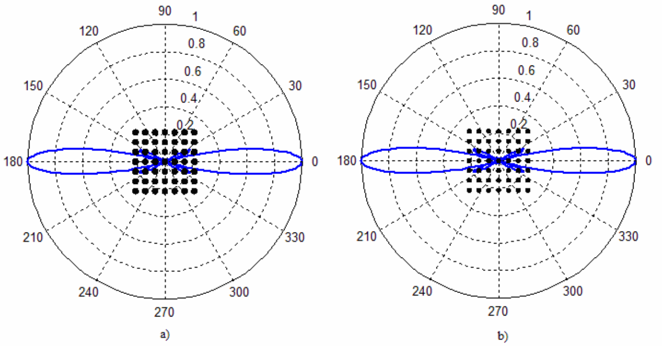


Fig. 6 Radiation patterns obtained in *H* plane for: a) planar array antennas with 7x7 elements and; b) 3D array antennas with 7x7x7 elements

As a comparison we used the opening angle of the main lobe. In case of identical currents injected in array elements the results are presented in Table I.

During the experiments we conclude that increasing the number of elements over 7 does not lead to significant increase of directivity, so we suggest that the maximum number of elements for an array should not be greater than 7. Using planar array antennas with 7 × 7 elements and introducing a phase shift of 30° and 165° between the currents injected in two consecutive elements, we obtain the radiation patterns represented in Fig. 7. For three – dimensional array antennas with 7 × 7 × 7 elements and with a phase shift equal to 105° and 165° respectively, the radiation patterns are represented in Fig. 8.

The 3D patterns for planar array and three – dimensional array are presented in Fig. 9. Because we used a distance between elements equal with *λ/2* for phase shift equal with 180° the obtained pattern is same as in case of no phase shift.

The values obtained for the opening angle of main lobe in case of array antennas formed by 7 elements with different phase shift between consecutive elements are presented in Table II.

We could observe that the opening angle for different structure array is same if the phase shift is complement to 180°, but for angles lower than 30° or higher than complement to 180° or for angles between 70° and 110° the pattern characteristic has two lobes, one in direction determined by phase shift and another to the direction complement to 180° attenuated with 20%.

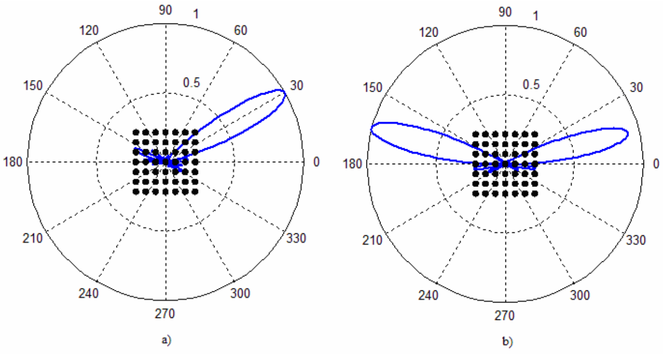


Fig. 7 Radiation pattern characteristic obtained in *H* plane for planar array antennas with: a) 30° phase shift, b) 165° phase shift

TABLE I.
OPENING ANGLE FOR ARRAY ANTENNAS FOR NUMBER OF ELEMENTS

| **Number of elements** | **Linear** | | **Planar** | | **Three – dimensional** | |
|---|---|---|---|---|---|---|
| | ***H*** | ***E*** | ***H*** | ***E*** | ***H*** | ***E*** |
| 2 | 59,6° | 97,2° | 58,4° | 47° | 58,4° | 27° |
| 3 | 35,5° | 57,3° | 35,5° | 32,1° | 35,5° | 32,1° |
| 4 | 25,3° | 23,5° | 26,3° | 24° | 26,3° | 12,6° |
| 5 | 19,6° | 47° | 20,6° | 19,5° | 20,6° | 19,5° |
| 6 | 16,1° | 18,3° | 17,1° | 16° | 17,1° | 9,1° |
| 7 | 13,7° | 40,1° | 13,7° | 13,7° | 13,7° | 13,7° |
| 8 | 11,4° | 14,9° | 11,4° | 11,4° | 11,4° | 6,8° |
| 9 | 10,3° | 39.2° | 10,3° | 10,3° | 10,3° | 10,3° |

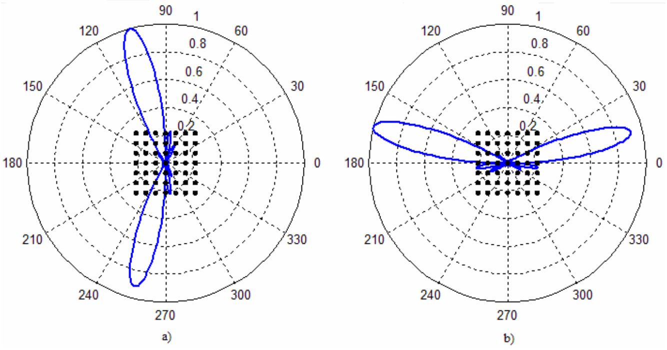


Fig. 8 Radiation pattern characteristic obtained in *H* plane for three – dimensional array antennas with: a) 105° phase shift, b) 165° phase shift

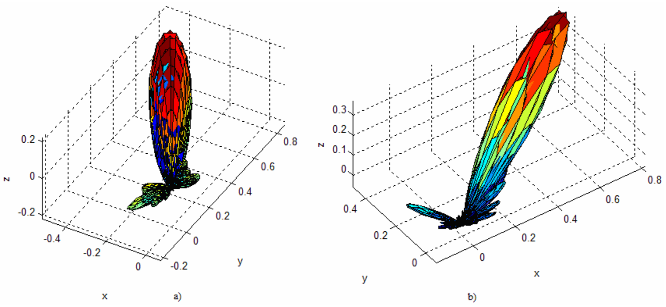


Fig. 9 3D radiation pattern characteristic for: a) planar array antennas with 120° phase shift, b) three – dimensional array antennas with 30° phase shift

## IV. WSN APPLICATIONS

One of the most interesting applications for directional antenna is Wireless Sensor Networks (WSN). WSN represent a spatially distributed of large number of autonomous sensors which collect data from environment and is able to send it to a main certain locations. WSN has many applications from different field. The main areas of applications of WSN are: military (monitoring forces, equipment and ammunitions; battlefield surveillance; targeting, etc.), environmental (forest fire detection; bio-complexity mapping of environment; flood detection; air and water pollution), health (telemonitoring of human physiological data, tracking and monitoring doctors and patients, drug administration), home and office (automation; smart environment), automotive (reduces wiring effects; measurements in chambers and rotating parts, remote technical inspections, conditions monitoring e.g. at a bearing) [8].

The architecture of the WSNs can take many forms starting from a star network to an advanced multi-hop wireless network. One possible solution for WSN communications architecture is presented in Fig. 10.

The components of the networks are called nodes; each of them is connected to one or, sometimes, several others. The main structure for sensor nodes is presented in Fig. 11.

In many cases in order to connect to send data it is necessary to obtain a maximum signal to noise ratio in transmission [9].

Because the transmission system has a consumption that depends on the signal to noise ratio, it is necessary to improve this parameter.

In a transmission system the signal to noise ratio could be improved without increasing energy if we are able to increase the gain of the antenna or if we are able to orient the antennas in a way in what the pattern characteristics are aligned. Using an array antenna we can control the orientation of the pattern characteristic without any energy consumption. This leads to the increased signal to noise ratio and to the reduced energy consumption because it is not necessary a mechanical orientation of the elements of the node in order to obtain maximum signal to noise ratio.

TABLE II.
OPENING ANGLE FOR ARRAY ANTENNAS FOR PHASE SHIFT

| Phase shift | Linear | | Planar | | Three – dimensional | |
|---|---|---|---|---|---|---|
| | ***H*** | ***E*** | ***H*** | ***E*** | ***H*** | ***E*** |
| $0^0$ | 13,7° | 40,1° | 13,7° | 13,7° | 13,7° | 13,7° |
| $15^0$ | 14,9° | 32,1° | 13,7° | 13,7° | 13,7° | 12,6° |
| $30^0$ | 13,7° | 11,4° | 13,7° | 13,7° | 13,7° | 10,3° |
| $45^0$ | 14,9° | 12,6° | 14,9° | 13,7° | 14,9° | 10,3° |
| $60^0$ | 14,9° | 13,7° | 14,9° | 13,7° | 14,9° | 10,3° |
| $75^0$ | 16° | 13,7° | 14,9° | 12,6° | 14,9° | 10,3° |
| $90^0$ | 17,2° | 13,7° | 13,7° | 10,5° | 13,7° | 12,6° |
| $105^0$ | 16° | 13,7° | 14,9° | 12,6° | 14,9° | 10,3° |
| $120^0$ | 14,9° | 13,7° | 14,9° | 13,7° | 14,9° | 10,3° |
| $135^0$ | 14,9° | 12,6° | 14,9° | 13,7° | 14,9° | 10,3° |
| $150^0$ | 13,7° | 11,4° | 13,7° | 13,7° | 13,7° | 10,3° |
| $165^0$ | 14,9° | 32,1° | 13,7° | 13,7° | 13,7° | 12,6° |
| $180^0$ | 13,7° | 40,1° | 13,7° | 13,7° | 13,7° | 13,7° |

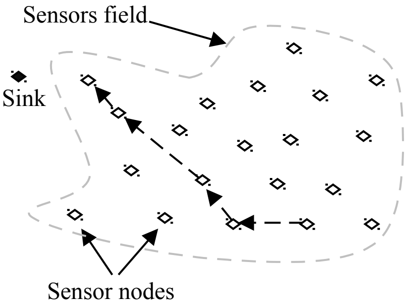


Fig. 10 WSN architecture

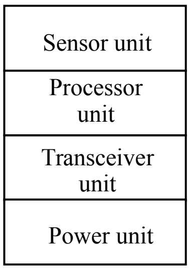


Fig. 11 Node architecture

## V. Conclusion

A brief overview of antenna array aspects has been given. We observe that in cases of planar array and three – dimensional array, the systems become more directly. Also, if a phase shift between current injected in the antenna elements is introduced the linear array become impracticable in terms of directivity pattern.

During the experiments we conclude that increasing the number of elements over 7 does not lead to significant increase of directivity, so we suggest that the maximum number of elements for an array should not be greater than 7.

In order to obtain an oriented pattern characteristic, only planar and 3D array could be used because a linear array has many secondary lobes, and in this case the gain decreases. Also in order to obtain a good directivity the phase shift should be between 30 and 150 degree.

Array antennas are increasingly used nowadays because using them can improve radio transmissions by:

- increasing signal coverage, by increasing capacity for data transfer,
- reducing interference due to multipath signal propagation, simultaneous multiple signals radiated to different users,
- better focus of the beam radiated to the direction of interest.

For a better orientation of the beam radiated to a user, the array antenna must be able to detect the direction from which the signal arrives from the user. Future work is oriented to investigate what is happened in case of changes the regularity of the array structure and also a new structure for planar array.

## Acknowledgment

The work of Horia Baltă was supported by a grant of the Romanian Ministry of Education, CNCS – UEFISCDI, project number PN-II-RU-PD-2012-3-0122. Horia Baltă is also with Valahia University of Târgovişte, Romania.

## References

[1] C.A. Balanis, *Antenna Theory: Analysis and Design.* 2nd ed., John Wiley & Sons, New York, 1997, pp. 27–48.

[2] W.L. Stutzmann, G.A. Thiele, *Antenna Theory and Design.* John Wiley & Sons, New York, 1998, pp. 31–35.

[3] R.J. Mailloux, *Phased Array Antenna Handbook.* 2nd ed., Artech House, New York, 2005, pp. 48–100.

[4] F.T. Ulaby, *Applied Electromagnetics.* Prentice Hall, Inc., New Jersey, 2007, pp. 100–125.

[5] C.A. Balanis, *Modern Antenna Handbook*, John Wiley & Sons, New York, 2008, pp. 60–95.

[6] K. K. Verma, K. R. Soni, “Theoretical study of 2x2 element planar array of equilateral triangular patch microstrip antenna in plasma medium,” *PRAMANA – Journal of Physics*, vol. 64, pp. 147–152, Jan. 2005.

[7] A. Ignea, E. Marza, A. De Sabata, *Antene şi Propagare (in romanian language).* Editura de Vest, Timişoara, 2002, pp. 289–310.

[8] W. Dargie, C. Poellabauer, *Fundamentals of wireless sensor networks: theory and practice*, John Wiley & Sons Ltd, New York, 2010, pp. 150-245

[9] P. Dorfinger, G. Panholzer, F. von Tüllenburg, M. Cristaldi, G. Tusa, and F. Böhm, Self-aligning wireless communication for first responder organizations in interoperable emergency scenarios, ICN 2015, Proceedings of The Fourteenth International Conference on Networks, April 19 - 24, 2015, Barcelona, Spain, pp.230-236,

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1291–1295

DOI: 10.15439/2015F298
ACSIS, Vol. 5

# Analysis of Inductively Coupled RFID Marker Localization Methods

Peter Vestenický
University of Žilina, Faculty of Electrical Engineering, Department of Control and Information Systems, Univerzitná 8215/1,
010 26 Žilina, Slovakia
Email:
peter.vestenicky@fel.uniza.sk

Tomáš Mravec
University of Žilina, Faculty of Electrical Engineering, Department of Control and Information Systems, Univerzitná 8215/1,
010 26 Žilina, Slovakia
Email:
tomas.mravec@fel.uniza.sk

Martin Vestenický
University of Žilina, Faculty of Electrical Engineering, Department of Telecommunications and Multimedia, Univerzitná 8215/1,
010 26 Žilina, Slovakia
Email:
martin.vestenicky@fel.uniza.sk

*Abstract* — **The presented paper is focused on analysis of two methods of marker localization. The markers are passive RFID transponders (without or with identification chip) consisting of tuned LC circuit and being used to mark and trace underground networks such as cables and pipes. Localization of the marker is based on evaluation of signal amplitude received from the excited marker, i. e. it is RSSI based localization method. The excitation of marker can be periodically repeated or continuous. In the first case the localization process consists of two stages – excitation and receiving of marker damped oscillations, in the second case the amplitude of continuously generated excitation signal is decreased by vicinity of the marker. Both localization methods are mathematically analyzed by modeling of their circuits using differential equations. The results of analysis are used to compare both methods and to evaluate their suitability for practical utilization.**

## I. Introduction

INDUCTIVELY coupled RFID (Radio Frequency Identification) systems [1] are now being widely used in many industrial applications. For example, the marking of goods by RFID technology enables the traceability of goods which is helpful to control the whole logistic chain from production to sale. In addition to these applications, the RFID transponders are being used for marking of underground facilities location. Such RFID transponders are called “markers”. The marker is a passive RFID transponder consisting of a tuned LC circuit without identification chip (1-bit) or with identification chip tuned on low frequency in 77 kHz – 170 kHz band.

For localization of some older underground facilities (cables, pipes etc.) a signal can be injected into their continual metal conductor and this signal can then be received on terrain surface and the cable or pipe can be traced. Today’s underground facilities are mostly constructed from plastic material so this simple localization and tracing method cannot be used, therefore in this case the marking of underground objects by RFID markers is the only useable method.

The unknown position of marker under the terrain surface can be estimated by a localization device (locator). Moreover, the depth of marker can be estimated by RSSI (Received Signal Strength Indication) similarly as described in [2], [3]. In [4] the authors describe marker localization methods based on marker damped oscillations and on continuous generating of electromagnetic waves. This work extends the analyses of RSSI based marker localization methods. The method based on damped oscillations of marker is extended by introduction of separate damping and sensing resistors and the analysis of method based on continuous marker excitation is done by applying differential equations instead of algebraic equations to analyze the transient phenomena in sensing circuit.

## II. Related Works

The localization of moving underground objects (for example animals) based on inductive coupling is described in [5]. This application assumes the use of sensor network consisting of transmitting coils fixed on terrain surface and the moving underground object equipped with receiver collects data transmitted from these coils.

Indoor localization based on triaxial coils applied in both transmitter and receiver with very low working frequency 2.5 kHz is published in [6]. Low frequency magnetic field is suitable for underground localization purposes, too, because it is not affected by ground properties.

Another approach is presented in [7]. This work assumes localization based on UHF RFID tags in mining industry, but it is performed in mining tunnels and localization from the terrain surface is not assumed so the UHF signals do not propagate through layer of ground.

## III. Mutual Inductance

An important quantity in the models presented in next chapters is the mutual inductance $M$ between marker and locator coils which depends on the geometrical arrangement of coils according to equation [1], [4], [8]:

$$M = \frac{\sqrt{L_R L_T}\, r_R^2 r_T^2 \cos\theta}{\sqrt{r_R r_T}\left(r_R^2 + x^2\right)^{\frac{3}{2}}} \tag{1}$$

978-83-60810-66-8/$25.00©2015, IEEE

where $r_R$ is radius of RFID locator antenna coil, $r_T$ is radius of marker coil, $x$ is distance between the locator antenna and the marker and $\theta$ is the angle between coil of locator antenna and marker coil (note that if $\theta$=0 ° then coils are parallel).

## IV. Mathematical Model Based On Damped Oscillation

This principle of localization assumes that the localization device periodically excites the marker LC circuit and in the pauses between excitation periods the response from marker in form of its damped oscillations is received. The simple model with one resistor was analyzed in [4]. In this chapter a more complex model with separate damping and sensing resistors will be analyzed. The first resistor $R_D$ is used for fast damping of locator $L_RC_R$ circuit damped oscillations and its value was calculated from equation (2). The second resistor $R_M$ is used for current sensing in the receiving stage of localization. The model is shown in Fig. 1.

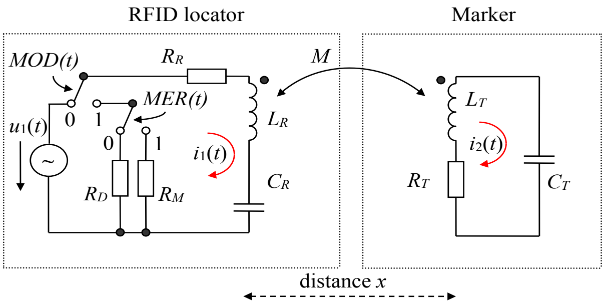


Fig. 1 Model of localization with separate damping and measuring resistors

$$R_D = 2\sqrt{\frac{L_R}{C_R}} - R_R \tag{2}$$

For this model the next system of equations (3) can be derived

$$L_R\frac{di_1(t)}{dt} + [(1-MER(t))MOD(t)R_D + MER(t)MOD(t)R_D + R_R]i_1(t) + \\ + \frac{1}{C_R}\int_0^t i_1(\tau)d\tau - M\frac{di_2(t)}{dt} = (1-MOD(t))u_1(t)$$

$$L_T\frac{di_2(t)}{dt} + R_Ti_2(t) + \frac{1}{C_T}\int_0^t i_2(\tau)d\tau - M\frac{di_1(t)}{dt} = 0 \tag{3}$$

The modulation function $MOD(t)$ is given by equation (4) and the sensing (measuring) resistor $R_M$ is switched by function $MER(t)$ given by the equation (5)

$$MOD(t) = \frac{\mathrm{Sign}\left(-\sin\frac{2\pi ft}{250}\right)+1}{2} \tag{4}$$

$$MER(t) = \frac{\mathrm{Sign}\left(-\sin\frac{2\pi f(t-\Delta T)}{250}\right)+1}{2} \tag{5}$$

i. e. it is binary square signal with a frequency 250 times lower than the frequency of excitation signal source and shifted (delayed) by $\Delta T$ in time against the modulation function $MOD(t)$. During the small delay $\Delta T$ the oscillations of locator tuned circuit has to be damped.

The excitation signal source is assumed harmonic, i. e.

$$u_1(t) = U_1\sin(2\pi ft) \tag{6}$$

The system of integrodifferential equations (3) was numerically solved after its conversion to the 1st order system of differential equations (7) by substitution $x_1(t)=i_1(t)$, $x_2(t)=\mathrm{d}i_1(t)/\mathrm{d}t$, $x_3(t)=i_2(t)$, $x_4(t)=\mathrm{d}i_2(t)/\mathrm{d}t$, $\omega=2\pi f$. Then we get:

$$\begin{aligned} \frac{dx_1(t)}{dt} &= x_2(t) \\ \frac{dx_2(t)}{dt} &= -a_1x_1(t) + a_2(1-MOD(t))\omega U_1\cos(\omega t) - a_3x_3(t) - a_4x_4(t) - \\ &\quad - a_2\left[R_R + [1-MER(t)]MOD(t)R_D + MER(t)MOD(t)R_M\right]x_2(t) \\ \frac{dx_3(t)}{dt} &= x_4(t) \\ \frac{dx_4(t)}{dt} &= -b_1x_1(t) + b_2(1-MOD(t))\omega U_1\cos(\omega t) - b_3x_3(t) - b_4x_4(t) - \\ &\quad - b_2\left[R_R + [1-MER(t)]MOD(t)R_D + MER(t)MOD(t)R_M\right]x_2(t) \end{aligned} \tag{7}$$

where the individual coefficients $a_1$, $a_2$, $a_3$, $a_4$ and $b_1$, $b_2$, $b_3$, $b_4$ are given by (8).

$$\begin{aligned} a_1 &= \frac{L_T}{C_R(L_RL_T-M^2)} & b_1 &= \frac{M}{C_R(L_RL_T-M^2)} \\ a_2 &= \frac{L_T}{L_RL_T-M^2} & b_2 &= \frac{M}{L_RL_T-M^2} \\ a_3 &= \frac{M}{C_T\left(L_RL_T-M^2\right)} & b_3 &= \frac{L_R}{C_T\left(L_RL_T-M^2\right)} \\ a_4 &= \frac{MR_T}{L_RL_T-M^2} & b_4 &= \frac{L_RR_T}{L_RL_T-M^2} \end{aligned} \tag{8}$$

The used values of $R_R$, $L_R$, $C_R$ and $R_T$, $L_T$, $C_T$ are listed in the next table:

Table I.
Values of Components Used in Numerical Calculations

| $R_R$ | $L_R$ | $C_R$ | $R_T$ | $L_T$ | $C_T$ |
|---|---|---|---|---|---|
| 15.7 Ω | 1 mH | 1.621 nF | 7.85 Ω | 1 mH | 1.621 nF |

Note that the values of $L_R$, $C_R$ and $L_T$, $C_T$ were selected so that the corresponding resonant frequencies are $f_R$=$f_T$=125 kHz.

The radiuses of both coils used in numerical calculations are $r_R$=$r_T$=0.1 m, distance between them is $x$=0.4 m and the angle $\theta$=0 °. Corresponding mutual inductance M is then calculated from (1). Amplitude of the excitation signal is $U_1$=10 V, its frequency $f$ is 125 kHz.

The damping resistor $R_D$=1533 Ω calculated from (2) in this case ensures the minimum time of $L_RC_R$ transient response. The used values of the measuring resistor are $R_M$=100 Ω for comparison with the results obtained in [4] and $R_M$=1 Ω to maximize the current response $i_1(t)$ from marker.

The time course of the current $i_1(t)$ for model of marker localization with two resistors is shown in Fig. 2 and Fig. 3 for $R_D$=1533 Ω, $R_M$=100 Ω and $R_M$=1 Ω, respectively.

Moreover, the dependence of maximum current amplitude $I_{1\_Max}$ on the distance $x$ was calculated for this model and it is shown in Fig. 4. Note that the current maximum was calculated in the time interval after the transients of the excitation current decay. For comparison of the presented analysis results and the results from simplified model calculated in [4] the case when $R_D$=$R_M$ was calculated, too. This case is identical with the simplified model and the comparison is shown in Fig. 4.

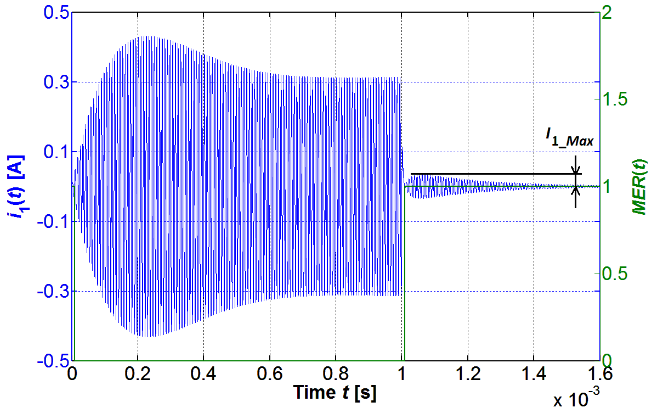


Fig. 2 Time course of the current $i_1(t)$ for $R_D$=1533 Ω and $R_M$=100 Ω

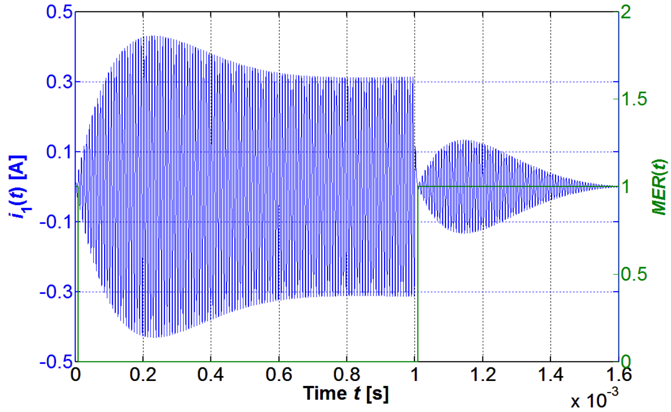


Fig. 3 Time course of the current $i_1(t)$ for $R_D$=1533 Ω and $R_M$=1 Ω

The measuring resistor $R_M$ influences the maximum current amplitude $I_{1\_Max}$ because this resistors acts as additional damping resistor for $L_RC_R$ circuit so that the ideal situation is when $R_M \rightarrow 0$.

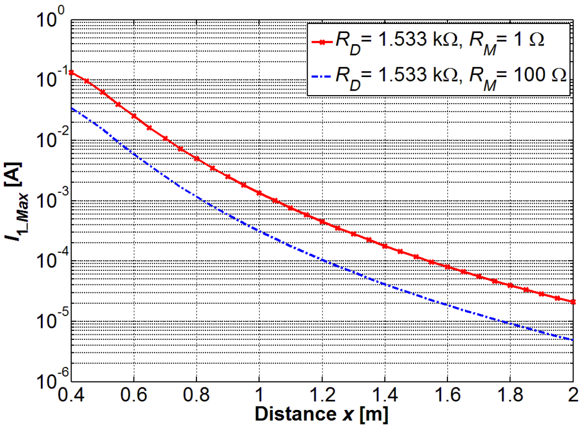


Fig. 4 Maximum amplitude of current response from marker with nominal resonant frequency $f_T$=125 kHz for two resistor model

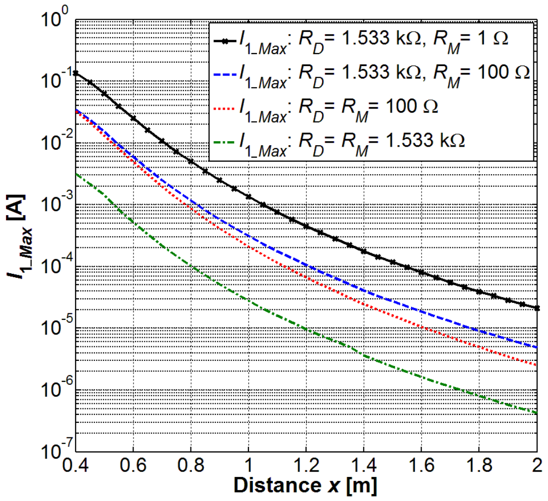


Fig. 5 Maximum amplitude of current response from marker for damped oscillation model

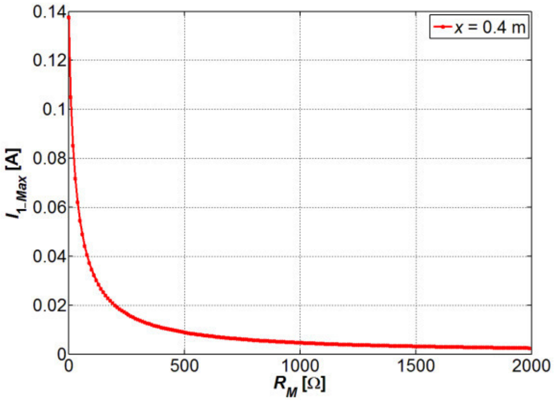


Fig. 6 Dependence of maximum current amplitude on the measuring resistor $R_M$ for distance $x$=0.4 m

## V. Mathematical Model Based On Continuous Excitation of Marker

This localization principle is used when the RFID locator continuously generates the magnetic field by its antenna coil $L_R$. The amplitude of current in locator circuit is then influenced by resonant circuit of marker so that the steady state current amplitude in locator circuit decreases if the marker is nearby.

The model is derived from the previous case by eliminating the switches. Schematic diagram of this model is shown in Fig. 7. In dependence on the distance *x* between coils the marker resonant circuit influences the amplitude of current $i_1(t)$ in locator antenna circuit.

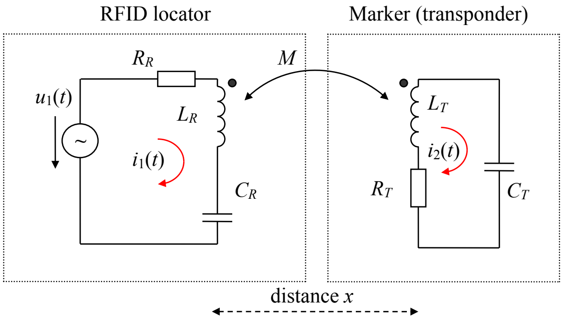


Fig. 7 Model of localization based on continuous excitation

This model can be analyzed by using of complex impedances of its components as it was performed in [4]. But such analysis did not show the transient phenomena in locator circuit and the results give only steady state solution. Therefore new model based on differential equations was created.

This model can be described by the following system of differential equations:

$$L_R\frac{di_1(t)}{dt}+R_Ri_1(t)+\frac{1}{C_R}\int_0^t i_1(\tau)d\tau-M\frac{di_2(t)}{dt}=u_1(\text{t})$$
$$L_T\frac{di_2(t)}{dt}+R_Ti_2(t)+\frac{1}{C_T}\int_0^t i_2(\tau)d\tau-M\frac{di_1(t)}{dt}=0 \qquad (9)$$

The signal source in Fig. 7 is harmonic as defined by (6). Similar as in previous chapter the system of integrodifferential equations (9) was transformed by substitution $x_1(t)=i_1(t)$, $x_2(t)=di_1(t)/dt$, $x_3(t)=i_2(t)$, $x_4(t)=di_2(t)/\mathrm{d}t$, $\omega=2\pi f$. Then we get the 1st order system of differential equations (10) which can be easily solved by standard mathematic software.

$$\begin{aligned}\frac{dx_1(t)}{dt}&=x_2(t)\\ \frac{dx_2(t)}{dt}&=a_1x_1(t)+a_2x_2(t)+a_3x_3(t)+a_4x_4(t)-a_5\omega U_1\cos(\omega t)\\ \frac{dx_3(t)}{dt}&=x_4(t)\\ \frac{dx_4(t)}{dt}&=b_1x_1(t)+b_2x_2(t)+b_3x_3(t)+b_4x_4(t)-b_5\omega U_1\cos(\omega t)\end{aligned} \qquad (10)$$

where individual coefficients are given by (11). Note that these coefficients are different from the coefficients given by (8).

$$\begin{aligned}a_1&=\frac{L_T}{C_R(M^2-L_RL_T)} & b_1&=\frac{M}{C_R(M^2-L_RL_T)}\\ a_2&=\frac{L_TR_R}{M^2-L_RL_T} & b_2&=\frac{R_RM}{M^2-L_RL_T}\\ a_3&=\frac{M}{C_T\left(M^2-L_RL_T\right)} & b_3&=\frac{L_R}{C_T\left(M^2-L_RL_T\right)}\\ a_4&=\frac{MR_T}{M^2-L_RL_T} & b_4&=\frac{L_RR_T}{M^2-L_RL_T}\\ a_5&=\frac{L_T}{M^2-L_RL_T} & b_5&=\frac{M}{M^2-L_RL_T}\end{aligned} \qquad (11)$$

In this case the detection of marker vicinity is more complicated as in the previous chapter because there is not the time stage in which only the response from excited marker can be received but the excitation signal in locator circuit is always present. The symptom of marker vicinity is decreasing of steady state current amplitude in locator circuit. When the distance between marker and locator is big ($x\rightarrow\infty$) then the mutual inductance calculated from (1) is very small ($M\rightarrow 0$). In this case the current $i_1(t)$ has maximum value $I_{1max}$, which is measured in steady state after the time $t$=2.5 ms. The marker vicinity then causes current drop $\Delta I_1$ (12) which can be calculated as difference between steady value of current $i_1(t)$ (see Fig. 8) and its maximum $I_{1max}$ measured in the same time point when no marker is nearby the locator.

$$\Delta I_1=I_{1max}-I_1 \qquad (12)$$

The time course of current $i_1(t)$ in the RFID locator circuit (Fig. 8) was numerically calculated from the system (10) for the same parameters as used in previous chapter. The dependence of current drop $\Delta I_1$ versus distance *x* is shown in Fig. 9.

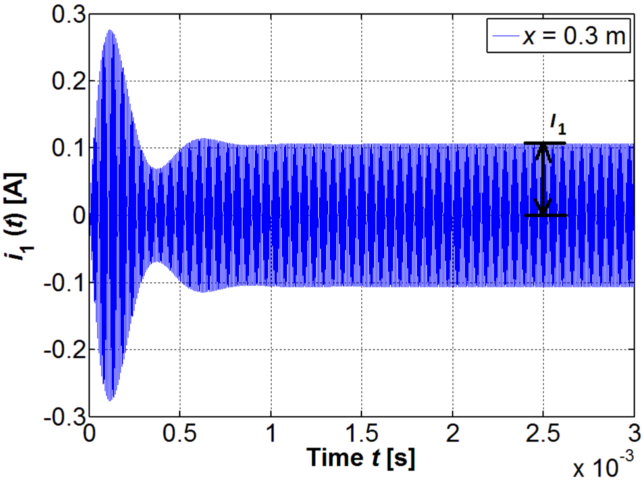


Fig. 8 The time course of current $i_1$(t), distance $x$=0.3 m

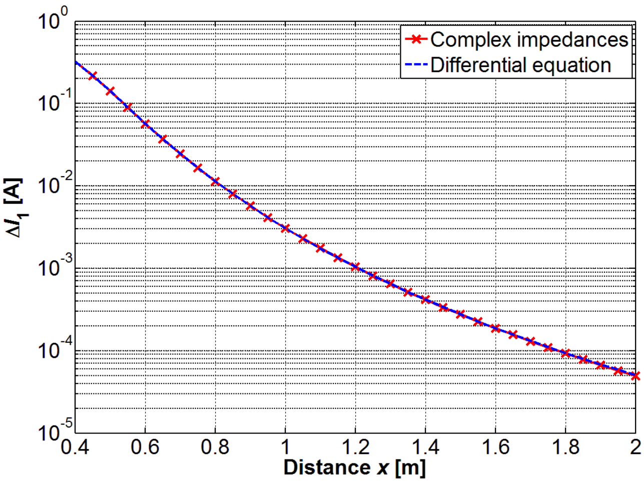


Fig. 9 Current drop $\Delta I_1$ versus distance $x$

The results calculated by analysis based on differential equations (10) were compared with the results obtained by analysis based on algebraic equations and complex impedances from [4]. This data is shown in the same graph in Fig. 9. As shown in this figure, both analyses give the same results.

## VI. Conclusion

Presented paper extends the mathematical analyses of marker localization principles. Two possible marker localization methods were analysed. For practical use the first method seems to be appropriate because the signal in receiving stage of localization can be directly processed.

The second method based on continuous marker excitation would require more complicated signal processing in locator. This complication is caused by "mixing" the excitation and response signals so that the marker vicinity cannot be simply detected by signal presence detection as in the first case but by detection of signal drop. Moreover the transient phenomena occur as documented in Fig. 8.

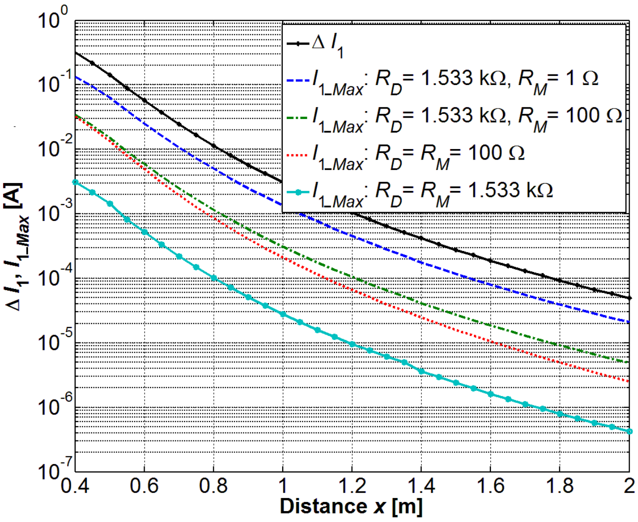


Fig. 10 Comparison of both localization models

Another criterion for evaluation of presented localization methods is based on their sensitivity to the marker vicinity. From comparison in Fig. 10 the second method seems to be more sensitive to the marker vicinity. Because the markers are typical near field application, the current amplitude in both cases decreases very rapidly when the distance between marker and locator increases.

The next research will be focused on the design of hardware needed to perform series of measurements so that the results of performed analyses will be compared with real measured data.

## Acknowledgment

This work has been supported by the Cultural and Educational Grant Agency of the Slovak Republic (KEGA) Number: 008ŽU-4/2015: "*Innovation of HW and SW tools and methods of laboratory education focused on safety aspects of ICT within safety critical applications of processes control*".

## References

[1] K. Finkenzeller, *RFID Handbook Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication,* Third Edition. John Wiley and Sons, Ltd. Chichester, UK, 2010, ISBN 978-0-470-69506-7. http://dx.doi.org/10.1002/9780470665121

[2] M. Y. Ahmad, A. S. Mohan, "RFID Reader Localization Using Passive RFID Tags". In: *APMC 2009, Asia Pacific Microwave Con*ference, Singapore, pp. 606 – 609, December 7th - 10th 2009, ISBN 978-1-4244-2802-1. http://dx.doi.org/10.1109/APMC.2009.5384152

[3] T. Van Haute, J. Rossey, P. Becue, E. De Poorter, I. Moerman, P. Demeester, "A Hybrid Indoor Localization Solution Using a Generic Architectural Framework for Sparse Distributed Wireless Sensor Networks". In: *FedCSIS 2014, Federated Conference on Computer Science and Information Systems*, September 7th – 10th, 2014, Warsaw, Poland, pp. 1009 – 1015, ISBN 978-83-60810-58-3. http://dx.doi.org/10.15439/2014F20

[4] P. Vestenický, T. Mravec, M. Vestenický, "Mathematical Modelling of Single-bit Passive RFID Marker Localization Methods". In: *ELEKTRO 2014, 10th International Conference*, May 19th – 20th, 2014, Rajecké Teplice, Slovakia, pp. 504 – 507, ISBN 978-1-4799-3720-2. http://dx.doi.org/10.1109/ELEKTRO.2014.6848946

[5] A. Markham, N. Trigoni, D. W. Macdonald, S. A. Ellwood, "Underground Localization in 3-D Using Magneto-Inductive Tracking". *IEEE Sensors Journal*, vol. 12, no. 6, pp. 1809 – 1816, June 2012, ISSN 1530-437X. http://dx.doi.org/10.1109/JSEN.2011.2178064

[6] T. E. Abrudan, A. Markham, N. Trigoni, "Poster Abstract: A Case for Magneto-Inductive Indoor Localization". In: *EWSN 2014, The 11th European Conference on Wireless Sensor Networks*, University of Oxford, Oxford, UK, pp. 18 – 19, February 17th -19th, 2014, ISBN 978-3-319-04651-8

[7] J. Hautcoeur, L. Talbi, M. Nedil, "High Gain RFID Tag Antenna for the Underground Localization Applications at 915 MHz Band". In: *APSURSI 2013, IEEE Antennas and Propagation Society International Symposium*, Orlando, FL, pp. 1488 – 1489, July 7th - 13th, 2013, ISBN 978-1-4673-5315-1. http://dx.doi.org/10.1109/APS.2013.6711403

[8] *RFID made easy*. Application note AN411, EM Microelectronic-Marin SA. Marin, Switzerland, 2002. [online] URL http://www.emmicroelectronic.com/webfiles/product/rfid/an/an411.pdf

DOI: 10.15439/2015F299
ACSIS, Vol. 5

# Drip Irrigation System using Wireless Sensor Networks

I. Bennis[*][†], H. Fouchal[†], O. Zytoune[*][‡], D. Aboutajdine[*]
[*]LRIT, unité associée au CNRST (URAC29), Université Mohammed V - Agdal, Rabat, Maroc
[†]Université de Reims Champagne-Ardenne, France
[‡]Université Ibn Tofail, Kénitra, Maroc
Email: ismail.bennis@etudiant.univ-reims.fr, hacene.fouchal@univ-reims.fr,
zytoune@univ-ibntofail.ac.ma, aboutaj@fsr.ac.ma

***Abstract*—Nowadays, adopting an optimized irrigation system has become a necessity due to the lack of the world water resource. Moreover, many researchers have treated this issue to improve the irrigation system by coupling the novel technologies from the information and communication field with the agricultural practices. The Wireless Sensor and Actuators Networks (WSANs) present a great example of this fusion. In this paper, we present a model architecture for a drip irrigation system using the WSANs. Our model includes the soil moisture, temperature and pressure sensors to monitor the irrigation operations. Specifically, we take into account the case where a system malfunction occurs, as when the pipes burst or the emitters block. Also, we differentiate two main traffic levels for the information transmitted by the WSAN, and we use an adequate priority-based routing protocol to achieve high QoS performance. Simulations conducted over the NS-2 simulator show promising results in terms of delay and Packet Delivery Ratio (PDR), mainly for the priority traffic.**

***Index Terms*—WSANs, Drip irrigation, Priority-based Routing.**

## I. Introduction

DURING the last decade, the Precision Agriculture (PA) has emerged as novel trend to enhance the agricultural practices. The principal aim of the PA is to monitor the spatio-temporal characteristics of the agricultural parcel [1]. By this way, the crops yield can be optimized while the natural, financial and energetic resources can be preserved. However, since the monitored agricultural regions are generally scattered and suffer from a variable environmental conditions, the need for accurate and real-time collected information is more pronounced. Also, the classical solution as the satellite imagery, aircraft or other systems based on the map cannot be supported by all farmers due to their heavy cost. To overcome this limitation, the Wireless sensor networks (WSNs) were introduced into the agricultural environment context [2].

Technically, the sensor nodes are deployed into the farmland. They start to collect environmental information and monitor soil characteristics. Then, they cooperate according to designed protocols to communicate the collected information to a central node. After that, this information is processed and treated to make an eventual decision.

The WSN have been explored in different ways for the agriculture field. As example, in [3] the authors have used four nodes types: soil, environmental, water and gateway to monitor the water content, temperature and soil salinity at farm located in Spain. Another work presented in [4] where authors have designed a node system for the collection of farmland information at different growth period of wheat, typically: seeding, jointing and heading. The study focuses in the optimal antenna height to use at the different growth period. Further examples presented in [5] and in [6] concern the greenhouse monitoring and the water saving irrigation using the WSN.

The security aspect is another example of how can the WSN improve the agricultural yield. In fact, crops are negatively affected by human or animal intruders. Also, the production process is still insufficiently controlled which lead to a potential product loss. To overcome this point, the video-surveillance nodes can be used to detect and identify intruders as well as to better take care of the production process [7]. In addition, the video-surveillance system allows the farmers to protect their sensors and equipment being installed in the crops from theft or potential damage.

One of the most important application of the WSNs in the PA is the irrigation system control. The interest comes naturally from saving water. For this aim, many researches were conducted to enhance the irrigation control system by coupling novel technologies with the agricultural practices. Among irrigation strategies, the drip irrigation system was considered as the most efficient policy to save water use. Moreover, combining this strategy with the WSNs leads us to have a great benefit from the farmlands. However, the irrigation system reliability need more attention, mainly in the case of general or partial dysfunction. For this aim, we present in this paper a model architecture for a drip irrigation system using the WSANs. Our model includes the soil moisture, temperature and pressure sensors to monitor the irrigation operations. Specially, we take into consideration the case when a system dysfunction occurs, as when the pipes are broken or the emitters are blocked. Also, we differentiate two main traffic levels for the information transmitted by the WSAN. Furthermore, based on our previous work [8], we can achieve a high QoS performance through an adequate priority-based routing protocol. The aim was to ensure an efficient and real-time communication between the different nodes type and the sink.

The remainder of this paper is organized as follows: in section II, we review some related works designed for an

978-83-60810-66-8/$25.00©2015, IEEE

efficient irrigation system. In section III, a description of our designed drip irrigation system is given. The priority-based protocol for DIS with simulation results are given in section IV. Finally, in section V, we draw the conclusion and give perspectives.

## II. Related work

To the best of our knowledge, monitoring the dysfunction of the drip irrigation system using the WSNs with an adequate priority-based routing protocol was never suggested before in the specialized literature. Therefore, in this section we summarize some related works for the irrigation system control.

In [9], authors propose an energy efficient method for the wireless sensor communication used in an automated irrigation system. This method is based on the Time Division Multiple Accesses (TDMA) scheduling that allows nodes to turn ON/OFF their radio according to scheduled slots. The main advantage of such scheme is saving the node's energy and reducing radio interference. Also, authors give a comparison between two methods to transmit the collected data to the sink node; namely the direct communication method and the data fusion method. For each method, the energy consumed and the data throughput are studied over the NS2 simulator.

To optimize water use in agricultural context, authors propose in [10] an automated irrigation system based in the WSNs technology. The developed system is composed of two kinds of sensors to collect soil-moisture and temperature information. The sensors are placed in the root zone of the plants. Also, a gateway was used to gather sensor information, triggers actuators, and transmits data to a web application. To control the water quantity, authors had programmed into a microcontroller an algorithm with threshold values of temperature and soil moisture. Concerning the energy, photo-voltaic panels are used to power the system. The entire system can be controlled through a web page which help to program an irrigation schedule and performs a data inspection.

In [11], authors present practical irrigation management system using a deployed WSN. This system includes a remote monitoring mechanism through a GPRS module to send SMS message containing land characteristic such as soil temperature and soil moisture, or the network performances such as packet delivery ratio, RSSI or the nodes energy level. The main contribution of this paper is to design and implement a low-cost efficient irrigation management system that combines sensors and actuators in a wireless sensor/actuator network. Authors conclude through this study that the deployment of the sensor nodes in the agricultural field is a critical issue. Furthermore, they suggest that the distance between sensor nodes has to be as short as possible in order to enhance the effectiveness of the system. However, the main weakness of this study is that authors employ only five sensors for the experiment.

We conclude for all referred works, that authors don't take into consideration the case of irrigation system dysfunction. Also they don't use the pressure sensor to monitor the irrigation flow rate. In addition, no priority-based protocol is designed to distinguish the importance of the communicated information. In the following section we present our proposed drip irrigation system that can overtake the dysfunction case.

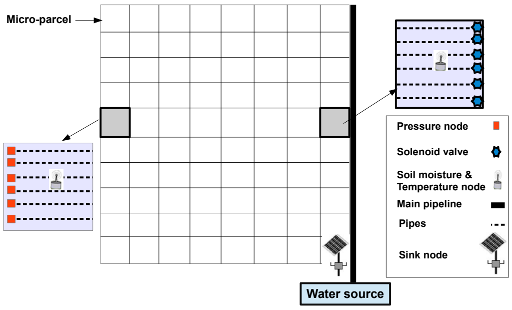


Fig. 1: Drip Irrigation System layout

## III. Proposed Drip Irrigation System Model

Recent practices in precision agriculture include two main micro irrigation methods which promote interesting water efficiency. The first method is the drip irrigation. It allows water to be dripped to the plants roots through pipes containing several emitters. This irrigation system is composed of the following components: water source (generally is a tank) which is connected with a main tube called main pipeline. To this line, several pipes are connected using manual or electrical valves that control the water flow. The pipes go through the field and distribute water for each plant.

The second method is the sprinkler irrigation which delivers water through a pressurized pipe network to the nozzles of sprinkler which spray the water into the air [12]. However, this method is less efficient than the drip one, since more water is losing due to evaporation and runoff. Therefore we choose the drip strategy for our design.

Our proposed model is a closed-loop model. As defined in [13], a system can be categorized as a closed-loop model if the response of the system is monitored and used to adjust the control. We note also that our proposed model is designed for a site-specific irrigation where the crops are characterized by a spatio-temporal variation of the irrigation requirements. The variability comes from the soil type, crop type, crop and meteorological conditions [13]. The main purpose of our design is to handle the dysfunctional situation of the drip installation. As discussed in [7], the crops are negatively affected by human or animals intruders. This is more critical in the case of drip irrigation installation. In fact, the pipes can be broken by rangers or by accident which can cause water waste and plants damage. Also the pipe emitters can be blocked due to environmental condition (sludge, sand) which can cause plant stress. To overtake these shortcoming the water flow rate into the drip installation must be monitored. For this aim, our proposed system include the following sensors and actuators:

- Soil moisture sensor: It is used to optimize irrigation and to warn of plant stress by controlling some parameters such as the electrical conductivity of soil or the underground volumetric water content (VWC). Measuring the soil moisture can help the farmers to manage their irrigation systems more efficiently by using less water to grow a crop and increasing quality and yields.

- Temperature sensor: It is used to monitor the ambient temperature. It can be analog or digital and help farmer to adjust their irrigation schedule according the temperature measured to avoid risk of evaporation.
- Pressure sensor: It is used to measure a pressure of gases or liquids and change it into a quantity that can be processed electronically. It generates a signal as a function of the pressure imposed. In irrigation application, this kind of sensor helps to monitor the abnormal pressure of pipe installation. In such case, by means of communication module (Zigbee/802.15.4), a message can be transmitted to the corresponding solenoid valve or the master valve (which control the main pipe) to shut down the system. A very low pressure value can be synonymous of a broken pipe or failure to open valves. Having a high pressure value can indicate that a valve is not closed correctly or some emitters are blocked.
- Solenoid valve: It is an electromechanical valve for use with liquid or gas controlled by running or stopping an electrical current through a solenoid, which is a coil of wire, thus changing the state of the valve [14]. Combined with a Zigbee module, the valve can be controlled trough wireless communication. Concerning the energy issue, the valve can have an external energy sources as solar panel.
- Sink node: It corresponds to the gateway of the system. All sensor nodes in the topology need to forward their gathered information to the sink node to be processed. Also, through this node, a request commands are generated to corresponding actuators or sensors.

An illustration of drip irrigation system with a deployment of the WSANs is shown in Fig .1

### A. Deployment strategy

Deploying the sensor nodes to monitor a farmland is crucial issue. In fact, many parameters must be considered to choose the most beneficial deployment, as the crops characteristics, the micro meteorological parameters, the sensors and nodes specification and obviously the farmer's budget. According to a generic guide proposed in [15] the coverage of the sensor nodes in agricultural WSN must be dense. By this way, all the required measurements can be gathered to have reliable knowledge of the monitored area. Authors in this guide argue that for a field with 100 $m^2$ size, at least 80-90 nodes are needed. They consider roughly 1 sensor node per 1 $m^2$. Of course, with such density we can reduce the sensors transmission power to the lowest level to save energy. In addition to have an adequate number of nodes, the topology formation must be determined. Among start, tree, or grid topology, the right choice depends to field's size and the plants formation. However, for middle or high surface, the grid topology remains the most suitable.

Based on the above discussion, we choose the grid topology for our drip irrigation design. We divide the field area into several equal micro parcel as suggested in [16]. The size of the parcel must be a trade-off between monitoring quality required, the communication coverage and the deployment cost. In the middle of each parcel we fix a soil moisture and temperature node. We make the assumption that the soil moisture and the temperature remain the same inside the parcel.

### B. Communication strategy

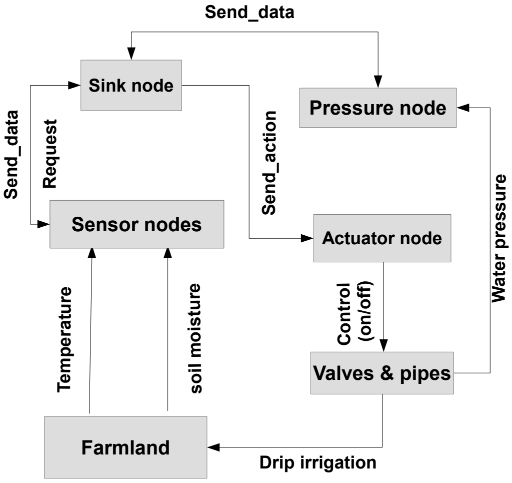


Fig. 2: Drip Irrigation System communication

In Fig. 2 we present a flowchart of the communication between all actors in the designed drip irrigation system. The sensor nodes gather the temperature and the soil moisture from the farmland periodically. According the value obtained, the sensor nodes decide to send the information to the sink or not. At the sink node, the abnormal information is processed and an eventual decision is taken to adjust the irrigation schedule according to the plant requirement. The same irrigation schedule is transmitted to the pressure nodes to be awakened at the same time of irrigation process. Once the actuators receive an action from the sink, they control their corresponding valves to be opened or closed. If the valves are opened, the water flow goes through the pipes and the pressure nodes start sensing. If any abnormal pressure value is gathered, an alert message is transmitted to the sink node to shut down the irrigation process and request an external human verification of the pipe installation.

We make the assumption that the sensor nodes communicate only with the sink node through a multi-hop protocol. Also, the actuators receive only actions from the sink. We assume also that the sink node can request some information from the sensor nodes at any time.

## IV. Priority-based DISM

### A. *Priority-based protocol*

As discussed in section III, we have two main traffic type gathered from sensors. The first one related to information gathered from temperature and the soil moisture sensors. We classify this traffic type as normal traffic since no need for an urgent intervention is required. The second traffic type is related to information gathered from pressure sensors. We classify this traffic type as priority traffic due to the need for an emergency resolution of the detected problem (shut off the main valve, require human intervention ... etc). Now, in the case when both traffics are active simultaneously, it is clear that the reliability and the timeliness of the priority traffic is more requested than those of the normal traffic.

However, in the wireless context, there are many troubles that can occur due to the sharing of the same communication medium. Among these problems we cite the interference problem, the exposed and the hidden problem [17]. Another problem that must be considered is the effect of the carrier sense range on communication performances. As discussed in our previous work [8], the carrier sense range is usually more larger than the transmission and the interference range. So carefully routing process must be applied to avoid any trouble between multiple sources and to satisfy the requested QoS for each traffic.

Let us take the example presented in Fig. 3. Two source nodes need to send their data to the sink. The first source node $A$ sends a priority traffic and the source node $B$ sends a normal traffic. We make the assumption that only one path is constructed from each source node. The circle presented around each node represents the transmission range. We avoid adding the carrier sense range in the figure to not overload it. As shown in Fig. 3, the black path refers to the path constructed from the node $A$ to the sink node. After that, the node $B$ needs to find out a valid path to reach the sink. If the red path is chosen, then all the nodes from the black path and the red one will be in concurrence to access to the communication medium which will degrade the final performance. The green lines represent relation between these nodes. To avoid such situation, the node $A$ must construct the blue path. Thus, even if the number of hops is higher the performances at the sink node are better. In what follow we will describe how the two paths can be constructed.

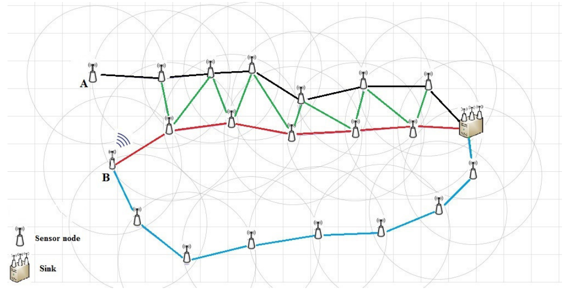


Fig. 3: Carrier sense effect in the case of multi-sources

### B. *Protocol description*

Based on our previous work [8], we design a routing protocol that can allow the priority source node (namely the pressure node) to construct an efficient routing path while avoiding the carrier sense range effect. In this work we make the assumption that nodes are aware of their positions and the position of the sink node. In the following, we give a short description of how the paths are constructed according to our approach.

When a priority source node seeks to communicate with the destination, it sets up a route discovery process by sending a priority forward agent (P-FAGT) to construct a short multi-hop path. The choice of the next hop node is based on the geographic information available at each node. For each selected node $i$, the node state is changed from free to busy, and a Hello message is broadcasted to all neighbors in the communication range to notify the new state of the node $i$. Every neighbor node $j$ of the node $i$ becomes a banish node, that means it cannot be selected for any communication. After that, each node $j$ broadcasts in its turn a hello message in their neighborhood. Now, if a normal source node needs to send some information, it constructs the routing path by sending a normal forward agent (N-FAGT) which must respect the following rules: the next hop must not be blocked, and must not be a banish node or having a banish node in its neighborhood. A node is in a blocked state when the destination is unreachable through this node. To avoid a blocking situation when a node cannot reach the destination, we use the same principle as in [18], called the step-back method. The same method is used by the agent when the selected next hop has a banish node in its neighborhood.

Once the destination is reached, the forward agent (either P-FAGT or N-FAGT) becomes a backward agent and an optimized reverse path is travelled. At each intermediate node, the agent records the valid next hop into the routing table, after that, it chooses from the reverse path the nearest neighbor to the current node. The same procedure is repeated until reaching the source node.

In the case where the P-FAGT finds an already constructed path (used by a normal source), it follows this path and changes the state information of all nodes involved in it. After that, the normal source is informed by a special agent to start another discovery phase to take into consideration the current priority source communication.

When the communication is ended, all the nodes altered by the communication process reset their state and become ready for further transmissions. In the remaining of this paper, we denote our approach by Carrier Sense Aware (CSA).

### C. *Simulation & result analysis*

#### *Working environment*

Our simulation scenario is based on the topology presented in Fig.1. The topology area is 200*200 $m^2$, and the total number of nodes is 280 (including pressure, temperature/soil moisture and valve nodes). We make the assumption that the micro parcel size is 20 $m^2$. Two random source nodes (pressure and temperature/soil moisture) are selected and start transmission at different instance but in the same interval time. To distinguish the two traffic in the simulation in the NS2 simulator [19], we choose for the temperature/soil moisture source node a constant bit rate (CBR) traffic with **X** packets per second, where:

$$X=\{8,16,24,32\}$$

For the pressure source node, we choose an exponential traffic (Exp) with a data rate equals to 20 Kbytes. The duration of communication is 30 s, and no mobility is supported in this scenario. For every value of **X**, 20 scenarios are generated and the average value of results is computed. We present the results with a confidence interval of 95%. According to the characteristic of the MicaZ node [20] and the two-ray-ground propagation model equation, we define the reception and the carrier sense threshold (RXThreshold and CSThreshold). Their respective value was $3.981{*}10^{-13}$ Watt and $3.981{*}10^{-14}$ Watt which represent nearly 20 m for the transmission range and 35 m for the carrier sense range. We compare our work with the Two Phase geographical Greedy Forwarding (TPGF) protocol [18] since it adopts also a geographical approach.

Table I summarizes the parameters used for simulation.

TABLE I: Main configuration parameters

| Parameters | value |
|---|---|
| link layer | LL |
| MAC layer | IEEE 802.15.4 |
| radio propagation | two ray ground |
| interface queue | PriQueue |
| ifqlen | 50 |
| antenna | omni-antenna |
| Antenna height (m) | 0.0864 |
| Frequency | 2.4 GHz |
| $CPThreshold$ (dB) | 10 |
| $CSThreshold$ (Watt) | $3.981 * 10^{-14}$ |
| $RXThreshold$ (Watt) | $3.981 * 10^{-13}$ |
| Pt (Watts) | 0.001 |
| Packet size | 100 Bytes |

*Result analysis*

We start our analysis by studying the PDR metrics for both normal and priority traffic. From Fig. 4, which represents the PDR for the normal traffic, we can see that, for both protocols, the PDR decreases as the number of Packet Per Second (PPS) increases. It is quite expected since more the traffic is higher more the collision likelihood is higher too. For the CSA protocol, it has the higher PDR between 8 and 24 PPS, after that, the PDR becomes slightly lower than that of the TPGF. We explain such behavior by the constructed paths of each protocol. In fact, the CSA protocol builds path for the normal traffic while avoiding any banish node as described in subsection IV-B. Concerning the TPGF path, it builds paths according to the greedy forwarding mechanism, so shorter paths are constructed. Therefore, the number of hops in the case of the CSA is higher than that of the TPGF. We can tolerate such performance, since the normal traffic is usually loss-tolerant.

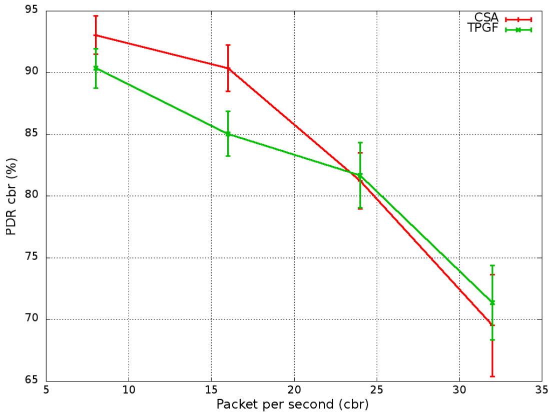


Fig. 4: Average PDR of normal traffic vs packet per second

For Fig. 5, which represents the PDR for the priority traffic, we can see also that, for both protocols, the PDR decreases as the number of PPS increases. However, it is clear that the CSA protocol achieves higher PDR than the TPGF. The PDR gain can reach 20%. For the TPGF, it provides a poor PDR value mainly when the traffic rate increases. Such performance cannot be acceptable for the priority traffic, which is almost loss-intolerant. In fact, as discussed in subsection IV-A, the carrier sense range effect occurs when a node cannot transmit when another node in its carrier sense range is already in transmitting phase. Therefore, when the number of PPS is higher, the nodes of all paths deprive mutually the channel access since there is a competition between them. Thus, the likelihood of loss packet is more pronounced.

In Fig. 6, the delay for both protocols in the case of priority traffic is depicted. As first observation, we can see that the delay increases as the number of PPS increases. We can see also that the CSA protocol provides a lower delay compared to the TPGF protocol. It is quite expected since the construction path process in the CSA protocol, ensures that the priority traffic will not be disturbed by any communication in the neighborhood. Such performance is suitable for the priority traffic which is usually delay sensitive.

## V. Conclusion and Future Work

In this paper, we have presented a model architecture for a drip irrigation system using the WSANs. Our model includes the soil moisture, temperature and pressure sensors to monitor the irrigation operations. Specially, we take into account the case where a system malfunction occurs, as when the pipes are broken or the emitters are blocked. Also, we differentiate two main traffic levels for the information transmitted by the WSAN, and based on our previous work, we achieve a high

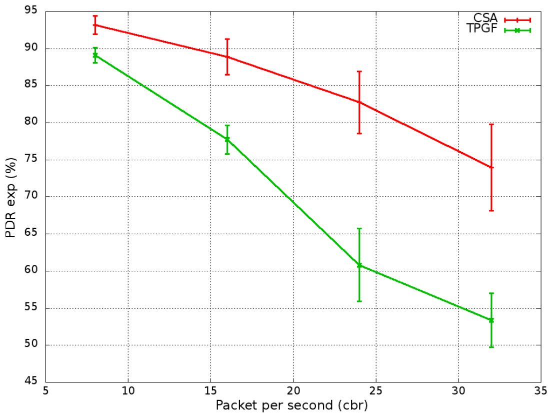


Fig. 5: Average PDR of priority traffic vs packet per second

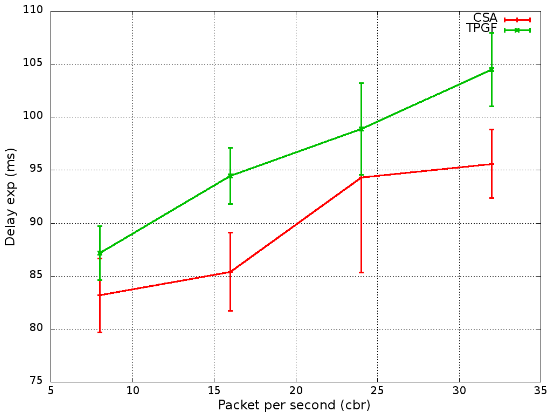


Fig. 6: Average delay of priority traffic vs packet per second

QoS performance through an adequate priority-based routing protocol.

We have performed extensive simulations. The results prove that our solution gives better performances in terms of delay, PDR for the priority traffic. As a future work, we intend to realize a real test-bed to investigate the effectiveness of our approach.

## Acknowledgment

This work is partially supported by ”Projet de coopération Maroco-Française : Contribution á l’optimisation de la qualité de service dans les réseaux de capteurs sans fil: Application à la supervision agricole 2013-2014”

## References

[1] Alberto Camilli, Carlos E. Cugnasca, Antonio M. Saraiva, André R. Hirakawa, and Pedro L.P. Corrêa. From wireless sensors to field mapping: Anatomy of an application for precision agriculture. *Computers and Electronics in Agriculture*, 58(1):25 – 36, 2007. Precision Agriculture in Latin America.

[2] Ning Wang, Naiqian Zhang, and Maohua Wang. Wireless sensors in agriculture and food industry?recent development and future perspective. *Computers and Electronics in Agriculture*, 50(1):1 – 14, 2006.

[3] J.A. López Riquelme, F. Soto, J. Suardíaz, P. Sánchez, A. Iborra, and J.A. Vera. Wireless sensor networks for precision horticulture in southern spain. *Computers and Electronics in Agriculture*, 68(1):25 – 35, 2009.

[4] Xiaoqing Yu, Pute Wu, and Zenglin Zhang. Design and test of nodes for field information acquisition based on wusn. In Baoxiang Liu and Chunlai Chai, editors, *Information Computing and Applications*, volume 7030 of *Lecture Notes in Computer Science*, pages 561–568. Springer Berlin Heidelberg, 2011.

[5] Minghua Shang, Guoying Tian, Leilei Qin, Jia Zhao, Huaijun Ruan, and Fengyun Wang. Greenhouse wireless monitoring system based on the zigbee. In Daoliang Li and Yingyi Chen, editors, *Computer and Computing Technologies in Agriculture VI*, volume 392 of *IFIP Advances in Information and Communication Technology*, pages 109–117. Springer Berlin Heidelberg, 2013.

[6] M. Nesa Sudha, M.L. Valarmathi, and Anni Susan Babu. Energy efficient data transmission in automatic irrigation system using wireless sensor networks. *Computers and Electronics in Agriculture*, 78(2):215 – 221, 2011.

[7] Antonio-Javier Garcia-Sanchez, Felipe Garcia-Sanchez, and Joan Garcia-Haro. Wireless sensor network deployment for integrating video-surveillance and data-monitoring in precision agriculture over distributed crops. *Computers and Electronics in Agriculture*, 75(2):288 – 303, 2011.

[8] Ismail Bennis, Hacene Fouchal, Ouadoudi Zytoune, and Driss Aboutajdine. Carrier sense aware multipath geographic routing protocol. *Wireless Communications and Mobile Computing*, pages n/a–n/a, 2015.

[9] M. Nesa Sudha, M.L. Valarmathi, and Anni Susan Babu. Energy efficient data transmission in automatic irrigation system using wireless sensor networks. *Computers and Electronics in Agriculture*, 78(2):215 – 221, 2011.

[10] J. Gutierrez, J.F. Villa-Medina, A. Nieto-Garibay, and M.A. Porta-Gandara. Automated irrigation system using a wireless sensor network and gprs module. *Instrumentation and Measurement, IEEE Transactions on*, 63(1):166–176, Jan 2014.

[11] Million Mafuta, Marco Zennaro, Antoine B. Bagula, Graham W. Ault, Harry Sam Harrison Gombachika, and Timothy Chadza. Successful deployment of a wireless sensor network for precision agriculture in malawi. *IJDSN*, 2013, 2013.

[12] H. Ali. *Practices of Irrigation & On-farm Water Management: Volume 2*. Practices of Irrigation & On-farm Water Management. Springer, 2011.

[13] AlisonC. McCarthy, NigelH. Hancock, and StevenR. Raine. Advanced process control of irrigation: the current state and an analysis to aid future development. *Irrigation Science*, 31(3):183–192, 2013.

[14] J.J. Haller, S.P. Glaudel, and G.J. Volz. System and method of operating a solenoid valve at minimum power levels, September 26 2013. US Patent App. 13/900,683.

[15] I. Mampentzidou, E. Karapistoli, and A.A. Economides. Basic guidelines for deploying wireless sensor networks in agriculture. In *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2012 4th International Congress on*, pages 864–869, Oct 2012.

[16] Wei An, Song Ci, Haiyan Luo, Dalei Wu, Viacheslav Adamchuk, Hamid Sharif, Xueyi Wang, and Hui Tang. Effective sensor deployment based on field information coverage in precision agriculture. *Wireless Communications and Mobile Computing*, pages n/a–n/a, 2013.

[17] Lu Wang, Kaishun Wu, and M. Hamdi. Combating hidden and exposed terminal problems in wireless networks. *Wireless Communications, IEEE Transactions on*, 11(11):4204–4213, November 2012.

[18] Lei Shu, Yan Zhang, LaurenceT. Yang, Yu Wang, Manfred Hauswirth, and Naixue Xiong. Tpgf: geographic routing in wireless multimedia sensor networks. *Telecommunication Systems*, 44(1-2):79–95, 2010.

[19] VINT. The network simulator ns-2.34, 2012.

[20] memsic. Micaz datasheet. http://www.memsic.com/userfiles/files/Datasheets/WSN/micaz_datasheet-t.pdf.

# Information Technology for Management, Business & Society

IT4MBS is a FedCSIS conference aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the disciplines of information technology and information systems. The IT4BMS area emphasizes the issues relevant to information technology and necessary for practical, everyday needs of business, other organizations and society at large. This area takes a sociotechnical view on information systems and relates also to ethical, social and political issues raised by information systems. Events that constitute IT4BMS are:

- **ABICT'15** - 6th International Workshop on Advances in Business ICT
- **AITM'15** - 13th Conference on Advanced Information Technologies for Management
- **ISM'15** - 10th Conference on Information Systems Management
- **IT4L'15** - 4th Workshop on Information Technologies for Logistics
- **KAM'15** - 21st Conference on Knowledge Acquisition and Management

# 6th International Workshop on Advances in Business ICT

ABICT focuses on Advances in Business ICT approached from a multidisciplinary perspective. It will provide an international forum for scientists/experts from academia and industry to discuss and exchange current results, applications, new ideas of ongoing research and experience on all aspects of Business Intelligence. ABICT will be also an opportunity to demonstrate different ideas and tools for developing and supporting organizational creativity, as well as advances in decision support systems.

We kindly invite contributions originating from any area of computer science, information technology and computational solutions for different applications areas, data integration and organizational implementation of ABICT, as well as practical ABICT solutions.

## Topics

Topics include (but are not limited to):

- Advanced technologies of data processing, content processing and information indexing
- Analytics as a service
- Big Data: benefits and challenges
- Business Analytics
- Business applications of social networks
- Business data mining and knowledge discovery
- Business Intelligence
- Business Rules
- Business-oriented time series data mining, analysis, and processing
- Cloud based Business Intelligence
- Creativity Support Tools
- Customer Relationship Management, social Customer Relationship Management
- Data driven marketing
- Data Warehousing
- Decision support
- Digital Business Strategy
- Enterprise Device Management
- ICT technologies in enterprise management
- Information forensics and security, information management, risk assessment and analysis
- Information Systems Design
- Internet of Things
- Knowledge Management (for better Decision Support, Collaboration and Competitiveness)
- Legal text processing
- Leveraging ICT for Transforming Organization
- M2M Device Management, M2M Solutions
- Semantic Web and Ontologies in Business ICT
- Virtual Enterprise
- Web 2.0 and Web 3.0 in fusing Business Intelligence systems and Decision Support Systems
- Web-Based Data Management Systems

## Event Chairs

**Mach-Król,** Maria, University of Economics in Katowice, Poland

**Olszak, Celina M.,** University of Economics in Katowice, Poland

**Pełech-Pilichowski, Tomasz,** AGH University of Science and Technology, Poland

## Program Committee

**Abramowicz, Witold,** Poznan University of Economics, Poland

**Badica, Amelia,** University of Craiova, Romania

**Berio, Giuseppe,** Universite de Bretagne Sud, France

**Chiu, Dickson K. W.,** Dickson Computer Systems, Hong Kong S.A.R., China

**Christozov, Dimitar,** American University in Bulgaria, Bulgaria

**Gaweł, Bartłomiej,** AGH University of Science and Technology

**Kacprzyk, Janusz,** Institute of Computer Science, Polish Academy of Sciences, Poland

**Khachidze, Manana,** Tbilisi State University, Georgia

**Konikowska, Beata,** Institute of Computer Science, Poland

**Korwin-Pawlowski, Michael L.,** Universite du Quebec en Outaouais, Canada

**Kulczycki, Piotr,** Systems Research Institute, Polish Academy of Sciences, Poland

**Loucopoulos, Peri,** Harokopio University of Athens, Greece

**Madeyski, Lech,** Wrocław University of Technology

**Ogihara, Mitsunori,** University of Miami, United States

**Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland

**Petryshyn, Lubomyr,** AGH University of Science and Technology, Poland

**Prasad, T. V.,** Visvodaya Technical Academy, India

**Pulvermueller, Elke,** University Osnabrueck, Germany

**Reimer, Ulrich,** University of Applied Sciences St. Gallen, Switzerland

**Rossi, Gustavo,** National University of La Plata, Argentina

**Salem, Abdel-Badeeh M.,** Ain Shams University, Egypt

**Sauer, Jurgen,** University of Oldenburg, Germany

**Szpyrka, Marcin,** AGH University of Science and Technology, Poland

**Teufel, Stephanie,** University of Fribourg, Switzerland

**Zieliński, Jerzy S.**

**Zurada, Jozef,** College of Business University of Louisville, Louisville

DOI: 10.15439/2015F250
ACSIS, Vol. 5

# Opportunities for Business Process Semantization in Open-source Process Execution Environments

Krzysztof Kluza and Krzysztof Kaczor and Grzegorz J. Nalepa and Mateusz Ślażyński
AGH University of Science and Technology
al. A. Mickiewicza 30, 30-059 Krakow, Poland
E-mail: {kluza,kk,gjn,mslaz}@agh.edu.pl

***Abstract*—Business Process models help to visualize the processes of an organization. There are open-source process execution engines that provide the environments for enacting process models. However, they lack semantization capabilities. In this paper, an overview of Business Process semantization techniques is provided. Moreover, we discuss the common architecture of the selected open-source process execution environments (Activiti, jBPM and Camunda) and provide the insights how they can be improved using semantization methods. We also present the use of the introduced techniques in the Prosecco (Processes Semantics Collaboration for Companies) research project.**

***Index Terms*—Business Process Model and Notation (BPMN), Business Process Semantization, Business Process Management**

## I. Introduction

BUSINESS Process Management (BPM) [1] is a modern approach to improve organization's workflow, which focuses on reengineering of processes in order to optimize procedures, increase efficiency and effectiveness by constant process improvement. A Business Process (BP) model constitute a graphical representation of a process in an organization, which is composed of related tasks that produce a specific service or product for a particular customer [2].

The BP runtime environment manages and monitors processes as they perform. It orchestrates the activities and interactions of the process with web services or other third-party applications. Such software also supports user tasks, handles exceptions or escalations by tracing the workflow in the model. There are more than 70 BPMN implementers[1],

As process models can be ambiguous, the top leader vendors that develop proprietary software introduced semantization techniques to their solutions. This helps to support more intelligent functions, like web services discovery or element name suggestions. These techniques use semantic annotations, which can be based on a formally specified ontology.

For the purpose of this paper, we focused on the selected open-source process execution environments that are freely accessible, use the BPMN 2.0 notation [3] on the executable level and can be easily extended with new functions. We analyze the possible ways of enhancing them with semantization methods. Some of these methods have been introduced in the Prosecco (Processes Semantics Collaboration for Companies) research project[2], which takes advantage of the Activiti execution engine.

The paper is structured as follows. In Section II, we present selected open-source process execution environments and their general architecture. Section III gives an overview of business process semantization approaches, especially focusing on the solution developed in the SUPER project and the SAP AG company. Section IV analyzes the possibilities of process semantization in the open-source process execution environments. The paper is concluded in Section V.

## II. Open-source Process Execution Environments

All three environments, Activiti[3] [4], jBPM[4] [5], and Camunda[5] [6], are light-weight BPM suites with extensible BPMN 2.0 process engines. They can be run in any Java environment, embedded in an application or as a service. The engines allow a user to execute business processes using the BPMN 2.0 specification, and are open-source software distributed under the Apache license.

All the projects include such components as:

- Process Engine – an execution engine that provides Process Virtual Machine; the engine uses BPMN 2.0 as the underlying XML format for the process definitions.
- Web-based Process Modeler – a web-based process editor for modeling business processes (mainly for analysts).
- Eclipse-based Process Designer – more complex process editor for modeling and implementing detail aspects of process models, it allows graphical modeling, development and debugging of models.
- Process Management Interface – a web application providing access to the runtime engine for all users of the system, this includes task and job management, process instance inspection, viewing reports based on statistical history data, etc.
- Process Repository – a repository for storing and managing process definitions.
- History Log – a log for storing history information about the process instances that are being executed, which can be further used to generate reports, etc.

The overview of this common architecture is presented in Figure 1 and comparison of their components is provided in Table I. Such an architecture is further considered for semantization of the execution environment.

[1] See: http://www.bpmn.org/#tabs-implementers.

[2] See: http://prosecco.agh.edu.pl.

[3] See: http://www.activiti.org/userguide/.

[4] See: http://www.jbpm.org/learn/documentation.html.

[5] See: http://docs.camunda.org/guides/user-guide/.

978-83-60810-66-8/$25.00©2015, IEEE

| L.p. | Component | Activiti | jBPM | Camunda |
|---|---|---|---|---|
| **1.** | **Process Engine** | Activiti Engine | Core Process Engine | Camunda Process Engine |
| **2.** | **Process Repository** | part of Activiti Engine | Process Repository | part of Camunda Process Engine |
| **3.** | **Web-based Process Modeler** | Activiti Modeler | Web-based Designer | Camunda-bpmn.js |
| **4.** | **Eclipse-based Process Designer** | Activiti Designer | Eclipse-based Editor | Camunda Modeler |
| **5.** | **Process Management Interface** | Activiti Explorer | jBPM Console | Camunda Cockpit/Tasklist |
| **6.** | **History Log** | part of Activiti Engine | History Log | part of Camunda Process Engine |

Table I
COMPARISON OF ACTIVITI, JBPM AND CAMUNDA COMPONENTS

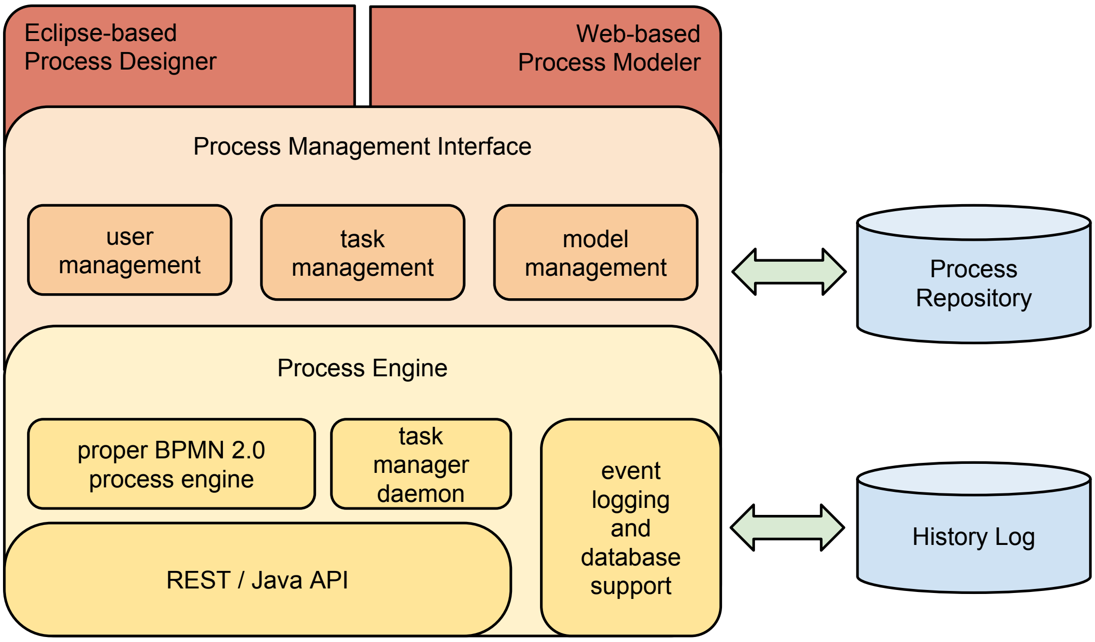

Eclipse-based Process Designer
Web-based Process Modeler
Process Management Interface
user management
task management
model management
Process Repository
Process Engine
proper BPMN 2.0 process engine
task manager daemon
event logging and database support
REST / Java API
History Log

Figure 1. The general architecture of the analyzed environments (based on our comparison of the Activiti, jBPM and Camunda environments)

## III. OVERVIEW OF BUSINESS PROCESS SEMANTIZATION POSSIBILITIES

This section presents the overview of the semantization approaches for business processes and business process runtime environments, based on the analysis of the research results in this field. It focuses on Business Process semantization in: the SUPER project, the BPM product of the SAP AG company, as well as some substitute of semantization which can be observed in Signavio Process Editor.

### A. *Business Process Semantization in the SUPER project*

The goal of the SUPER project [6] was to create tools for business process semantization by describing process models using concepts from the ontology.

A BPM system that uses ontology as a common language of communication can facilitate clear expressing the statement by the people from the business and provide a method of unambiguous communication between the system, IT people and non-technical people associated with the business.

*The SUPER project architecture*

The elements of the architecture of the SUPER environment is presented in Figure 2. WSMO Studio is a stand-alone application (also available as a plugin for the Eclipse integrated development environment), which provides the following functions: enables customers to create ontologies, specify goals, web services and mediators, as well as provide appropriate interfaces for these elements. Additionally, the environment provides dedicated editors, including SAWSDL editor to annotate semantics to WSDL.

*Ontologies in the SUPER project*

In the SUPER project ontology the following elements can be distinguished (the relationship between them is presented in Figure 3):

- Web Service Modeling Ontology [7] specifies formally the terminology of the information used by all other components and provides the semantic description of web services (their functional and non-functional properties, and their interfaces).
- Business Domain Ontologies related to the business domain knowledge (Business Functions, Business Process Resources, Business Roles and Business Modeling Guidelines Ontologies).

[6] The website of the SUPER project http://www.ip-super.org/content/view/196/163/ is no longer maintained. However, some pieces of information about the results can be found at: http://www.sti-innsbruck.at/results/movies/sbp-execution-developed-in-super as well as in the project publications.

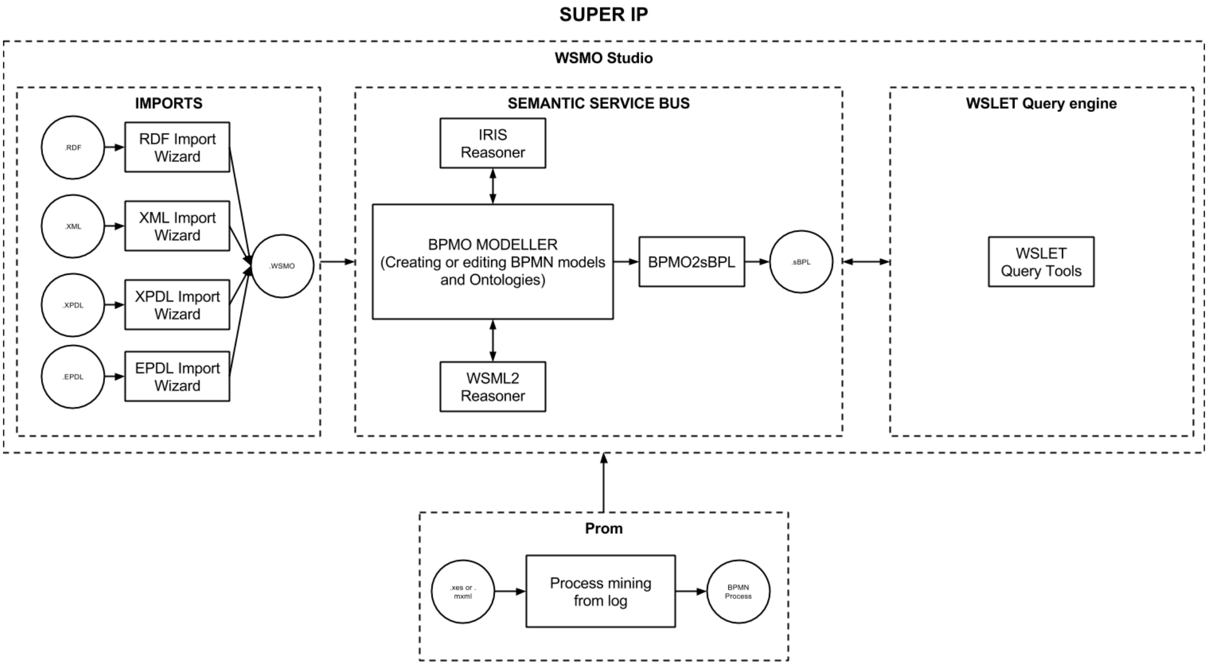


Figure 2. The elements of the basic architecture of the SUPER project (based on the description from the SUPER project website)

- SUPER Ontology Stack:
  - Upper-Level Process Ontology (UPO) is the top-level ontology that aims to represent high-level concepts for Business Process modelling.
  - Business Process Modelling Ontology (BPMO) acts as a bridge between the business level and the processes execution level and is used for representing high-level business process workflows.
  - Semantic Event-driven Process Chains notation Ontology (sEPC) supports the annotation of process models created with EPC tools.
  - Semantic Business Process Modeling Notation Ontology (sBPMN) for formalization of the core subset of the BPMN notation.
  - Semantic BPEL Ontology (sBPEL) extends the ontology of BPEL with a Semantic Web Services model.
  - Behavioral Reasoning Ontology (BRO) for reasoning over the business processes behaviours using WSML axioms.
  - Events Ontology (EVO) that constitutes a reference model for capturing logging information used by the execution engines and the analysis tools.

*The SUPER project process life cycle*

The methodology of the SUPER project defines four phases that form the business process life cycle [9]. For these phases the appropriate methods and techniques for business process semantization were developed. In the following paragraphs, these phases are elaborated.

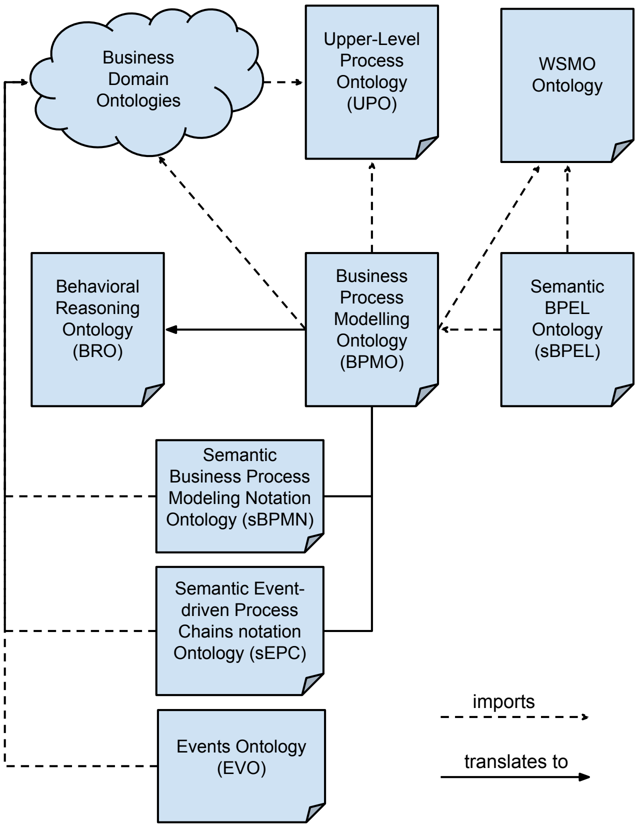

Business Domain Ontologies
Upper-Level Process Ontology (UPO)
WSMO Ontology
Behavioral Reasoning Ontology (BRO)
Business Process Modelling Ontology (BPMO)
Semantic BPEL Ontology (sBPEL)
Semantic Business Process Modeling Notation Ontology (sBPMN)
Semantic Event-driven Process Chains notation Ontology (sEPC)
Events Ontology (EVO)
imports
translates to

Figure 3. The overview of the ontologies in the SUPER project (based on [8])

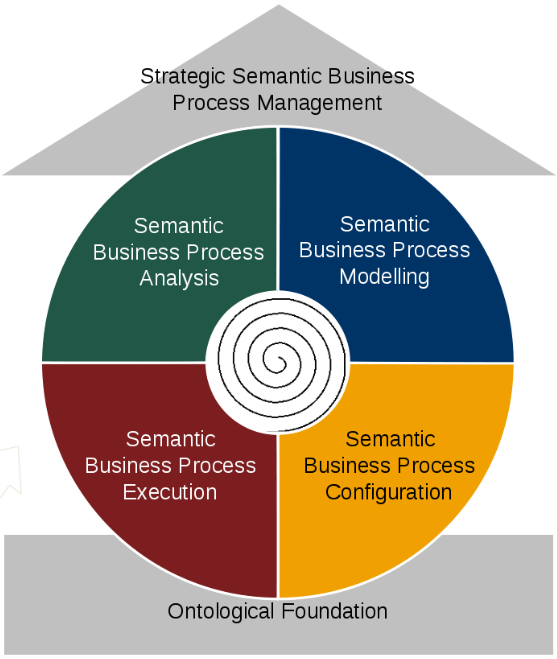

Strategic Semantic Business Process Management
Semantic Business Process Analysis
Semantic Business Process Modelling
Semantic Business Process Execution
Semantic Business Process Configuration
Ontological Foundation

Figure 4. Phases in the SUPER project methodology [9]

*Phase 1: Semantic Business Process Modelling*

The first phase of the life cycle in the SUPER methodology consists in developing business process models based on the BPMO ontology. It uses the environment for semantic modeling (WSMO Studio tools with the integrated BPMO editor). The business process model is based on the domain ontologies specified for the particular company as well as on Semantic Web Services and Goals. Its source can be implicit knowledge of business analysts or analysis of reports from the previous Semantic Business Process (SBP) Analysis phase.

*Phase 2: Semantic Business Process Configuration*

Semantic Business Process Configuration (SBPC) is a second phase in the SUPER methodology life cycle that uses the semantic business process models which are the output from the previous phase. During this phase, the semantic business process models are configured.

The configuration phase consists of deriving an sBPEL ontology from a BPMO instance, discovering the possible Semantic Web Services (SWS) [10], identifying the potential data mismatches, and based on them creating the interface mappings and data mediators. Lastly, the process is validated in terms of the correctness of the semantic process description before the execution and potentially refined.

*Phase 3: Semantic Business Process Execution*

In the third phase of SUPER methodology, modeled and configured processes are executed and processed. During the runtime, data, which will be used for analysis, are collected. As this phase is performed without user interactions, it minimizes the time required for its completion. In this pahse, process execution is supported by the semantic BPEL (BPEL4SWS) and detection and execution of Semantic Web Services (SWS).

In Figure 5, the scenario of semantic business process execution is presented. This scenario involves the following seven steps [9]:

1) *Request Service* – in order to initialize a semantic BPEL process, a user have to send request through the Semantic Service Bus to SBPELEE.
2) *Achieve Goal* – invocation of SWS is delegated to SEE by SBPELEE which passes the WSMO Goal to it.
3) *Discover Service* – SEE queries the Semantic Web Services repository to discover the desired SWS.
4) *Invoke Service* – SEE invokes the discovered SWS.
5) *Engine Return Result* – SEE returns the result received from SWS to SBPELEE.
6) *User Return Result* – After the process execution has been finished, the result is returned to the user.
7) *Process Tracking* – During the execution, execution events are published to Execution History for persistence and to the Monitoring Tool for tracking process executions.

The most important benefits of using such an approach are:

- flexible use of Web Services,
- supplier matching supported by Semantic Web Service discovery and invocation from within semantic business processes,
- more flexible traffic routing,
- automates supplier matching and traffic routing process taking into account all existing suppliers,
- minimizes time-to-offer.

*Phase 4: Semantic Business Process Analysis*

The last phase of the process life cycle concerns the analysis of the executed processes. In this phase, various analysis goals are supported, such as: overview over process usage, detecting business and technical exceptions, etc.

Thanks to this phase, it is possible to get an overall overview about system usage, finding out exceptions within process flow and bottlenecks, as well as get necessary information needed to apply 6 Sigma methodology.

*Data, Information, and Process Integration with SWS*

The SUPER project took advantage of the experience of the DIP[7] (Data, Information, and Process Integration with Semantic Web Services) [11] platform.

The aim of the DIP platform was extending the semantic web technologies and web services in order to create a new technical infrastructure – Semantic Web Services (SWS).

The DIP platform provides: Web Service Modelling Ontology (WSMO), Web Service Modeling Language (WSML) as the language for modeling web services, and Web Service Execution Environment (WSMX) as software framework for runtime binding of service requesters and service providers.

[7] See: http://dip.semanticweb.org/.

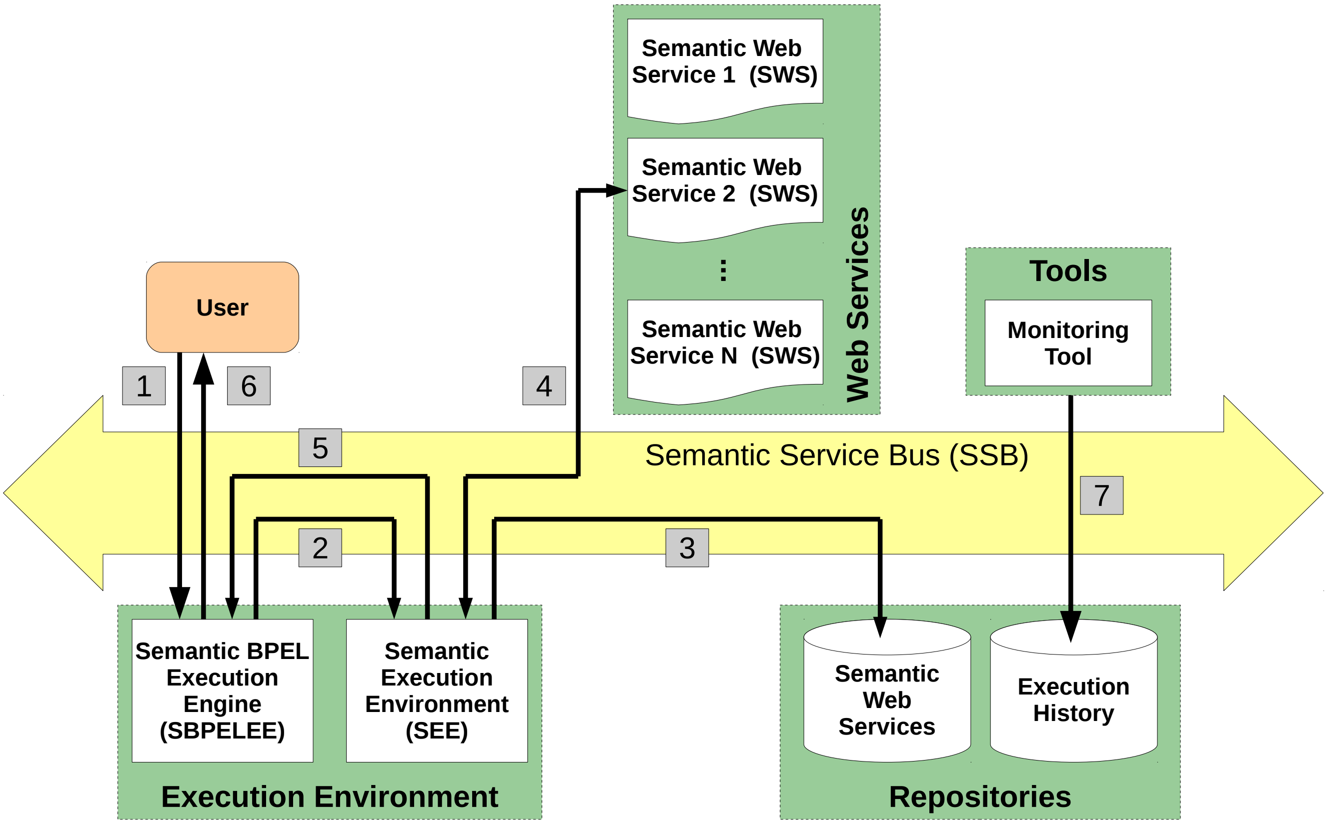

Semantic Web Service 1 (SWS)
Semantic Web Service 2 (SWS)
⋮
Semantic Web Service N (SWS)
Web Services
Tools
Monitoring Tool
User
1
6
4
5
Semantic Service Bus (SSB)
7
2
3
Semantic BPEL Execution Engine (SBPELEE)
Semantic Execution Environment (SEE)
Execution Environment
Semantic Web Services
Execution History
Repositories

Figure 5. The execution scenario in the SUPER methodology (based on the D10.2 SUPER Showcase Presentation http://slideplayer.com/slide/741902/)

### B. *Business Process Semantization in SAP AG*

Another more extended solution related to business process semantization is an approach developed [12] and patented[8] by the SAP AG company.

Semantization of business processes in the SAP AG uses the semantic descriptions for business process artifacts. The integration consists in linking the identified semantic pieces of information described in the form of ontology with the elements of business process models.

The approach also supports semantic modeling by matching elements of a process model to concepts from ontologies and using fitting functions for choosing proper semantic annotations. This is achieved by comparing the given context and text description with instance domain ontology. In the approach, three goals are achieved: support for modeling, exploring relevant services, and searching the process model repository.

Figure 6 shows the main components of the process modeling tool semantic extension. The BPMN data objects are used for describing activities by defining the related objects and state transitions. For such activities, a user can graphically specify their preconditions and postconditions, as well as define the related objects with the specification of the object state changes before and after the execution of the activity.

An ontology in this approach contains the information about objects, states, state transitions, and actions related to the domain. For each object the possible states and state transitions are defined and they form the object life cycle. These kinds of domain ontologies support semantic process modeling by using their concepts in model elements specification, especially by suggesting relevant concepts or instances of objects. For suggesting the relevant components (data objects, activities, associations and states), a combination of different algorithms associated with the text matching was used. The algorithms take advantage of contextual information related to the process model as well as domain knowledge ontology.

Thanks to the domain knowledge, the names of tasks can be suggested based on the object life cycle. The object life cycle can also be used to exclude re-using the task names that have already been modeled.

The system also supports the semantic description of data flow. The object status can be visualized directly in the diagram. The “less than” sign (<) denotes the object status before and the “greater than” (>) denotes the object status after performing the associated tasks.

Such semantization supports consistency checking and extends the capabilities of semantic searching. Compared to the approach of the SUPER project, it supports more flexible and accurate semantic annotations by referring directly to the elements from the defined domain ontology.

[8] See: The patent “Semantic extensions of business process modeling tools” number US 8112257 B2: https://www.google.com/patents/US8112257.

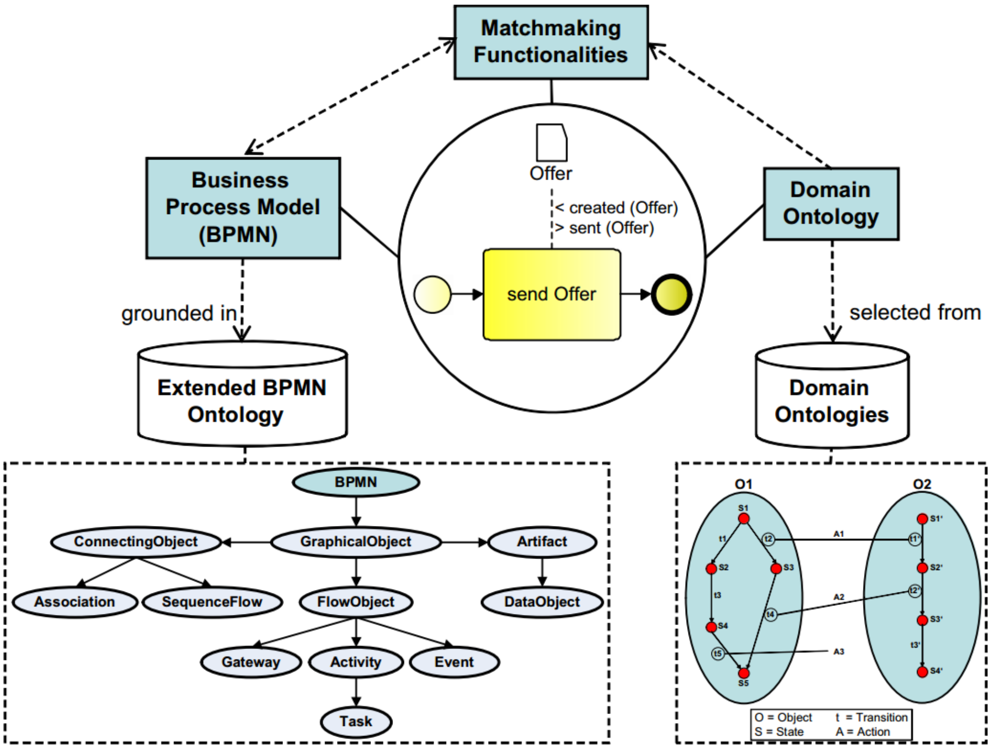


Figure 6. The main components of the semantic extension tools for modeling business processes in the SAP platform [12]

### C. Other possibilities of semantization

As the considered open-source process execution environments use REST interfaces, the possibility of semantization of the REST interface is presented in [13]. The authors compare the different semantic annotation languages for REST interfaces and show how to take advantage of them by creating a website which combines online applications from different sources (in particular internet services) - the so-called mashup. They also proposed a new language SAWADL, based on WADL (Web Application Description Language), for REST Web Services semantization.

In the Signavio Process Editor[9], some basic semantization in the form of a dictionary can be observed. In the dictionary one can define the concepts, assign them to one of the 6 categories (Organizational Unit, Documents, Activities, Events, IT System, Other), add the appropriate descriptions, as well as assign to them additional documents (links to them). Then, these concepts can be used to describe the elements of the BPMN model, e.g. during choosing a name for the particular task in the process (see Figure 7).

Although the tool does not support the formal semantic description in the form of the ontology, it supports multilingual description of the same concepts what allows users to work with the same model in different countries in their own languages.

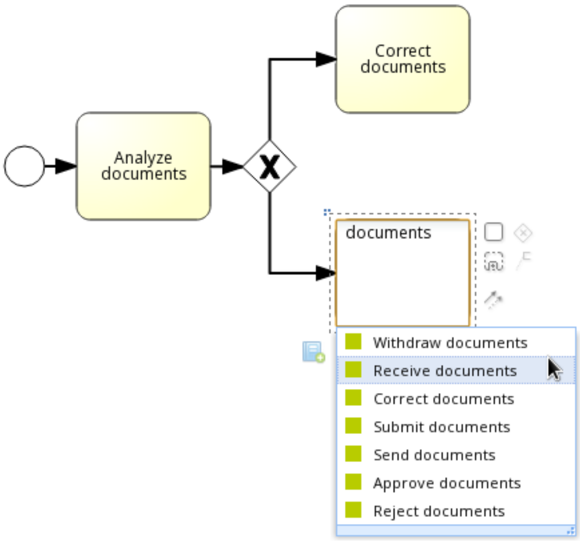


Figure 7. Using dictionary during modeling in the Signavio Process Editor

Other works related to semantic business process modeling can be found in the papers [14], [15], where the processes in the form of Petri nets are connected with the ontology describing the network. Their objective is to standardize the terminology, in particular with regard to the level of abstraction of the labels used in the model. This allows for validating models and detecting the incorrect names of the elements.

[9] See: http://academic.signavio.com.

# IV. Analysis of semantization of selected runtime environments

## A. Prosecco System

The Prosecco (Processes Semantics Collaboration for Companies) project aims at addressing the needs and constraints of Small and Medium Enterprises (SME) by designing methods that will significantly improve BPM systems by simplification of the system design and configuration, targeting the management quality and competitiveness improvement, fostering decisions making and strategic planning in the SME market sector. The Prosecco system involves three technologies for specification of business logic: business processes, rules and ontologies. Additionally, these technologies are supported by external services providing additional functionalities and integrating core of the system with external tools.

The architecture of the system is oriented towards services what significantly improves its portability and high versatility. Such architecture also enables integration with external tools that provide their functionality as a service. As each element of the architecture provides own data management, there is no centralized repository for all models. Therefore, Prosecco repository consists of several repositories for various models:

1) Repository for Business Processes is called *prosecco-business* and is managed by Activiti engine. It stores information concerning existing processes and their instances, variables and other data processed by the engine. Additionally, the repository contains components that can be used for creating new processes.
2) *Prosecco-knowledgebase* is a repository for rules. It is divided into two parts. The first part stores rules processed by Drools rule engine. The second contains rules learned according to decisions made by the users that are traced on the business process level.
3) Ontology is stored in the OWL format. The POJO model, which is suitable for process and rule engines, can be generated from the OWL representation.
4) *Prosecco-profilemanager* repository stores information related to users and ACL (Access Control List).
5) System History is managed by the Cassandra tool and stores information concerning operations performed within the system.

In turn, due to the fact of usage of the ontology, the project assumes that all data types and their instances existing within the system are consistent with the ontology. Because of the separate data repositories, the object types and existing instances have to be continuously synchronized with the ontology. For example for rule engine the POJO (Plain Old Java Object) model is generated according to the ontology, and for external tools dedicated integration interfaces providing type alignment were developed.

The Prosecco system uses Activiti as the process engine. Apart from the system types generated according to ontology, the Activiti uses also other semantization techniques. The possible scenarios of semantization of the open-source process engines are considered in the following section.

## B. Analysis of semantization scenarios of Activiti, jBPM and Camunda runtime environments

In Section II, six components of the open-source Activiti, jBPM and Camunda environments are distinguished. Semantization of each of them allows for disambiguation of data description and controlling their integrity. Moreover, it may extend their functionality and possible use scenarios in the following way:

1) Process Engine – may allow for invoking semantically matched web services, rule-based components or subprocesses.
2) Repository of Processes – may enable semantic search of models in the repository.
3) Web-based editor of Processes – may provide semantic-based recommendation of the elements during the modeling process.
4) Eclipse-based editor of Processes – may suggest names of a process elements, artifacts, etc.
5) Process Management System – may support semantic search in the system e.g. running process instances.
6) Log – logging of system events described semantically.

Semantization of these components can bring advantages. Some of these semantization scenarios are considered in the Prosecco project.

Moreover, Prosecco also extends Activiti with a module supporting additional data types. Activiti provides poor data type system[10]; natively, it supports only five data types (`String`, `Long`, `Enum`, `Date`, `Boolean`) that are often insufficient[11]. Therefore, the developed module provides dedicated form-based user interface that allows for entering the values and provides validation mechanism for such new data types as:

- `Month` and `Season`, `Text`,
- `Double` (floating point number),
- `PIN` (Personal Identification Number),
- `PESEL` (Polish national identification number),
- `Country` (list of countries).

Semantization of this module may include additional descriptions of these types as well as their instances, e.g.:

- `Month` – number of days,
- `Season` – months or days in particular season,
- `Text` – minimal or maximal length of the text,
- `Double` – number of significant digits, unit, displaying format, etc.,
- `PIN` – number of digits, number of attempts to enter correct value, additional security information,
- `PESEL` – interpretation of digits in the number (e.g. birth date, gender),
- `Country` – geographical location, polity, etc.

According to the project assumptions, all the types as well as their semantic descriptions that are used within process must be consistent with the ontology.

[10] See: http://www.activiti.org/userguide/#formProperties.

[11] The Activiti environment provides also two additional advanced data types: `User` and `ProcessDefinition`, which are not listed in the user guide, but available in the environment [4].

## V. Summary and future works

The paper gives an overview of business process semantization approaches in the existing proprietary software. It also presents the selected open-source process execution environments (such as Activiti, jBPM and Camunda) and outlines their general architecture, as well as analyzes the possibilities of process semantization in these open-source environments.

Some of the proposed semantization methods have already been introduced in the Prosecco (Processes Semantics Collaboration for Companies) research project, which takes advantage of the Activiti execution engine. Activiti was chosen because it can be easily extended and provides a good documentation [4]. Activiti is the winner of a 2013 Best of Open Source Software Awards (BOSSIE)[12].

Our future works will focus on development of new semantization techniques that can improve business process management environments not only with simple semantic annotations [16]. This can be used for extending recommendation methods in Activiti [17] or semantization of rules in the Wiki environment integrated with processes [18]. Moreover, it is possible to take advantage of semantization in business process verification [19].

## References

[1] M. Weske, *Business process management: concepts, languages, architectures*. Springer Science & Business Media, 2012.

[2] A. Lindsay, D. Dawns, and K. Lunn, "Business processes – attempts to find a definition," *Information and Software Technology*, vol. 45, no. 15, pp. 1015–1019, Dec 2003, elsevier.

[3] OMG, "Business Process Model and Notation (BPMN): Version 2.0 specification," Object Management Group, Tech. Rep. formal/2011-01-03, January 2011.

[4] T. Rademakers, T. Baeyens, and J. Barrez, *Activiti in Action: Executable Business Processes in BPMN 2.0*, ser. Manning Pubs Co Series. Manning Publications Company, 2012.

[5] M. Salatino, *jBPM Developer Guide*. Packt Publishing Ltd, 2009.

[6] J. Freund and B. Rücker, *Real-life BPMN: Using BPMN 2.0 to analyze, improve, and automate processes in your company*. Camunda, 2012.

[7] D. Roman, U. Keller, H. Lausen, J. de Bruijn, R. Lara, M. Stollberg, A. Polleres, C. Feier, C. Bussler, D. Fensel *et al.*, "Web service modeling ontology." *Applied ontology*, vol. 1, no. 1, pp. 77–106, 2005.

[8] A. Filipowska, M. Hepp, M. Kaczmarek, and I. Markovic, "Organisational ontology framework for semantic business process management," in *Business information systems*. Springer, 2009, pp. 1–12.

[9] I. Weber, J. Hoffmann, J. Mendling, and J. Nitzsche, "Towards a methodology for semantic business process modeling and configuration," in *Service-Oriented Computing - ICSOC 2007 Workshops*, ser. Lecture Notes in Computer Science, E. Di Nitto and M. Ripeanu, Eds. Springer Berlin Heidelberg, 2009, vol. 4907, pp. 176–187. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-93851-4_18

[10] D. Fensel, F. M. Facca, E. Simperl, and I. Toma, "What are sws good for? dip, super, and soa4all use cases," in *Semantic Web Services*. Springer, 2011, pp. 299–324.

[11] D. Fensel, H. Lausen, A. Polleres, J. de Bruijn, M. Stollberg, D. Roman, and J. Domingue, *Enabling semantic web services: the web service modeling ontology*. Springer Science & Business Media, 2006.

[12] M. Born, F. Dörr, and I. Weber, "User-friendly semantic annotation in business process modeling," in *Web Information Systems Engineering – WISE 2007 Workshops*, ser. Lecture Notes in Computer Science, M. Weske, M.-S. Hacid, and C. Godart, Eds. Springer Berlin Heidelberg, 2007, vol. 4832, pp. 260–271. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-77010-7_25

[13] A. Malki and S. M. Benslimane, "Building semantic mashup," in *ICWIT*, 2012, pp. 40–49.

[14] A. Koschmider and E. Blanchard, "User assistance for business process model decomposition," in *In First IEEE International Conference on Research Challenges in Information Science*, 2007, pp. 445–454.

[15] A. Koschmider and A. Oberweis, "Ontology based business process description." in *EMOI-INTEROP*, 2005.

[16] Y. Liao, M. Lezoche, H. Panetto, N. Boudjlida, and E. R. Loures, "Semantic annotation for knowledge explicitation in a product lifecycle management context: A survey," *Computers in Industry*, vol. 71, pp. 24–34, 2015.

[17] S. Bobek, G. J. Nalepa, and O. Grodzki, "Integration of activity modeller with bayesian network based recommender for business processes," in *Proceedings of 10th Workshop on Knowledge Engineering and Software Engineering (KESE10) co-located with 21st European Conference on Artificial Intelligence (ECAI 2014), Prague, Czech Republic, August 19 2014*, ser. CEUR Workshop Proceedings, G. J. Nalepa and J. Baumeister, Eds., vol. 1289, 2014. [Online]. Available: http://ceur-ws.org/Vol-1289/kese10-05_submission_10.pdf

[18] K. Kluza, K. Kutt, and M. Woźniak, "SBVRwiki (tool presentation)," in *Proceedings of 10th Workshop on Knowledge Engineering and Software Engineering (KESE10) co-located with 21st European Conference on Artificial Intelligence (ECAI 2014), Prague, Czech Republic, August 19 2014*, G. J. Nalepa and J. Baumeister, Eds., 2014. [Online]. Available: http://ceur-ws.org/Vol-1289/

[19] I. Weber, J. Hoffmann, and J. Mendling, "Beyond soundness: on the verification of semantic business process models," *Distributed and Parallel Databases*, vol. 27, no. 3, pp. 271–343, 2010.

[12] See: http://www.infoworld.com/slideshow/119652/bossie-awards-2013-the-best-open-source-applications-226975#slide23

DOI: 10.15439/2015F78
ACSIS, Vol. 5

# Validation and Verification of Temporal Knowledge as an Important Aspect of Implementing a Temporal Knowledge Base System Supporting Organizational Creativity

Maria Mach-Król
University of Economics
ul. Bogucicka 3, 40-226 Katowice,
Poland
Email: maria.mach-krol@ue.katowice.pl

Krzysztof Michalik
University of Economics
ul. Bogucicka 3, 40-226 Katowice,
Poland
Email:krzysztof.michalik@ue.katowice.pl

*Abstract*— **The paper is devoted to the problem of temporal knowledge validation and verification during the process of implementing a system supporting organizational creativity. The motivation for implementing a temporal knowledge base system is presented, the implementation methodology is outlined, and the V&V (validation&verification) process is described in detail, using an example of the Logos reasoning tool. The main achievements of the paper are: elaborating a new implementation methodology for a temporal knowledge base system, and elaborating detailed V&V steps**.

**Keywords: organizational creativity, temporal knowledge base system, validation, verification, implementation methodology, Logos tool..**

## I. Introduction

ORGANIZATIONAL creativity is a relatively new concept in the theory of management, which has partially arisen on the ground of knowledge management.

There are many definitions of organizational creativity, but it is commonly perceived as a team, dynamic activity, responding to changing features of organization's environment, a team process – see e.g.[15], [1].

The organizational creativity is therefore to be perceived in the context of organizational dynamics, because it depends on the situational changes and is composed of processes. Therefore while discussing the question of computer support for organizational creativity, the temporal aspects should not be omitted.

Such a way of formulating this problem – underlining its dynamic aspect – justifies a proposal of using an intelligent system with a temporal knowledge base, as a tool supporting creation and development of organizational creativity, which is understood as organizational asset (see e.g. [7], [14]).

By the system with a temporal knowledge base we will understand (slightly modifying the definition given in [8]) an artificial intelligence system, which explicitly performs temporal reasoning. Such a system contains not only fact base, rule base, and inference engine, but also directly addresses the question of time. For an intelligent system to be temporal, it should contain explicit time representations in its knowledge base – formalized by the means of temporal logics – and at least in the representation and reasoning layers. In the paper we use an example of the Logos tool – a reasoning system constructed by the authors within the frame of the research project under the same name. One of the assumptions of the Logos project is the possibility of using it for different scientific researches and experiments, among others using it for building temporal knowledge base research prototype.

The main aim of the paper is to present a new implementation methodology for a temporal knowledge base system supporting organizational creativity, and to present – in detail – the steps performed during validation and verification of the temporal knowledge embedded in the system.

## II. Motivation

While reading many authors' discussions on the essence of organizational creativity, one sees that this is primarily team activity. The effect of this activity may be referred to as "creative knowledge", which itself generates new ideas, concepts, and solutions. To do so, the creative knowledge must be first codified, and next disseminated. This justifies the use of a knowledge base system. But the creative knowledge changes in time, due to several reasons.

First, organizational creativity is a process, therefore its effects are subject to change. Moreover, the process encompasses solving problems that also change, because the organization's environment changes [8] p. 13-15, [3] p. 150 and next, 176 and next.

Second, each knowledge – including the creative one – changes simply with the passage of time, with the flow of new information about objects [2].

Third, organizational creativity is linked with dynamics, which can be seen e.g. in the assets approach to this creativity

This paper has been supported by a grant: „Methodology for Computer Supported Organizational Creativity" from National Science Centre in Poland, 2013/09B/HS4/00473.

978-83-60810-66-8/$25.00©2015, IEEE

or in the requirement of adapting creative knowledge to situational context.

All the above leads to conclusion that a knowledge base system is not enough to support organizational creativity, because classical knowledge bases do not support time. Therefore in this paper we propose the use of a temporal knowledge base system, as defined earlier. Such system is able to perform the tasks arising from the characteristics of organizational creativity and its artifacts.

An important element of the implementation methodology of the proposed system is the process of validation and verification of temporal knowledge embedded in the system. The system is a rule-based one (strictly speaking: a temporal rule-based one), thus in order to run properly, its knowledge base must be correct. And a "correct" temporal knowledge base means that it has been verified and validated to ensure that there are no anomalies such as: redundant rules, subsuming rules, contradictory rules, unused attributes, unused values, recursive rules (inference loop, circularity),

Thus, the procedures of validation and verification (V&V) are an important fragment of implementation methodology for the proposed system. In more detailed manner we describe this problem in section 5.

## III. Related Work

In the literature, there are many methodologies concerning expert systems, among others [9], p. 136:

- blackboard architecture,
- KADS and CommonKADS,
- HyM for hybrid systems,
- Protégé,
- CAKE.

It must be however noted, that the above mentioned methodologies have been created for expert system, while the proposed temporal knowledge base system is not a typical ES. In the context of its architecture, it is worthy of considering the blackboard architecture, which is by some authors understood as a knowledge engineering methodology [6]. It enables to explicitly represent knowledge and its structure in a rule-based system (and the proposed temporal knowledge base system is a rule-based one). It may be acknowledged that a postulated division of system's knowledge base into several sub-bases means implementing the blackboard architecture and achieving its assumptions.

The second interesting methodology is CAKE (Computer Aided Knowledge Engineering), elaborated by Michalik. The detailed description of CAKE may be found e.g. in [11], [9]. Its advantages are similar to those of the blackboard methodology:

- use of the blackboard systems methodology,
- easy management of heterogeneous knowledge sources,
- support of groupworking,
- automatic control of formalized creative knowledge code,
- knowledge base editor,
- a package of wizards facilitating the coding process of the acquired knowledge.

It has to be pointed out, however, that the knowledge coding formalism, embedded in the CAKE system, has no temporal references. A sample diagram of knowledge base anomalies can be found e.g. in [16] and more detailed discussion on verification and validation in [12]. Some introductory concepts concerning V&V in Logos system being subject of our presentation can be found in [10]. Very interesting remarks on V&V in the context of knowledge engineering in the CommonKADS Methodology can be found in [13]. Authors differentiate between internal validation for both internal and external meaning, e.g. saying that some people use the term verification for internal validation and apply validation concept against user requirements ("is it the model right?").

## IV. System Implementation Methodology

It is also important that the aforementioned methodologies relate mainly to building expert systems, while the methodology needed for a temporal knowledge base system has to take into account also the processes of implementing the system in a creative organization. Therefore it is not possible to directly use any of the aforementioned methodologies, and a new one has to be developed, suited to the task of supporting organizational creativity by a temporal knowledge base system. Two very important questions thus arise.

First, the proposed system is an intelligent one, containing at least one temporal knowledge base, therefore the implementation methodology has to make use of (but not copying directly) existing methodologies for implementing such systems, as e.g. expert ones.

Second, the main aim of the system is to support organizational creativity, therefore the most important system elements are user interface and knowledge base. The first enables both adding creative knowledge to the system, and querying this kind of knowledge, the second is needed in representation and reasoning layers. The proposed methodology should accommodate also these elements.

The implementation methodology for a temporal knowledge base system has to be conformable to temporal knowledge base system's lifecycle. We propose the following lifecycle for the system (adapted from [5]):

1. Problem identification, and definition of users' needs;
2. System's formal specification, encompassing dialogues with users;
3. Definition of knowledge sub-bases', and general knowledge base structure and scope, choice of knowledge representation technique(s), creation of reasoning algorithm;
4. Creative knowledge acquisition;
5. Prototype creation and verification;
6. System coding and testing;

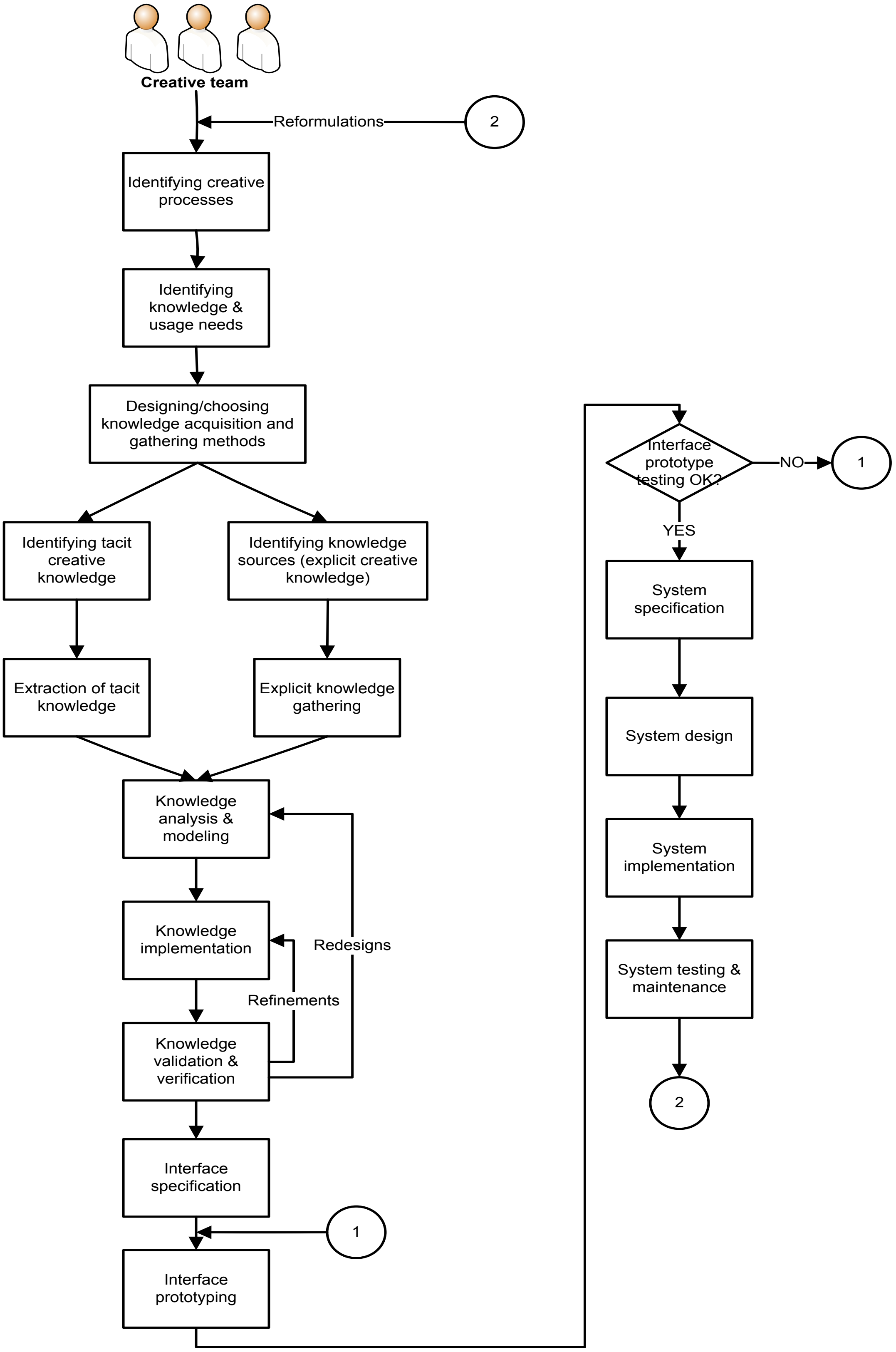

Creative team
Reformulations
2
Identifying creative processes
Identifying knowledge & usage needs
Designing/choosing knowledge acquisition and gathering methods
Identifying tacit creative knowledge
Identifying knowledge sources (explicit creative knowledge)
Extraction of tacit knowledge
Explicit knowledge gathering
Knowledge analysis & modeling
Knowledge implementation
Knowledge validation & verification
Redesigns
Refinements
Interface specification
1
Interface prototyping
Interface prototype testing OK?
NO
1
YES
System specification
System design
System implementation
System testing & maintenance
2

Fig. 1. Schema of temporal knowledge base system implementation methodology

Source: Own elaboration.

7. System maintaining and development – principally the creative knowledge bases and user interface.

The implementation methodology for the temporal knowledge base system covers points 1,2, 4-6, and 7 of the proposed system's lifecycle.

The methodology should be focused primarily on the creative knowledge, and on user interface. Therefore its main elements are creative knowledge engineering, and system engineering, with emphasis put on interface design and prototyping. The general structure of the proposed methodology is presented in fig. 1.

The proposed methodology has been inspired by other knowledge engineering methodologies for the process of knowledge management – particularly by the work [18] – and by classical, fundamental methodologies for implementing expert systems: [17], p. 135-139, and [4], p. 139. Obviously, it was not possible to directly merge the existing models, the proposal concerning knowledge engineering had to be remodeled in the context of organizational creativity process, while methodologies for implementing expert systems had to be adapted to the temporal knowledge base system, and its main task.

The methodology for implementing a temporal knowledge base system starts with the group of activities concerning capturing, and modeling of creative knowledge. At this stage it is essential to discover creative processes running within the team of employees involved in organizational creativity. This will enable to identify needs concerning the creative knowledge, and its usage by an organization (or team). Having this information, the next step of the methodology is to choose and/or design tacit creative knowledge acquisition methods, as well as to acquire explicit creative knowledge. This is so because we assume that the creative knowledge, as any other kind of knowledge, may be divided into tacit and explicit one. Next, it is necessary to identify (with the aid of previously gathered information) tacit knowledge, and sources of explicit knowledge, and to acquire both types of creative knowledge. Only then it is possible to model and analyze the creative knowledge, which is to be incorporated in the temporal knowledge base system.

During each stage of the proposed methodology, especially during the creative knowledge engineering stage, it is indispensable to closely cooperate with system users, that is the employees involved in the process of organizational creativity. Without them it is impossible to identify, and to acquire tacit knowledge. Moreover, the system will be useful only if people want to use it.

Activities concerning creative knowledge modeling, implementation, and verification are absolutely crucial, therefore in the proposed methodology there is a possibility to return to previous stages, in order to refine knowledge representation and implementation, or even to completely change the design of the knowledge model.

It also has to be explained why activities concerning system's specification, design, and implementation are placed at the end of the methodology, which differs from classical implementation methodologies for intelligent systems. As it has been already said, the main task for the temporal knowledge base system is to support organizational creativity, so its most essential elements are temporal creative knowledge base(s) and GUI. Thus the methodology is focused on these elements. Activities concerning system's engineering are also important, but are of ancillary nature regarding temporal KB, creative knowledge management, and GUI design.

As it has been already said, an important element (step) in the proposed methodology concerns temporal knowledge validation and verification, to ensure that the knowledge base is correct. The V&V step is preceded by knowledge analysis&modeling, and knowledge implementation. These three steps may be followed several times, continuously refining the knowledge base. Due to the importance of the V&V procedure, it is described in detail in the next section.

## V. Validation and verification of temporal knowledge during the process of implementing a temporal knowledge base system – the Logos example

As mentioned, we build the Logos reasoning system, being software platform for our experiments with temporal knowledge bases as well as with temporal reasoning. Additionally, while developing this system, we take into account the very important factor of V&V. At present main procedures discovering some anomalies in knowledge bases are ready. What have to be done – according to our thesis concerning V&V in temporal knowledge bases – is creating algorithms to discover some specific temporal anomalies. The thesis is that in temporal knowledge bases some new, very specific anomalies and errors, may theoretically appear. On the other hand, most of V&V methods already built-in in Logos for conventional (not temporal) knowledge bases are also useful and even necessary. The reason is that temporal knowledge bases may include the same kind of anomalies as the conventional ones. Most of the temporal anomalies we plan to detect as a first step while building Logos relate to incorrect time dependencies as declared in knowledge base. As we mentioned in section 3, the correct temporal knowledge base means that it has been verified and validated to ensure that there are no anomalies such as the following [11], [9]:

**Redundancy**

Two rules we regard as redundant, if for two rules:
$R_i \leftarrow W_{i1} \wedge .. \wedge W_{in}$ and
$R_j \leftarrow W_{j1} \wedge .. \wedge W_{jn}$, where $i \neq j$,
holds: $\{ W_{i1}..W_{in}\} = \{ W_{j1}..W_{jn}\}$

**Subsuming rules**

If for two different rules:
$R_i \leftarrow W_{i1} \wedge .. \wedge W_{im}$ and
$R_j \leftarrow W_{i1} \wedge .. \wedge W_{in,}$ where $i \neq j$,
holds $\{ W_{i1},..,W_{im} \} \subseteq \{ W_{i1},..,W_{in} \}$, then we say, that rule $R_j$ subsumes $R_j$.

**Contradictory rules**

Two rules we regard as contradictory if
$R_i \leftarrow W_1 \wedge .. \wedge W_n$ and

$\neg R_j \leftarrow W_1 \wedge .. \wedge W_n$, where $i \neq j$.

**Inconsistent rules**

Two rules we regard as inconsistent if
$R_i \leftarrow W_1 \wedge .. \wedge W_n$ and
$R_j \leftarrow W_1 \wedge .. \wedge W_n$, where $i \neq j$ and $R_i \neq R_{j,.}$
In our opinion, in such case system should give warning for knowledge engineer about possibility of inconsistency as understood in logic (here we use terminology of our system). Such situation, may in fact potentially lead to a contradiction or can be result of any other mistake of knowledge engineers they should be aware.

**Incompleteness**

We assumed that temporal knowledge base is complete when contains all possible combinations of attributes and their allowable values in rules' antecedents and consequents. It should be noticed that in practice not all combinations are required, so this kind of verification is only warning for knowledge engineer that some rules could be missing.

**Missing rules**

The anomaly called here as missing rules can be treated as a special case of incompleteness. While creating V&V module of Logos system we consider missing rules as the case when some of decision-making attributes are not present in any of the rule antecedents. This situation can appear as side effect of rapid prototyping and incremental method of knowledge base development.

**Unused attributes and values**

Our system in order to be able to detect some anomalies and errors requires explicit declaration of attributes and values being used in the knowledge base. When given attribute or value is never used in any of rules then Logos gives warning addressed to knowledge engineer because it may be information about serious anomaly in knowledge base. On the other hand, similarly as in the case of missing rules it can be side effect of using methodology of rapid prototyping and incremental development of knowledge base.

**Recursive rules (inference loop, circularity)**

Recursion in most of rule-based systems is very important anomaly in knowledge base with serious consequences and sometimes very difficult to detect by knowledge engineer without software support, e.g. as implemented in our Logos system. Recursion in this context may take a variety of patterns. In the simplest case can be like this:

$$K_i \leftarrow W_1 \text{ and } K_i = W_{1.}$$

This type of recursion (loop) is very easy to detect for knowledge engineer, even without special algorithms. The sign '=' does not mean simple equality but may have more complex semantics of ability of two expressions to match. Other variants of the same direct recursive call of the conditions to the conclusions of the rule may take one of the following schemes:

$K_i \leftarrow W_1 \wedge .. \wedge W_n$ and $K_i = W_1$

$K_i \leftarrow W_1 \wedge .. \wedge W_n$ and $K_i = W_n$

$K_i \leftarrow W_1 \wedge ..W_j.. \wedge W_n$ and $K_i = W_j$

While recursion in rules of logic programs is very useful and correct situation (provided their correct semantics), the recursion in expert systems is generally treated (as mentioned) as serious knowledge base anomaly.

We rejected using recursions in Logos for many reason, the main is that logic programming is first of all programming formalism (e.g. Prolog) and not knowledge representation language for temporal knowledge bases.

Much more difficult situation to detect by the knowledge engineer is that of indirect recursion. Sometimes, in large knowledge bases with several levels of inference, it can be practically not possible to detect in reasonable time. Then software support as e.g. that we implemented in Logos is absolutely necessary. In such case of indirect recursive call it does not occur at the same level of a given rule to its conclusion, within a single rule. This can be illustrated by the following example:

$R_1$: $K_1 \leftarrow W_{11} \wedge .. W_{1j} .. \wedge W_{1x}$
$R_i$: $K_i \leftarrow W_{i1} \wedge .. W_{il} .. \wedge W_{iy}$
$R_n$: $K_n \leftarrow W_{n1} \wedge .. W_{nm} .. \wedge W_{nz}$
where: $W_{1j} = K_i$ i $W_{il} = K_{n.}$ i $W_{nm} = K_1$ and '=' means matching/unification

In this case, the recursive call is related to a lower level in the hierarchy of rules, making its location is difficult to detect for a knowledge engineer. NB: Incidentally, defined earlier contradiction of rules – as mentioned - can also be caused by indirect inference.
Direct contradiction:
$p \leftarrow q$
$\neg p \leftarrow q.$
Indirect contradiction appearing during inference process:
$\neg p \leftarrow r$
$p \leftarrow q$
$q \leftarrow r.$

## VI. Concluding Remarks

Our main goal at this stage of research was design and development of software platform to make experiments with temporal knowledge bases. In this paper we focused on the problem on V&V, which is always present while building knowledge bases, but sometimes underestimated. The consequences of badly evaluated knowledge bases in this respect can be very serious, and for example some anomalies can't be detected for long time giving bad decisions. The more so, as we showed some of the anomalies, as for example recursions (inference loops) can be sometimes very difficult to detect by knowledge engineer. So our objective since the beginning of the project Sphinx was to build computer aided knowledge engineering system automatically supporting knowledge engineer especially helpful in such difficult to analyze cases. The next step in our researches concerning temporal knowledge bases will be identification of the specialized anomalies typical only for temporal systems. We suspect that beside the simple related to time errors we may discover special kind of temporal anomalies. Even some of already identified and described in our paper some anomalies can take the new character in relation to time. We have also additional hypothesis that temporal knowledge bases have kind of anomalies which are completely specific and different from that described. Verification of theses hypothesis we take as our next goal of our researches.

## References

[1] Andriopoulos, C. and Dawson, P., 2014. *Managing Change, Creativity and Innovation.* Second Edition. Los Angeles/London/New Delhi/Singapore/Washington DC: SAGE Publications.

[2] Benthem van, J., 1995. *Temporal Logic.* In: D. M. Gabbay, C. J. Hogger and J. A. Robinson, Eds. *Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 4: Epistemic and Temporal Reasoning.* Oxford: Clarendon Press, pp. 241-350.

[3] Czaja, S., 2011. *Czas w ekonomii. Sposoby interpretacji czasu w teorii ekonomii i w praktyce gospodarczej.* Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego.

[4] Hayes-Roth, F., Waterman, D. and Lenat, D. Eds., 1983. *Building Expert Systems.* Reading, Mass.: Addison-Wesley Publishing Company.

[5] Infernetica, 2012. *Systemy ekspertowe dla biznesu.* [Online] Available at: http://infernetica.com/systemy-ekspertowe/ [Accessed: 07 04 2014].

[6] Kendal, S. and Creen, M., 2007. *An Introduction to Knowledge Engineering.* London: Springer.

[7] Krupski, R., Ed., 2011. *Rozwój szkoły zasobowej zarządzania strategicznego.* Wałbrzych: Wałbrz.Wyż.Szk.Zarz. i Przedsięb.

[8] Mach, M. A., 2007. *Temporalna analiza otoczenia przedsiębiorstwa. Techniki i narzędzia inteligentne.* Wrocław: Wydawnictwo AE.

[9] Michalik, K., 2014. *Systemy ekspertowe we wspomaganiu procesów zarządzania wiedzą w organizacji.* Katowice: Wydawnictwo Uniwersytetu Ekonomicznego.

[10] Michalik, K., 2015. *Validation and .Verification of Knowledge Bases in the Context of Knowledge Management. Logos Reasoning System Case Study,* [w:] Technologie wiedzy w zarządzaniu publicznym'13, red. J. Gołuchowski, A. Frączkiewicz-Wronka, Wydawnictwo UE, Katowice (in print)

[11] Michalik, K., Kwiatkowska, M. and Kielan, K., 2013. *Application of Knowledge-Engineering Methods in Medical Knowledge Management.* In: R. Seising and M. E. Tabacchi, Eds. *Fuzziness and Medicine: Philosophical Reflections and Application Systems in Health Care.* Berlin Heidelberg: Springer, pp. 205-214.

[12] Owoc M., Ochmańska M., Gładysz T., 1999. *On Principles of Knowledge Validation,* [in:] *Validation and Verification of Knowledge Based Systems: Theory, Tools and Practice,* eds. A. Vermessan, F. Coenen, Kluwer Academic Publishers, Boston.

[13] Schreiber et al., 2000. *Knowledge Engineering and Management, The CommonKADS Methodology,* The MIT Press, Cambridge MA,.

[14] Sirmon, D. G., Hitt, M. A., Ireland, R. D. and Gilbert, B. A., 2011. *Resource Orchestration to Create Competitive Advantage: Breadth, Depth, and Life Cycle Effects.* Journal of Management, September, Vol. 37 (No. 5), pp. 1390-1412.

[15] Unsworth, K. L., 2001. *Unpacking Creativity.* Academy of Management Review, Vol. 26(No. 2), pp. 286-297.

[16] Vermessan A.I., 1998. *Foundation and Application of Expert System Verification and Validation,* [in:] *The Handbook of Applied Expert Systems,* ed. J. Liebowitz, CRC Press, New York 1998.

[17] Waterman, D., 1986. *A Guide to Expert Systems.* Reading, Mass.: Addison-Wesley Publishing Company.

[18] Yazdanpanah, A. and Sadri, R., 2010. *Proposed Model for Implementation Expert System for the Planning of Strategic Construction Projects as a Tool for Knowledge Management.* Istanbul, IPMA.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1321–1326

DOI: 10.15439/2015F323
ACSIS, Vol. 5

# Unifying Business Concepts for SMEs with Prosecco Ontology

Grzegorz J. Nalepa*, Mateusz Ślażyński†, Krzysztof Kutt‡, Edyta Kucharska§, Adam Łuszpaj¶
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059, Krakow, Poland
*gjn@agh.edu.pl, †mslaz@agh.edu.pl, ‡kkutt@agh.edu.pl, §edyta@agh.edu.pl, ¶adam@softhis.com

*Abstract*—**Knowledge management in business information systems often requires a unified dictionary of business concepts, that allows for a transparent integration of such systems. Thanks to it sharing the conceptualization between users becomes possible, and better decision support facilities can be provided. The Prosecco project is a research and development project aims to address the needs and constraints of small and medium enterprises by designing methods that will significantly improve BPM systems. In this paper we focus on the development of ontology-based mechanisms allowing for creating taxonomies of business logic concepts unifying system objects. Building a taxonomy of business concepts shared by number of SMEs targeted in the project and then turning it into a formalized ontology integrating the software components is a major challenge. The paper demonstrates how this ontology is used to unify vocabulary of business processes and rules. The original contribution of this research discussed in the paper is the design and implementation of the ontology, and the demonstration of its practical use in the system.**

## I. Introduction

Providing a unified dictionary of business concepts is often a critical aspect of knowledge management in business information systems. Such a dictionary allows for a transparent integration of these systems. Moreover, it allows for sharing the conceptualization between different users providing better decision support facilities. For over a decade, it has been a common approach to support this task with Semantic Web technologies, including ontologies. Specifically formalized ontologies e.g. in OWL are of great practical importance. They become not only a tool for capturing the conceptual description of a business system, but also provide a technical backbone for software modules it is composed of. From the technical point of view, building ontologies is a knowledge engineering task that is currently mostly well supported. However, number of challenges remain, including practical integration of dedicated ontologies in a business information systems, often including business process and business rules (BR) management modules. Possible unification of these through an ontology is a non trivial task of great importance.

The *Prosecco* (Processes Semantics Collaboration for Companies) project [1] is a 32 month research and development project funded by NCBR (2012-2015). Provisioning of Business Process Management (BPM) systems is an important activity of main IT vendors. However, such systems are dedicated mainly for large companies, organizations and agencies. The motivation of the project is to address the needs and constraints of small and medium enterprises (SME) by designing methods that will significantly improve Business Process Management (BPM) systems. The main goal is to provide technologies that improve and simplify the design and configuration of BPM systems integrated with Business Rules Systems, targeting the management quality and competitiveness improvement. Moreover the project aims at fostering decision making and strategic planning in the SME market sector (mainly in the selected services sector). Specific objectives of the project include: a) development of business process modelling methods taking into account semantic dependencies between business process models and rule models, b) providing recommendation methods for analysis of semantically described business process models, and even more importantly, c) the development of ontology-based mechanisms allowing for creating taxonomies of business logic concepts unifying system objects.

In this paper we focus on the above mentioned objective c). In fact, building a taxonomy of business concepts shared by number of SMEs targeted in the project and then turning it into a formalized ontology integrating the software components is a major challenge. The paper demonstrates how this ontology is used to unify vocabulary of business processes and rules. The original contribution of this research discussed here is the design and implementation of the ontology. Furthermore, we demonstrate its practical use in the Prosecco BPM system.

The rest of the paper is composed as follows: In Section II we briefly discuss the architecture of the Prosecco system. Then in Section III we describe the design process of the ontology. Section IV demonstrates the capturing the semantics of business logic components of the system. Related work is included in Section V. The paper ends with a brief evaluation and summary in Section VI.

## II. Prosecco System Architecture

From the technical point of view, some of the main objectives of the Prosecco system that meets the project goals include the development of the:

1) integrated business logic model composed of business processes and rules,
2) runtime environment suitable for execution of the model,
3) recommendation modules for the design and use of business artifacts,

[1] See http://prosecco.agh.edu.pl for the project website.

978-83-60810-66-8/$25.00©2015, IEEE

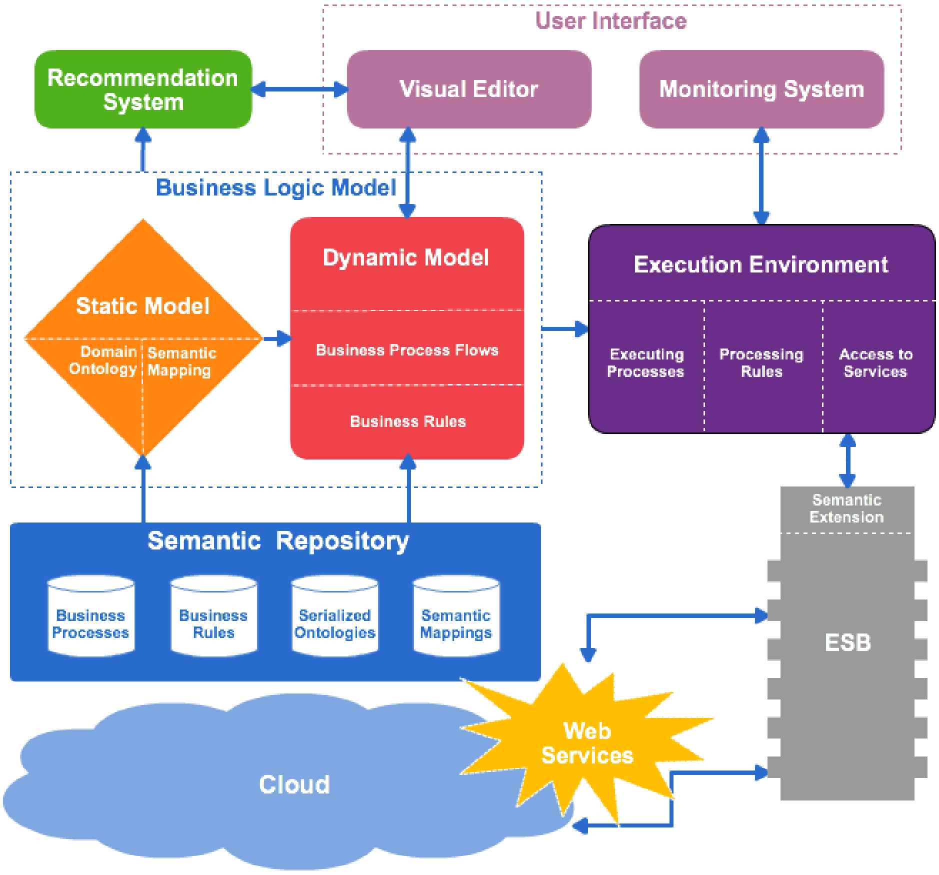


Figure 1. Outline of the Prosecco system architecture.

4) repository of business objects,
5) enterprise service bus (ESB) integrating the system components in a cloud environment, and
6) system ontology based on the taxonomy of shared business concepts.

The outline of the system can be observed in the Fig. 1.

The end users of the Prosecco system are SMEs, in fact selected employees as well as management of companies. The system aims at supporting carrying out of the main business processes of a company. Moreover, these processes are accompanied by business rules, capturing the details of the business logic, including lower level processing, as well as high-level constraints. Within the system these are modeled with the help of BPMN (Bussiness Process Model and Notation), and appropriate BR models identified with SBVR (Semantics of Business Vocabulary and Business Rules) and implemented with the help of the Drools BRMS as well as the HeaRT rule engine.

The process models and rule models are based on concepts captured during initial structured interviews with SMEs. The resulted taxonomy of concepts was used to design and implement the Prosecco ontology that works as the main unifying backbone of the system. It contains all the business terms needed for expressing and capturing the artifacts in business process models and rules. Moreover, it allows to monitor the execution of these models on a semantic level. The execution environment uses the ontology, so users can actually trace how these concepts are used in the executed processes and rules.

From the point of view on the integrators of the system, the ontology supports recommendation mechanism that allows for adaptation of business process and rule models to the needs of specific companies. Furthermore, the repository of business objects uses semantic annotation based on this ontology. It makes it possible for an easy retrieval of needed objects based on semantic queries. In the next section the design of the Prosecco ontology is described.

## III. Design of the Prosecco Ontology

### A. Requirements for the Ontology

The ontology that will be a part of a management system has to identify key concepts and relations describing static aspect of opertions of considered SMEs. Furthermore, it must be integrated with business processes and rules that describe the dynamic aspect of SME management.

Based on Prosecco system architecture described above, some assumptions for the Prosecco Ontology were established:

- Ontology should be modularized: each module should describe different domain (different topic).
- Ontology is designed to be used as dictionary of concept describing elements of business processes and rules.

- Ontology should be defined in a simple description logic language (OWL Lite$_A$, OWL 2 QL, or eventually OWL 2 RL).
- Each concept and role must be documented with short description (what exactly the term means) and possible connections with other ontology elements.
- Ontology is designed for Polish companies so all concepts should be written in Polish language and adjusted to Polish law.
- There should be a possibility to extend ontology by Prosecco system users.

### B. How the Ontology was built

There are some methods and tools that support ontology engineering, but there are no standard approaches how to develop ontologies in general [1]. Brief survey [2] indicates that:

- Existing methods are relatively old.
- The methods can be grouped into categories: incremental and iterative, or more comprehensive ones.
- Most methodologies consists of same main steps: assessment, deployment, testing and refinement.
- Most studies suffer from lack of information about tools.
- Only few recent studies suggests decrease in research activity in this field.

These conclusions encouraged Prosecco analysts to review the most significant and best-established methodologies for creating and managing ontologies [1], [2]. Three main approaches were considered: TOVE (*Toronto Virtual Enterprise*) [3], Enterprise Model Approach [4] and METHONTOLOGY [5]. Analysis of these methodologies resulted in specifying number of steps that are required to develop proper ontology. These steps designate the sequence of work in Prosecco project:

1) *motivating scenarios*: exemplary problems that ontology should resolve,
2) *competency questions*: which questions the ontology should answer,
3) *knowledge acquisition*: interviews with experts, domain texts analysis,
4) *conceptualisation*: extraction of concepts, objects and relations between them,
5) *integration*: consideration of existing ontology reuse,
6) *implementation*: expressing ontology in terms of a formal system,
7) *evaluation*: validation and verification, check for completeness, redundancy and contradictions,
8) *documentation*: comments and documents describing the concepts and relations in ontology.

In Prosecco project each step was performed incrementally in few iterations.

Besides comprehensive methodologies mentioned above, there are projects that provide "good practices" for selected steps in ontology engineering process, e.g. Ontolingua, CommonKADS, KACTUS, PLINIUS, ONIONS, Mikrokosmos, MENELASPHYSSYS, SENSUS (for overview see [1]).

Examined methodologies and "good practices" postulate separation of informal and formal part of ontology development. Some of them suggest introducing a middle representation, that will be a connector between unstructured text and a set of formal axioms. This layer can use structured language or simple graphical language (see IDEF5 methodology [6]). In Prosecco project three "formalization levels" were used:

1) *Unstructured notes*, gathered during interviews with experts (SME workers).
2) *Taxonomy in structured language*, prepared during conceptualisation step. Written down using Prolog language and visualised using custom scripts (see Fig. 2).
3) *Formal ontology*, written in OWL 2 language.

There are three main approaches to defining taxonomies [7]:

- *top-down*: the most general concepts are defined at the beginning and then more specific ones are gradually determined,
- *bottom-up*: low-level concepts and relations are defined at first and then they are generalized,
- *middle-out*: conceptualisation begins with identifying the most relevant concepts; more general and specified ones are determined as needed. It was selected as the best suitable approach for the Prosecco project.

### C. Description of the Ontology

Prosecco Ontology was partitioned into several parts. Each represents a different area in SME management:

1) Project artifacts (PL: *Artefakty*) – components connected with planning and implementation of the project.
2) Organizations (PL: *Organizacje*) – types of companies and their main properties (e.g. e-mail, VAT number).
3) Organization structure (PL: *StrukturyOrganizacji*) – elements that describe company structure, e.g. customer care department.
4) Documents (PL: *Dokumenty*) – concepts and relations associated with various kinds of documents.
5) People (PL: *Osoby*) – depiction of people: key properties (e.g. name, surname) and occupation.
6) Methodologies (PL: *Techniki*) – things connected with methods and tools used in company, e.g. code repository.
7) Resources (PL: *Zasoby*) – grouping into human resources, tangible and intangible resources.
8) Events (PL: *Zdarzenia*) – event types and their properties.
9) Prosecco (PL: *Prosecco*) – main module, the parent of other modules that integrates them and adds additional values common to all of them, e.g. uid or name.

Each part consists of at least one concept, object or dataproperty from another module. These nine parts constitute modules of the Prosecco Ontology that consists of: 86 classes, 70 object properties and 64 data properties. All of them are described by 1042 axioms. Classes are arranged in a hierarchy using `rdfs:subclassof` properties. Besides this simple generalization/specialization relations, each class can be inferred from object and data properties that this class has, e.g.

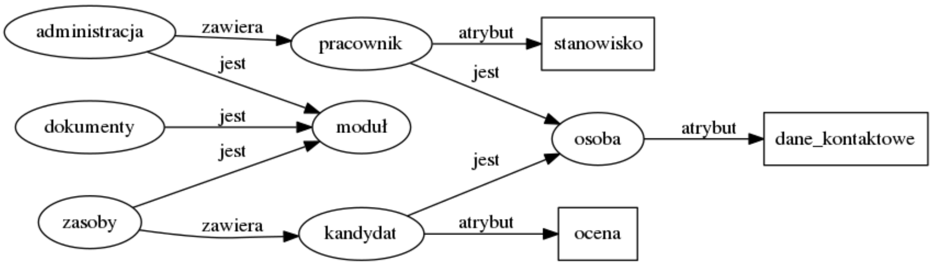
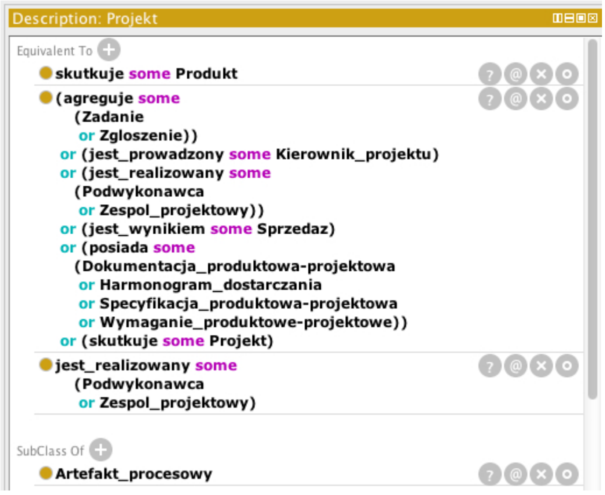


Figure 2. Visualisation of middle layer taxonomy.

`Project` is something that aggregates some `Tasks` or something that is managed by the `Project leader`. Axioms allows inference only in one way: if something aggregates `Tasks`, it must be a `Project`. Fact that something is a `Project` does not infers conclusion that it must aggregates `Tasks`. Example (in Polish) is presented on Fig. 3. For axioms that connect classes with object and data properties, cardinality was also defined as it is presented in Table I.



Description: Projekt

Equivalent To

```
skutkuje some Produkt
(agreguje some
    (Zadanie
     or Zgloszenie))
  or (jest_prowadzony some Kierownik_projektu)
  or (jest_realizowany some
    (Podwykonawca
     or Zespol_projektowy))
  or (jest_wynikiem some Sprzedaz)
  or (posiada some
    (Dokumentacja_produktowa-projektowa
     or Harmonogram_dostarczania
     or Specyfikacja_produktowa-projektowa
     or Wymaganie_produktowe-projektowe))
  or (skutkuje some Projekt)
jest_realizowany some
    (Podwykonawca
     or Zespol_projektowy)
```

SubClass Of

```
Artefakt_procesowy
```

Figure 3. Exemplary concept definition in Prosecco Ontology.

| Cardinality | Protégé code | Example |
|---|---|---|
| $0, 1$ | `max 1` | `home_address max 1 string` |
| $1$ | `exactly 1` | `created exactly 1 dateTime` |
| $[0, \infty)$ | `only` | `description only string` |
| $[1, \infty)$ | `some` | *Not used yet. Defined for future use* |

Table I
CARDINALITY DEFINED IN AXIOMS THAT CONNECTS CLASSES WITH OBJECT AND DATA PROPERTIES.

### D. Tools and resources used

Different tools were used in different steps during the development process of the Prosecco Ontology.

*Taxonomy in structured language:*

- *SWI-Prolog*[2] for interpreting concepts and rules that were recorded with prepared templates for concepts `c(?Id, ?Name)`, relations `r(?Subject, ?Predicate, ?Object)` and attributes `a(?Class, ?AttributeName, ?ListOfPossibleValues)`
- *Custom scripts* for validation and visualisation using *Graphviz* tool[3] (see Fig. 2).

*Formal ontology:*

- *Protégé*[4] editor was used for preparing Ontology in OWL 2 language. It was selected because of its documentation, users support, official handbooks (e.g. [8]) and huge amount of design patterns [9].
- *Git repository* for collaborative work with following iterations of Ontology. *OntoCVS*[5] plugin was used for better tracking of Ontology changes.
- *Ontology modularization tool.* Ontology was implemented as a coherent model. In last step it was divided into modules. Different ontology modularization techniques were analysed. One of the most important is the Atomic Decomposition method [10]. It is based on atoms that consist of a set of axioms that occur together. Correlation between axioms inside atom is so strong, that this axioms have to be together. Single module is a single atom or a set of atoms. Using atomic decomposition alorithm [11] results in a set of modules. For Prosecco Ontology this algotihm generates 62 modules, where one of them consists of 574 axioms (of overall number 1042 axioms). These modules were not functional and decision about manual modularization was made.

The whole development process of the ontology lasted 12 months, including preliminary analytic meetings with SMEs. The development team included 4 business analytics and 4 developers; a 2 person evaluation team was also included.

## IV. Capturing the Semantics of Business Logic

In order to integrate the complete ontology model into the Prosecco architecture some additional works had to be performed. Different formal models, designed to capture dynamics of business processes, were created by separate teams, and they had to be matched with the ontology's concepts. This refinement process can be divided into three separate steps.

Firstly, both BPMN and rule models had very uneven granularity levels: abstract concepts were often mixed with detailed and concrete names of particular tools and documents. Due

[2] See: http://www.swi-prolog.org/.

[3] See: http://www.graphviz.org/.

[4] See: http://protege.stanford.edu/.

[5] See: https://code.google.com/p/ontovcs/.

to these inconsistencies integration of both dynamic models was nearly impossible. During the first step, all the concepts in business processes and rules were generalized or rejected where necessary to represent the same level of abstraction as ontology. The resulting models were smaller and more general than previous versions:

- Exemplary rule that was rejected due to the limitation only to selected SMEs: `It is necessary that each Employee has the available vacation days.`
- Exemplary rule before: `It is possible that a` *`programmer`* `can change a task status [...]` and after generalization: `It is possible that a` **`specialist`** `can change a task status [...].`

Secondly, due to the incoherences in concepts' naming, both models have been aligned with names provided in the ontology. Every noun and verb in BPMN and rule models was compared against the ontology: both name and the possible usage described by object and data properties. If equivalent concept exists, it replaced word used in dynamic model. Furthermore ontology had to be filled with lacking concepts. In particular there were added many data type attributes due to the low-level characteristics of rules – every attribute used in conditional or decisive part had to be formally specified inside the related class. It was a very important step to achieve an automatic execution of hybrid BPMN/rule models.

As a result, both dynamic models of business logic shared the same taxonomy, along with attributes' types and relations between the concepts. Unfortunately, neither Prosecco's external services nor environments used to execute and validate these models do not support semantic data directly and there had to be introduced a solution to integrate them with ontology in an indirect manner. The proposed solution is based on automatically generated Plain Old Java Objects (so-called POJO), which represent ontological concepts in serializable and executable manner. The ontology is used to infer: a) types of classes (corresponding to the ontological classes), b) types of the class attributes (corresponding to data type properties), and c) relations between the classes (corresponding to object properties).

Within the practical implementation of the architecture depicted in Fig.1 we integrated the Prosecco ontology as the core of the system. The Prosecco system uses Activiti as the business process engine [12]. Apart from the type system generated according to ontology, Activiti uses also other semantization techniques. The Prosecco repository consists of several sub-repositories, including business process models for Activiti, rule models for Drools, user information in ACL, and system history managed by the Cassandra tool. Data repositories are separate from the type system and existing instances are continuously synchronized with ontology. Thanks to the use of the ontology the project assumes that all data types and their instances existing within the system are consistent with the ontology.

## V. Related Work

During the development of the Prosecco Ontology, existing organization ontologies were analysed and compared with Ontology discussed in this paper. Four models were considered:

1) *An organization ontology* [13] – developed by W3C and Epimorphics Ltd. and implemented in simple description logics language ($\mathcal{SIF}$(D)). It is aimed at describing basic organizational information in a number of domains. Comparing to the Prosecco Ontology, this model is less acurate and more general.
2) *IntelLEO Organization Ontology* [14] – created during IntelLEO project and currently not further developed. It models organization structures using people responsibilities and relations between them. This ontology models access rights what is outside the scope of Prosecco Ontology. On the other hand, it is not suitable for describing business processes.
3) *Ontology for organizations* [15] – is a part of a larger one that is used to annotate The Gazette[6] contents. It is best suitable for characterize activities of the organization (e.g. if it is a government or charity organization). This ontology does not provide concepts for describing organization structure, what was one of main goals of Prosecco Ontology.
4) *PROTON (PROTo ONtology)* [16] – upper-level ontology that is simple and well-documented. This is the biggest one among analysed models (it consists of 250 classes). As Prosecco Ontology, it describes projects, documents and organization structure. It lacks the support for business processes modelling.

Three more models draw attention, but due to the fact they are not public, there was no possibility to analyse them insightful and compare them to the Prosecco Ontology:

1) *Unified Enterprise Modelling Ontology (UEMO)* [17] – is based on UEML (Unified Enterprise Modeling Language) and depicts companies and information systems. It is coupled with BPM (Business Process Management).
2) *O-CREAM-v2, a core reference ontology for the CRM domain* [18] – is a very detailed CRM (Customer Relationship Management) ontology. One of its drawbacks is the lack of emphasis on services.
3) *WeCoTin* [19] – ontology designed to modelling process of matching the offer to the customer's requirements.

## VI. Evaluation and Summary

Knowledge management in business information systems often requires a unified dictionary of business concepts, that allows for a transparent integration of such systems. Thanks to it sharing the conceptualization between users becomes possible, and a better decision support facilities can be provided. In this paper we focused on the development of ontology-based mechanisms allowing for creating taxonomies of business logic concepts unifying system objects considered

[6] See: https://www.thegazette.co.uk/.

in the Prosecco project. We demonstrated how this ontology is used to unify vocabulary of business processes and rules in the Prosecco BPM system. The practical contribution discussed in the paper is the design and implementation of the ontology, and the demonstration of its practical use in the system.

The Prosecco Ontology, whose development and structure were described in the earlier sections, fulfills all the requirements of the project. As a result of careful design and iterative refactoring it can be considered as a formal definition of data and concepts appearing in the other parts of the Prosecco system, especially BPMN models and business rules. In comparison to less formalized data models it has certain advantages like mature tools, unambiguous semantics and possibility to formally prove coherence thanks to the available reasoners. Moreover, the formal approach to design system around a semantic model has proven to be a valid option among the standard solutions – the Prosseco Ontology is already successfully used to generate model layer of the system's services and future plans include usage of semantic queries to enhance data retrieval and storage.

A certain limitation of the current version of the ontology is the fact that it is based on the Polish vocabulary. However, it was a conscious decision and the requirement of the project related to the Polish regulations. In general most of the modules of the ontology could be easily adapted to other SMEs so the ontology would be suitable for international use. However, from the practical point of view an adaptation to some specific regional regulation (e.g. EU) should be considered. Certain parts specific to the Polish law would have to be replaced.

The biggest disadvantage of the proposed solution is lack of a direct integration between different modeling techniques and technologies; due to this limitation there was proposed an intermediate translation of semantic classes in a form of the corresponding Java classes. Currently the works are focused on the creation of environments which could help to model business logic enhanced by use of the ontological background.

Our future work include the use of the ontology to improve the usability of the execution environment. We are considering how semantic annotations could enhance intelligibility of the business logic. Thanks to the already integrated domain ontology, it would be easy to enhance rules with semantic annotations, leading to more meaningful design of the system, and possibly porting it to mobile platforms [20], and simplyfying a formalized modeling [21]. We are also working on a tighter integration of the components of the BPM system, including semantic tracking of business objects and a BP editor [22]. Moreover, we consider possible extensions of the ontology towards the needs of other SMEs from different sectors.

## Acknowledgment

The paper is supported by the Prosecco Project funded by NCBR, and co-supported by the AGH University Grant.

## References

[1] D. Jones, T. Bench-Capon, and P. Visser, "Methodologies for ontology development," in *Proceedings of IT&KNOWS Conference of the 15th IFIP World Computer Congress*, 1998.

[2] M. Bergman, "A brief survey of ontology development methodologies," Aug 2010. [Online]. Available: http://www.mkbergman.com/906/a-brief-survey-of-ontology-development-methodologies/

[3] M. Gruninger and M. S. Fox, "The design and evaluation of ontologies for enterprise engineering," in *Workshop on Implemented Ontologies, European Workshop on Artificial Intelligence, Amsterdam, The Netherlands*, 1994.

[4] M. Uschold and M. King, "Towards a methodology for building ontologies," in *IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Canada*, 1995.

[5] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, "Methontology: from ontological art towards ontological engineering," in *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series*. American Asociation for Artificial Intelligence, 1997.

[6] P. C. Benjamin, C. P. Menzel, R. J. Mayer, F. Fillion, M. T. Futrell, P. S. deWitte, and M. Lingineni, "IDEF5 method report," Knowledge Based Systems, Inc, Tech. Rep., 1994.

[7] M. Uschold and M. Gruninger, "Ontologies: Principles, methods and applications," *The knowledge engineering review*, vol. 11, no. 02, pp. 93–136, 1996.

[8] M. Horridge, H. Knublauch, A. Rector, R. Stevens, and C. Wroe, "A practical guide to building OWL ontologies using the Protege-OWL plugin and CO-ODE tools edition 1.0," Aug. 2004. [Online]. Available: http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf

[9] "Ontologydesignpatterns.org." [Online]. Available: http://ontologydesignpatterns.org/

[10] C. Del Vescovo, B. Parsia, U. Sattler, and T. Schneider, "The modular structure of an ontology: Atomic decomposition," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 3, 2011, pp. 2232–2237.

[11] D. Tsarkov, C. Del Vescovo, and I. Palmisano, "Instrumenting atomic decomposition: Software apis for owl," in *Proceedings of OWLED'13: the 10th International Workshop on OWL: Experiences and Directions*, 2013.

[12] T. Rademakers, T. Baeyens, and J. Barrez, *Activiti in Action: Executable Business Processes in BPMN 2.0*, ser. Manning Pubs Co Series. Manning Publications Company, 2012.

[13] D. Reynolds, "The organization ontology," W3C, Recommendation, Jan. 2014. [Online]. Available: http://www.w3.org/TR/vocab-org/

[14] J. Jovanovic and M. Siadaty, "IntelLEO organization ontology," IntelLEO, Working Draft, Apr. 2011. [Online]. Available: http://intelleo.eu/ontologies/organization/spec/

[15] J. Tennison, "The london gazette ontology, organisation module," 2008. [Online]. Available: https://www.thegazette.co.uk/def/organisation.owl

[16] K. Simov, A. Kiryakov, I. Terziev, D. Manov, M. Damova, and S. Petrov, "Proton ontology (proto ontology)," 2005. [Online]. Available: http://www.ontotext.com/documents/proton/protontop.ttl

[17] A. L. Opdahl, G. Berio, M. Harzallah, and R. Matulevičius, "An ontology for enterprise and information systems modelling," *Applied Ontology*, vol. 7, no. 1, pp. 49–92, 2012.

[18] D. Magro and A. Goy, "A core reference ontology for the customer relationship domain," *Applied Ontology*, vol. 7, no. 1, pp. 1–48, 2012.

[19] J. Tiihonen, M. Heiskala, A. Anderson, and T. Soininen, "Wecotin–a practical logic-based sales configurator," *AI Communications*, vol. 26, no. 1, pp. 99–131, 2013.

[20] G. J. Nalepa and S. Bobek, "Rule-based solution for context-aware reasoning on mobile devices," *Computer Science and Information Systems*, vol. 11, no. 1, pp. 171–193, 2014.

[21] M. Szpyrka, G. J. Nalepa, A. Ligęza, and K. Kluza, "Proposal of formal verification of selected BPMN models with Alvis modeling language," in *Intelligent Distributed Computing V. Proceedings of the 5th International Symposium on Intelligent Distributed Computing – IDC 2011, Delft, the Netherlands – October 2011*, ser. Studies in Computational Intelligence, F. M. Brazier, K. Nieuwenhuis, G. Pavlin, M. Warnier, and C. Badica, Eds. Springer-Verlag, 2011, vol. 382, pp. 249–255. [Online]. Available: http://www.springerlink.com/content/m181144037q67271/

[22] K. Kluza, K. Kaczor, and G. J. Nalepa, "Enriching business processes with rules using the Oryx BPMN editor," in *Artificial Intelligence and Soft Computing: 11th International Conference, ICAISC 2012: Zakopane, Poland, April 29–May 3, 2012*, ser. Lecture Notes in Artificial Intelligence, L. Rutkowski and [et al.], Eds., vol. 7268. Springer, 2012, pp. 573–581. [Online]. Available: http://www.springerlink.com/content/u654r0m56882np77/

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1327–1333

DOI: 10.15439/2015F267
ACSIS, Vol. 5

# Assessment of query execution performance using selected Business Intelligence tools and experimental agile oriented data modeling approach

Radek Němec
VŠB – Technical University of
Ostrava, Faculty of Economics,
701 21 Ostrava 1, Czech Republic
Email: radek.nemec@vsb.cz

□

*Abstract*—**The paper deals with the assessment of an experimental data modeling approach which is intended to support the agile oriented data modeling. The approach is based on the Anchor Data Modeling technique and is applied on a multidimensional data model. The assessed approach is expected to facilitate more effective execution of queries in the data mart environment. The emphasis is placed on the comparison of the query execution performance using database schemas, each built using traditional and the experimental approach. The tests are done in the environment of selected modern Business Intelligence tools, and using two test queries with varying output dataset sizes. The results show that the use of the database schema, created according to the experimental data modeling approach, had positive impact on the querying performance in several cases. The magnitude of impact on the querying performance, however, varied depending on each query's respective resulting dataset size.**

## I. Introduction

Multidimensional data modeling principles are one of the cornerstones of the Business Intelligence (BI) system's design and development process. These principles were introduced more than 3 decades ago by Ralph Kimball and then developed to today's well-known bus architecture and dimensional modeling methodology [1]. While the BI system is usually a critical decision-making and management support system, it is crucial to devote the best of care to its development and management. Although current trends strongly promote big data as the new Holy Grail for today's CIOs, the fact is that standard relational data marts will still be a thing in the coming future. Vast amounts of company's historical business data are stored predominantly in traditional relational database environments. Moreover, the business process management will always rely on the analysis of historical facts in relation to current real-time data. This is also supported in [2] where the authors state that the data and information integration are the most fundamental issues in the integration of decision support systems into processes, to enhance decision support performance.

The dynamics of event incidence and subsequent changes in the business world produce new business process management related requirements. Therefore, high impact on almost all aspects of the data mart architecture design, usage and its continuous adaptation and evolution is experienced almost on a regular basis. According to [3], one of the information system's success dimensions is the information timeliness and currency with respect to related business processes. The adaptation to the time aspect is therefore very important and it has an impact on the quality of data used in the decision-making process. This is supported also by [4] where the perception of data warehouse system's success dimensions by its users is studied. The authors determined that the quality of the underlying data is emphasized as an important antecedent of information quality and indirectly also of the system quality as important aspects of information system's success. Other aspects, like system performance (including query execution performance), service quality and usefulness of the BI system's toolset were also studied (the query execution performance is one of the main themes in this paper).

Changes usually encompass a set of data updates and schema changes, constraint modifications (keys, value containment) or metadata adjustments [5]. Kimball has already introduced and developed methods for the time-aware evidence of critical data in business dimensions represented in a multidimensional database schema. In terms of the data mart development, the standard Kimball's approach offers several non-destructive methods of capturing changes in values of dimensional attributes (i.e. horizontal changes) [6]. Although these methods are actively used in practice, some space is still left for further research in this part of the well-established data mart development paradigm [7]. Drawbacks of these methods usually reveal themselves sooner or later during most BI projects. These include mainly notable data volume increases and obvious limits in the length of change

□ This paper was made with financial support of the European Social Fund within the project CZ.1.07/2.3.00/20.0296.

978-83-60810-66-8/$25.00©2015, IEEE

capture time period. The assessment of an innovative approach addressing these issues in the data modeling practice are main research interests of the author of this paper.

Issues of business change adaptation dynamics are closely related to the agile development paradigm. Agile principles have already built a strong position in the software development area. In the data modeling practice, during the information system development life cycle, there is also an effort put into the adoption of such practices and principles. Generally, the agile approach emphasizes intense cooperation with customers as a vital asset. A core principle which is promoted by the "agile", and is especially relevant to the data modeling, is the modularity principle. The modularity facilitates easier adaptation of a software solution to changes in business requirements.

Also the multidimensional data model must reflect the most current business requirements and easy adaptation of the data model to these changes is an issue that needs to be solved most effectively. [8], [9] write about such agile oriented data modeling techniques applicable in the data warehouse development field. These techniques, although referenced as very effective in the practice, pose, however, rather agile oriented data modeling process management solutions than particular agile oriented data definition solutions. This issue is particularly relevant also to the development of the data mart architecture and is addressed by the experimental approach.

In recent works [10], [11], [12] an experimental (hybrid) data modeling approach was introduced. The proposed approach is focused on representing database schema of the multidimensional data model (designed using standard Kimball's dimensional modeling process) in an agile oriented fashion – i.e. as an implicitly modular data model. For this purpose, a data modeling technique, called the anchor data modeling (ADM), is leveraged and its principles and guidelines are adapted to the needs of the multidimensional data modeling field. The ADM technique was created by Olle Regardt, Lars Rönnbäck and their colleges, and fully described in [13].

In this paper, the emphasis is placed on the assessment of the usage of the experimental data modeling approach in the environment of selected client-side Business Intelligence tools. The goal of this paper is then to compare query execution performance in selected tools using the ADM based multidimensional database schema on one side, and a more traditional multidimensional database schema on the other side (both schemas are derived from a sample multidimensional data model described further in the text). The results will help to build more evidence of the possible applicability of the mentioned experimental data modeling approach. The mentioned approach, along with other methodological background, is described in the section II and the query performance assessment is presented in section III.

## II. Methodology

### A. Short overview of the Anchor Data Modeling technique

By definition, the ADM technique's application should lead to the creation of a highly decomposed database schema. The relations in such schema can change in time separately, both in terms of attribute values and structure in a more effective fashion. The implicit modularity feature brings the possibility of passing changes in the semantic background of the data model to changes in the final database schema without breaking the structure of source entities. This can make the data modeling more flexible and possibly promote non-destructive ways of managing changes even in multidimensional data models, according to most recent changes in the business environment.

The ADM technique is based on the usage of several distinctive conceptual constructors that are designed for easy and understandable representation of semantic terms and the evolution of the database schema. The resulting database schema is called the anchor database schema by default. The technique is intended to be a part of the relational data management paradigm, but the usage can easily span also in the object-oriented data modeling field [13]. Each entity is represented by one *Anchor* constructor and a set of *Attribute* constructors – both terms conceptually represent basic semantic features of an entity (name and properties). Each *Attribute* is dedicated to serve as a representation of each entity's property (i.e. attribute), containing usually only identifier (composite key) and an actual value of the entity's property. *Anchors* are connected using *Ties* which represent, by default, $M$ *to* $N$ relationships between entities (i.e. *Anchors*).

### B. Description of the experimental data modeling approach and differences from the traditional approach

Let a simple multidimensional data model be defined as a set of $i$ dimensions $DIM_i$ and a fact records subset. The fact records subset is matrix of records comprised of $n$ quantitative process performance measurements $M$ so that each fact record is represented as $facts = \{M_1, M_2, \dots, M_n\}$. Each dimension has $j$ dimensional properties $P$, i.e., each $DIM_i = \{P_{i,1}, P_{i,2}, \dots, P_{i,j}\}$. The description of the difference between traditional and the anchor data modeling (ADM) based database schemas follows.

Let the traditional multidimensional relational database schema, derived from the simple multidimensional data model ("trad."), be a set of $i$ relational tables, each for one dimension $DIM_i$, and one relational table for the fact records subset (the fact table). Each relational table $DIM_i$ consists of $j$+1 columns $C$, one column $C_K$ for the primary key and others for each property $P_j$. The whole tuple of the $i$-th dimension is then $DIM_i = \{C_{i,K}, C_{i,1}, C_{i,2}, \dots, C_{i,j}\}$. All dimensional properties are then included in one relational table. The fact table consists of $n$ respective metrics $M$ and a subset of composite key columns $K^*_{DIM}$, each referencing the $C_K$ column in one of the $DIM_i$. The whole tuple of the fact table of the "trad." schema is then $facts = \{K^*_{DIM_1}, K^*_{DIM_2}, \dots, K^*_{DIM_i}, M_1, M_2, \dots, M_n\}$.

Let the ADM based multidimensional relational database schema, derived from the simple multidimensional data model ("ADM"), be defined as a set of $i$ *Anchor* relational tables $A_i^{DIM}$ (one for each dimension $DIM_i$) and a set of $i \times j$ *Attribute* relational tables $Attr_{i,j}^{DIM}$ (one for each property $P_j$ of $i$-th dimension $DIM_i$). Each $A_i^{DIM}$ contains only identifying columns $C_{\mathrm{K}}$ of $i$-th dimension, i.e. $A_i^{DIM} = \{C_{i,\mathrm{K}}\}$. Inclusion of more attributes is, however, possible, e.g. to store ETL (ELT) process metadata (e.g. data quality related information). Each *Attribute* relation should contain only identifying composite key column and another column $C^*$ for $j$-th property of $i$-th dimension, i.e. $Attr_{i,j}^{DIM} = \{C^*_{i,\mathrm{K}}, C_{i,j}\}$. The primary key of the *Attribute* relation can then be extended with additional datetime columns if the historization is to be applied to the respective dimension's property. So one less *Attribute* relation must be created to represent the same dimension as it is in the traditional schema, but each *Attribute* should have identifying column, which is the obvious drawback of the approach, in this regard. The composite key column in the *Attribute* relation relates to the primary key column in the *Anchor* relation. Most modern database systems can perform elimination of tables in joins in query optimization procedures. This feature mitigates join demands of a hierarchically more extensive query by excluding tables from join from which no columns are selected to be used in the output of a query [13]. The fact table again consists of $n$ respective metrics $M$ and then a subset of composite key columns $K^*_{A^{DIM}}$, each referencing the $C_{\mathrm{K}}$ column in one of the $A_i^{DIM}$ relations. The whole tuple of the fact table is then, in fact, the same as for the "trad." schema, therefore $facts = \{K^*_{A_1^{DIM}}, K^*_{A_2^{DIM}}, \dots, K^*_{A_i^{DIM}}, M_1, M_2, \dots, M_n\}$.

The modular nature of the resulting anchor database schema implies normalization of relations into the 5th normal form. The schema can then be qualified as a highly decomposed[1]. Moreover, if the historization of *Attribute* values is applied, the *Attribute* relation is in the 6th normal form [14] and the primary key of the *Attribute* relation is complemented with a time validity aspect (i.e. column with a date and time data type).

The normalization of all relations in the anchor database schema into the 6th normal form is implied by original ADM technique's guidelines [13]. However, the results in previous related publication [15], concerning one of the first applications of the ADM in the multidimensional data modeling field, indicated that the normalization of the fact table leads to severe querying performance problems. Therefore, in the presented experimental approach, only Attribute relations in the schema should be normalized into the 6th normal form, if it is desired.

One of the expected benefits from using the ADM technique, when constructing a multidimensional database schema, relates to typical user behavior in the analysis and reporting – users typically select only few dimensional properties in their queries. Also, relevant values of these properties are usually filtered so that only a limited set of values is used to calculate the final results of the query. The high degree of normalization that is applied in the ADM based multidimensional database schema, then leads to the elimination of the need for scanning whole rows of a dimensional table (potentially large one). This aspect is then mirrored in expected query execution performance benefits. Also the straightforwardness and ease of applying implicit *Attribute* values' historization (each attribute can be historized separately) is a notable asset of the experimental approach. The high degree of dimensional properties' decomposition offers also a possibility of effective use of specific query processing optimization techniques, including compression and table data pre-ordering. This may bring some enhancements into already used relational data management environments, similar to those known to be present in columnar storage engines.

### C. Description of the sample multidimensional data model

The multidimensional data model that was used as a source for the creation of the sample database schemas is a typical banking model with 6 dimensions, as presented by Kimball and Ross [1]. Fact records represent monthly account data with cardinality of 50 million rows. Dimensions also come from the same publication and provide descriptive data to Accounts (220 000 rows), Households (i.e. Customers, 200 000 rows), Banking products (20 rows), Branches (1000 rows) and Account states (3 rows). The time dimension defining the monthly granularity of facts spans over 12 years (2000–2012; 156 rows). Fig. 1 provides a conceptual outlook on the structure of the sample multidimensional data model.

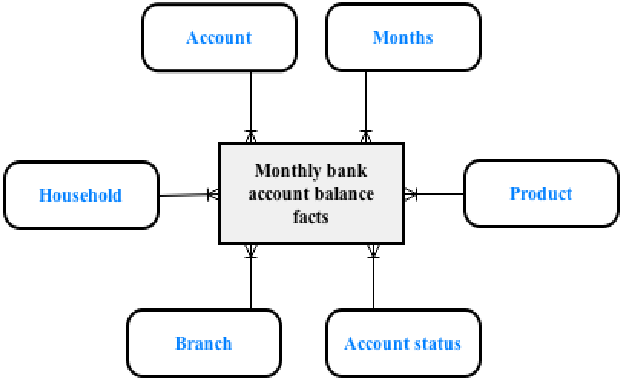


Fig. 1 Conceptual view on the sample multidimensional data model

### D. Description of the query testing process

None of the client-side tools used, however, support usage of an aggregated SQL query definition of data source for a

[1] From a conceptual point of view, the topology of the data model (and the database schema in the end) gets close to the snowflake topology, while the traditional data model stays with the classical star topology.

report. Both test queries therefore had to be stripped of aggregate functions and group by statements. This paper will, therefore, deal with query execution time (QET) results related to full dataset extraction because this mode is supported by all of the BI tools used in the test. The time dimension filtering condition was set to filter 5 specific calendar years (2008 to 2012) along with additional filtering options. In the appendix, there are both SQL queries listed, in both versions for both schema variants (with mentioned additional filtering options). Each query was then executed in each BI tool and query performance results were gathered using Microsoft SQL Profiler. This tool was used to get data on QET results, as well as actual CPU demands and the count of logical data reads performed during the execution of each query in each tool. Also the Apache jMeter 2.10 tool was used for the execution of specified test queries directly on the database server.

The testing process was done using a database server with Microsoft SQL Server 2012 64-bit bundle installed, with following server hardware configuration: Intel Xeon Quad-core 2.66 GHz CPU and 4 GB RAM. Although the server hardware is not the best-of-breed in the data warehousing field, I was able to get meaningful query execution performance results with it. Client-side tools were installed and run on a standard PC with 8 GB RAM and Intel Quad-Core processor.

### *E. Business Intelligence client-side tools for the assessment*

Gartner Research [16] published the 2015 version of their annual 'Magic Quadrant for Business Intelligence and Analytics Platforms' research report. The list of tools in this analytical report was a main source of tools that were assessed for the use in the process of query execution performance analysis. BI tool selection process respected following criteria:

1) the software is a client-side and stand-alone tool, without the need for installation of any application and/or OLAP server,
2) the software has a trial or free desktop version available for download,
3) the software is a BI reporting/dashboard management tool,
4) the software allows seamless connection to a relational database (SQL Server 2012 specifically).

The final list (Table I) includes 6 tools that passed the criteria and were successfully installed and run without stability and connectivity issues (both queries were executed without crashes and/or timeout problems). All tools are modern solutions, allowing for fast and intuitive data discovery and visualization. User interfaces offer self-service functionalities and also several data analysis features. Database connectivity features include connection to structured as well as unstructured data management solutions. The Tableau Desktop software also offers very fast on-the-fly in-memory computing capabilities. According to the Gartner Research report [16], all vendors of these tools, except for Tibco, are recognized as 'Leaders' in the magic quadrant chart (Tibco is viewed as a 'Visionary' and is located very close to the 'Leaders' quadrant). So the way how the data is handled and processed internally in each tool's software environment, and using respective connectivity interfaces, is the main differentiating factor for all 6 tools.

TABLE I.
SELECTED BI TOOLS AND DATABASE CONNECTIVITY PROPERTIES

| Short code | BI tool name | Database connectivity interfaces |
|---|---|---|
| Tibco | Tibco Spotfire Desktop 7.0 | .NET data provider |
| Tableau | Tableau Desktop 8.3 | ODBC |
| Lumira | SAP Lumira 1.23 | JDBC |
| MSTR | Microstrategy Analytics Desktop 9.4 | ODBC |
| Qlik | QlikSense Desktop 1.0 | ODBC, OLE DB |
| MSPP | Microsoft PowerPivot for Excel 2013 | OLE DB, .NET data provider, ODBC |

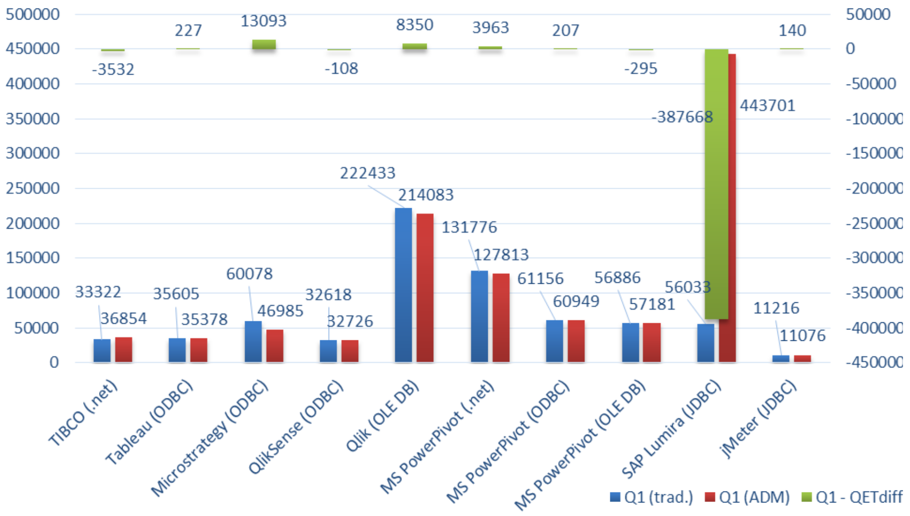


Fig. 2 QET results for the query Q1 in selected tools with QET differences between schema variants

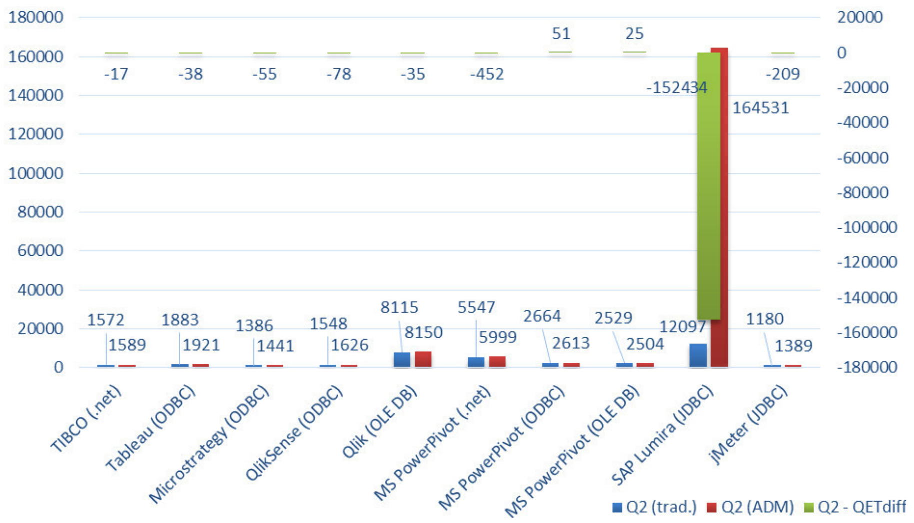


Fig. 3 QET results for the query Q2 in selected tools with QET differences between schema variants

## III. Results of Query Execution Performance Tests and Discussion

The results of the query performance testing process, which took into consideration both the highly decomposed ADM based database schema of the sample multidimensional data model and the traditionally structured counterpart provided interesting insights.

Fig. 2 (Q1) and Fig. 3 (Q2) show full dataset extraction QET results acquired for both queries after its execution in each tool as well as the on-server result obtained using the jMeter. Figures show also difference values $QET_{diff}$ that were calculated according to the formula: $QET_{diff} = QET_{trad.} - QET_{ADM}$.

The results for the query Q1 show that the usage of the ADM based database schema was beneficial in 4 cases as seen in the Fig. 2. The Tibco tool had bigger problems with the ADM based schema. The Qlik had slight problems too, but the $QET_{diff}$ was only -108 ms in this case. The usage of the OLE DB interface in the MSPP resulted in the $QET_{diff}$ = -295 ms which was also still relatively close to zero.

Since the query Q1 results in a large dataset to be extracted from the database (9 615 712 rows), the greatest difference may lie in the way of how each tool processes large datasets are being extracted. The Lumira is doing a bad job in this matter because resulting $QET_{diff}$ values were noticeably greater in comparison with other tools. The difference between QET value obtained for the ADM based schema and the traditional one came out largely in favor of the traditional schema. The client-side usage of the JDBC interface may have some part in the magnitude of QET value differences along with the connection to the database server through LAN. All in all, the Lumira, as a stand-alone client-side BI tools, is clearly not beneficial for usage with the ADM based schema. Table II shows a comparison of actual CPU demands and the amount of data reads performed during the usage of each tool.

Table II.
Physical characteristics of query Q1‘s execution

| BI tool | Interface | Q1 (trad.) | | Q1 (ADM) | |
|---|---|---|---|---|---|
| | | CPU | Reads | CPU | Reads |
| Tableau | ODBC | 13530 | 199556 | 14547 | 199586 |
| Tibco | .NET d. p. | 8827 | 199556 | 11391 | 199586 |
| MSTR | ODBC | 10109 | 199575 | 10938 | 199562 |
| Qlik | ODBC | 9141 | 199542 | 12812 | 200395 |
| | OLE DB | 12482 | 200081 | 12687 | 199626 |
| MSPP | .NET d. p. | 11857 | 199954 | 14033 | 200137 |
| | ODBC | 11936 | 199542 | 14124 | 199562 |
| | OLE DB | 12860 | 200107 | 14238 | 200375 |
| **mean** | | **11343** | **199739.1** | **13096** | **199853.6** |
| **std. dev.** | | **1761** | **259.1** | **1370.9** | **379.9** |
| jMeter | JDBC | 8360 | 199825 | 10000 | 199848 |
| Lumira | JDBC | 51641 | 38662071 | 429640 | 230975841 |

In the table II, it is visible that each tool’s internal algorithm of a large dataset processing plays an important role when using a more normalized data source (besides the perceived problem with Lumira).

Results for the query Q2 (Fig. 3) show that the usage of the ADM based database schema was beneficial in 2 cases (if we take into account only positive $QET_{diff}$ values). The lowest $QET_{diff}$ value is indicated only in the case of MSPP using .NET data provider interface ($QET_{diff}$ = -452 ms). In almost all cases, however, the $QET_{diff}$ was relatively low (in absolute terms and again excluding the Lumira case). The QET result obtained using the jMeter was very close to 4 tools which may, however, be mainly due to the nature of the query Q2 and resulting row count. Nevertheless, the $QET_{diff}$ = -209 ms which is relatively lower than most client-side tools (excluding cases of Lumira and MSPP using .NET interface). The size of the Q2’s resulting dataset (507 978 rows) points out on one possible effect that stems from the high normalization of the ADM based database scheme – if the size of the dataset gets lower the $QET_{diff}$ differences get closer to zero (except some cases).

Table III shows a comparison of actual CPU demands and the amount of data reads performed during the usage of each tool and query Q2.

TABLE III.
PHYSICAL CHARACTERISTICS OF QUERY Q2‘S EXECUTION

| BI tool | Interface | Q2 (trad.) | | Q2 (ADM) | |
|---|---|---|---|---|---|
| | | CPU | Reads | CPU | Reads |
| Tableau | ODBC | 1781 | 16033 | 1516 | 13959 |
| Tibco | .NET d. p. | 1281 | 15471 | 1595 | 13407 |
| MSTR | ODBC | 1514 | 15579 | 1418 | 13484 |
| Qlik | ODBC | 1655 | 16057 | 1671 | 13435 |
| | OLE DB | 1688 | 16145 | 2141 | 14456 |
| MSPP | .NET d. p. | 1968 | 16033 | 1874 | 13937 |
| | ODBC | 1640 | 15451 | 1891 | 13379 |
| | OLE DB | 1936 | 16033 | 1936 | 16042 |
| **mean** | | **1683** | **15850.3** | **1755.3** | **14012.4** |
| **std. dev.** | | **222.9** | **294.4** | **244.3** | **902.8** |
| jMeter | JDBC | 1595 | 15451 | 1717 | 13379 |
| Lumira | JDBC | 11625 | 12529763 | 94125 | 3824425 |

For this relatively smaller dataset (in comparison to the Q1's resulting dataset), it is evident that each tool's internal algorithm of the extracted dataset processing played relatively less important role, i.e. when using a more normalized data source (again besides the already perceived problem with Lumira's usage).

As for the related works, there are, regrettably, still no other works yet that deal specifically with the assessment of ADM technique's use. Besides the original paper [13] there are books [8], [9] that address agile oriented data modeling process guidelines in the traditional development of a data warehouse. However, there are works that deal with certain data modeling issues that are also addressed by the experimental approach (although mostly the conceptual part of the data modeling process is handled). Evolutionary aspects of the data in the data warehouse are dealt with in [17] and [18]. These works deal with the addition of temporal aspects into the UML based logical multidimensional data model (i.e. class model in which the conceptual model decomposition can be handled quite easily). In the paper [18], a rather general solution using only specific classes with temporal properties (a prototype based example of the application is presented) was proposed. In the paper [17], a more complex conceptual modeling approach was proposed. Along with modelling time-varying dimensional property objects, the authors propose also a solution for modelling time-varying hierarchies and hierarchy levels. Moreover, the way of mapping the resulting UML class schema to the entity relationship model of a data mart is proposed, although the paper lacks a practical example. In the paper [19], a graph theory based hybrid modeling method is introduced. The approach decomposes entity properties (attributes) and mutual relationships into graph nodes and edges. The schematic representation of data sources and requirements is then combined with the graph based representation and a conceptual multidimensional data model is derived. The changes in entities of the data model are induced by checking requirement-derived constraints, but time-validity aspects are not considered in the approach. The authors also present a short application example.

## IV. CONCLUSION

The query execution performance results were analyzed with the conclusion that the ADM based schema performed better in specific cases (BI tools). The performance differences were more in favor of the ADM based schema if the source dataset was larger rather than smaller. The larger dataset related results had, however, wider spread in maximum/minimum QET results. These facts will be studied in more detail in further research, also in contrast with the aggregated query execution performance results. Also, differences in the use and demands of particular physical query processing operators will provide important insights on how are the test queries internally processed.

The results indicate that the ADM based hybrid data modeling approach has certain future potential, although more evidence will be vital to justify its practical usefulness. The potential may be further increased if the application of more advanced query optimization techniques will have a positive effect on the query execution performance – both synthetic on-server and in-tool test results will be compared in this matter.

Also, further research effort will focus on using other BI tools, especially those that can or need to use separate OLAP/application server which may provide additional benefits. The expectation here is that the direct communication of the database server and the OLAP/application server may have additional benefits regarding query execution performance when using the ADM based multidimensional database schema.

## REFERENCES

[1] R. Kimball a M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 3rd ed., New York: Wiley, 2013.

[2] S. Liu, A. H. B. Duffy, R. I. Whitfield a I. M. Boyle, „Integration of decision support systems to improve decision support performance," *Knowledge Information Systems,* vol. 22, no. 3, pp. 261-286, 2010. DOI: 10.1007/s10115-009-0192-4.

[3] W. H. DeLone a E. R. McLean, „Measuring Success: Applying the DeLone & McLean Information Systems Success Model," *International Journal of Electronic Commerce*, vol. 9, no. 1, pp. 31-47, 2004.

[4] R. R. Nelson, P. A. Todd a B. H. Wixom, „Antecedents of Information and System Quality: An Empirical Examination Within the Context of Data Warehousing," *Journal of Management Information Systems*, vol. 21, no. 4, pp. 199-235, 2005. DOI: 10.1080/07421222.2005.11045823.

[5] E. A. Rundensteiner, A. Koeller a X. Zhang, „Maintaining Data Warehouses over Changing Information Sources," *Communications of the ACM*, vol. 43, no. 6, pp. 57-62, 2000. DOI: 10.1145/336460.336475.

[6] T. Torey, S. Lightstone, T. Nadeau a H. Jagadish, *Database Modeling and Design: Logical Design*, 5th ed., Burlington: Morgan Kaufmann, 2011.

[7] S. Rizzi, „Conceptual Modeling Solutions for the Data Warehouse," In *Data Warehouses and OLAP : Concepts, Architectures and Solutions*, Hershey, IGI Global, 2007, pp. 1-26.

[8] S. Ambler, *Agile Database Techniques: Effective Strategies for the Agile Software Developer*, New Jersey: Wiley, 2003.

[9] L. Corr, *Agile Data Warehouse Design*, Leeds: DecisionOne Press, 2014.

[10] R. Němec a F. Zapletal, „The Design of Multidimensional Data Model Using Principles of the Anchor Data Modeling: An Assessment of Experimental Approach Based on Query Execution Performance," *WSEAS Transactions on Computers,* vol. 13, pp. 177-194, 2014.

[11] R. Němec a F. Zapletal, „Analysis of Query Execution Performance Factors in the Anchor Multidimensional Database Schema Environment," in *Selected Paper of MEKON 2014 Conference*, Ostrava, 2014, pp. 105-117.

[12] R. Němec, „The Analysis of Historization Technique in Context of Handling Changes in Dimensions in Multidimensional Model and Anchor Data Modeling," in *Proceedings of the 10th International Conference on Strategic Management and Its Support By Information Systems SMSIS 2013*, Ostrava, 2013, pp. 135-146.

[13] O. Regardt, L. Rönnbäck, M. Bergholtz, P. Johannesson a P. Wohed, „Anchor Modeling: An Agile Modeling Technique Using the Sixth Normal Form for Structurally and Temporally Evolving Data," in *Conceptual Modeling - ER 2009 (Lecture Notes in Computer Science 5829)*, Rio Grande do Sul, 2009, pp. 234-250. DOI: 10.1007/978-3-642-04840-1_19.

[14] C. J. Date, H. Darwen a N. A. Lorentzos, *Temporal Data and the Relational Model: A Detailed Investigation into the Application of Interval and Relation Theory to the Problem of Temporal Database Management*, Oxford: Elsevier LTD., 2003.

[15] R. Němec, „The Comparison of Anchor and Star Schema from a Query Performance Perspective," in *World Academy of Science, Engineering and Technology, issue 71*, Paris, 2012, pp. 1718-1722.

[16] R. L. Sallam, B. Hostmann, K. Schlegel, J. Tapadinhas, J. Parenteau and T. W. Oestreich, "Magic Quadrant for Business Intelligence and Analytics Platforms 2015," Gartner Research, 2015, 2015-02-23, URL: http://www.gartner.com/technology/reprints.do?id=1-2ACLP1P&ct=150220&st=sb.

[17] E. Malinowski a E. Zimányi, „A conceptual solution for representing time in data warehouse dimensions," in *Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling*, Darlinghurst, 2006, pp. 45-54. DOI: 10.1145/1151855.1151861.

[18] F. Ravat, O. Teste a G. Zurfluh, „Towards Data Warehouse Design," in *Proceedings of the eighth international conference on Information and knowledge management*, Kansas City, 1999, pp. 359-366. DOI: 10.1145/319950.320028.

[19] F. Di Tria, E. Lefons a F. Tangorra, „GrHyMM: A Graph-Oriented Hybrid Multidimensional Model," in *Advances in Conceptual Modeling (ER 2011 Workshops, LNCS 6999)*, Brussels, 2011, pp. 86-97. DOI: 10.1007/978-3-642-24574-9_12.

## APPENDIX

Both queries Q1 and Q2, in both versions for both multidimensional database schema variants, follow:

```
/*Q1 – trad.*/
SELECT
  bproduct_nazev, /*SUM*/ pocet_transakci, kal_rok, mesic_nazev,
  bproduct_typ
FROM
  FACTmesicni_stav_uctu_snimek
  INNER JOIN DIMmesic ON FACTmesicni_stav_uctu_snimek.monthID
  = DIMmesic.monthID
  INNER JOIN DIMbankovni_produkt ON
  FACTmesicni_stav_uctu_snimek.bproductID =
  DIMbankovni_produkt.bproductID
WHERE
   kal_rok IN (2012, 2011, 2010, 2009, 2008)
   AND bproduct_typ IN ('běžný účet', 'hypotéka', 'leasing', 'termínovaný
   vklad', 'spotřebitelský úvěr')
```

```
/*Q1 – ADM*/
SELECT
  BP_BPN_bproduct_nazev, /*SUM*/ pocet_transakci,
  ME_KRO_kal_rok, ME_MNA_mesic_nazev, BP_BPT_bproduct_typ
FROM
  FACTmesicni_stav_uctu_snimek
  INNER JOIN BP_DIMbankovni_produkt ON
  FACTmesicni_stav_uctu_snimek.BP_ID_byBProdukt =
  BP_DIMbankovni_produkt.BP_ID
  INNER JOIN BP_BPN_bproduct_nazev ON
  BP_BPN_bproduct_nazev.bp_id = BP_DIMbankovni_produkt.bp_id
  INNER JOIN BP_BPT_bproduct_typ ON BP_BPT_bproduct_typ.bp_id
  = BP_DIMbankovni_produkt.bp_id
  INNER JOIN ME_DIMmesic ON
  FACTmesicni_stav_uctu_snimek.ME_ID_byMesic =
  ME_DIMmesic.ME_ID
  INNER JOIN ME_KRO_kal_rok ON ME_KRO_kal_rok.me_id =
  ME_DIMmesic.ME_ID
  INNER JOIN ME_MNA_mesic_nazev ON
  ME_MNA_mesic_nazev.me_id = ME_DIMmesic.ME_ID
WHERE
   ME_KRO_kal_rok IN (2012, 2011, 2010, 2009, 2008)
   AND BP_BPT_bproduct_typ IN ('běžný účet', 'hypotéka', 'leasing',
   'termínovaný vklad', 'spotřebitelský úvěr')
```

```
/*Q2 – trad.*/
SELECT
  pobockaID, /*COUNT*/ ucetID, DIMmesic.kal_rok,
  DIMpobocka.pobocka_adresa_mesto, DIMmesic.mesic_nazev
FROM
  FACTmesicni_stav_uctu_snimek
  INNER JOIN DIMucet ON FACTmesicni_stav_uctu_snimek.ucetID =
  DIMucet.ucetID
  INNER JOIN DIMmesic ON FACTmesicni_stav_uctu_snimek.monthID
  = DIMmesic.monthID
  INNER JOIN DIMpobocka ON
  FACTmesicni_stav_uctu_snimek.pobockaID = DIMpobocka.pobockaID
WHERE
   Year (DIMucet.ucet_otevren) IN (2008, 2009, 2010, 2011, 2012)
   AND DIMmesic.kal_rok IN (2012, 2011, 2010, 2009, 2008)
  AND DIMpobocka.pobocka_adresa_mesto IN ('Praha', 'Bohumín',
    'Ostrava', 'Jindřichův Hradec', 'Olomouc')
```

```
/*Q2 – ADM*/
SELECT
  PB_ID_byPobocka, /*COUNT*/ UC_ID_byUcet, ME_KRO_kal_rok,
  PB_PME_pobocka_adresa_mesto, ME_MNA_mesic_nazev
FROM
  FACTmesicni_stav_uctu_snimek
  INNER JOIN UC_DIMucet ON
  FACTmesicni_stav_uctu_snimek.UC_ID_byUcet = UC_DIMucet.uc_id
  INNER JOIN UC_UOT_ucet_otevren ON UC_UOT_ucet_otevren.uc_id
  = UC_DIMucet.uc_id
  INNER JOIN ME_DIMmesic ON
  FACTmesicni_stav_uctu_snimek.ME_ID_byMesic =
  ME_DIMmesic.me_id
  INNER JOIN ME_KRO_kal_rok ON ME_KRO_kal_rok.me_id =
  ME_DIMmesic.me_id
  INNER JOIN ME_MNA_mesic_nazev ON
  ME_MNA_mesic_nazev.me_id = ME_DIMmesic.me_id
  INNER JOIN PB_DIMpobocka ON
  FACTmesicni_stav_uctu_snimek.PB_ID_byPobocka =
  PB_DIMpobocka.pb_id
  INNER JOIN PB_PME_pobocka_adresa_mesto ON
  PB_PME_pobocka_adresa_mesto.pb_id = PB_DIMpobocka.pb_id
WHERE
  Year (UC_UOT_ucet_otevren) IN (2008, 2009, 2010, 2011, 2012)
  AND ME_KRO_kal_rok IN (2012, 2011, 2010, 2009, 2008)
  AND PB_PME_pobocka_adresa_mesto IN ('Praha', 'Bohumín', 'Ostrava',
  'Jindřichův Hradec', 'Olomouc')
```

DOI: 10.15439/2015F226
ACSIS, Vol. 5

# Enhanced simulation performance through parallelization using a synthetic and a real-world simulation model

Tommy Baumann*¶, Bernd Pfitzinger‡§ , Dragan Macos†, Thomas Jestädt‡
*Andato GmbH & Co. KG, Ehrenbergstraße 11, 98693 Ilmenau, Germany. tommy.baumann@andato.com
¶Hochschule Aalen – Technik und Wirtschaft, Beethovenstraße 1, D73430 Aalen.
‡Toll Collect GmbH, Linkstraße 4, 10785 Berlin, Germany. {bernd.pfitzinger|thomas.jestaedt}@toll-collect.de
§FOM Hochschule für Oekonomie & Management, Zeltnerstraße 19, 90443 Nürnberg, Germany.
†Beuth Hochschule für Technik Berlin, Luxemburger Str. 10, 13353 Berlin, Germany. dmacos@beuth-hochschule.de

***Abstract*—Taking an existing large-scale simulation model of the German toll system we identify possibilities for parallelization in order to enhance simulation performance. We transform parts of the model from its current serial implementation to a parallel implementation. Afterwards we evaluate the achieved performance enhancement and compare the results to a synthetic benchmark model.**

## I. Introduction

As technology advances in electronics, systems and processes with higher complexity, interconnectedness and heterogeneity can be developed. Simultaneously, user requirements are constantly increasing. Unfortunately “more is different” [1] as it is summarized in the definition of a “distributed system” as “one in which the failure of a computer you didn’t even know existed can render your own computer unusable” [2]. Modeling and simulation techniques are applied to design, analyze, evaluate, validate, and optimize such complex systems. Especially in specification and design stage executable models deliver tremendous value by lowering the system design uncertainty – even expert advice is known to be over-confident [3], a well-known cognitive bias that needs to be mitigated by the system design process. Yet in many parts of everyday life people depend [4] on software-intensive systems.

The use of simulation models is one way to increase the specification speed and quality [5]. Hence, current system design approaches like Simulation Driven Design [6] are characterized by applying executable models to a large extend. A prerequisite to apply executable models is a so called execution domain: In our context Discrete Event Simulation (DES) [7] has gained significance. DES is used in many industries, e.g. energy, telecommunications, production, logistics, avionics, automotive, business processes, and system design. Inter alia DES is applied for dimensioning of resources, to answer questions about topology (e.g. [8]), scalability and performance regarding operational scenarios, to predict system behavior, and to estimate risks. Increasingly the performance in defining and executing models becomes vital due to the increased complexity of systems and processes as well as the customer requirement to create holistic, integrated, high accuracy models up to real world scale. Several use cases of simulations are only possible once the simulation performance is ‘good enough’: simulating the long-term dynamic behavior, iterative optimization loops, automatic test batteries, real-time models (higher reactivity to market demands and changes), and automated specification and modeling processes (including model transformation/generation) [9]. In this context Parallel Discrete Event Simulation (PDES) [10] helps to provide the necessary simulation performance. In the article we identify possibilities for parallelization of a large-scale simulation model of the German toll system implemented in MSArchitect [11]. We transform parts of the model from the current serial, nonparallel implementation to a parallel implementation. Afterwards we evaluate the achieved performance enhancement.

The outline of the article is as follows: Section II gives an overview of the automatic German toll system and the corresponding simulation model. Typical use cases involve simulations at a scale of 1:1 spanning time periods of up to one year – necessitating a high-performance simulation model. The aim of this article is to investigate parallelized simulation models. To that end section III introduces the simulation framework architecture followed in section IV by a synthetic benchmark model for the use in PDES simulation. Section V describes the parallelized simulation model of the German automatic toll system and evaluates the simulation performance achieved. Section VI summarizes the results and describes future work and applications of our simulation model.

## II. Simulation model of the German toll system

The German automatic toll system is a typical example of a state-of-the-art tolling system [12, 13] based on global navigational satellite systems (GNSS). At present it is the largest system of its kind in operations – collecting more than 4.3 bn€ annually [14–16]. It was the first large-scale GNSS-based tolling system with more than 800 000 on-board-units (OBUs) deployed at present. More than 90% of the tolls are

978-83-60810-66-8/$25.00©2015, IEEE

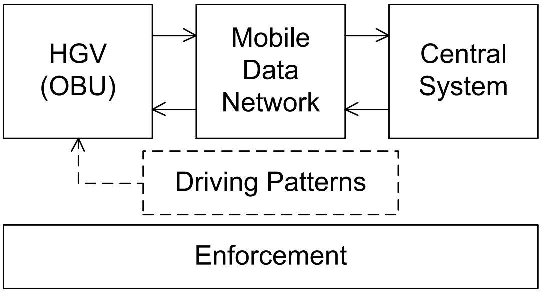


Fig. 1. High-level system design of a GNSS-based electronic tolling system and its dependency on the user interaction (driving patterns).

collected automatically the remainder using a manual process via internet or one of around 3 500 toll-station terminals.

The generic architecture of a GNSS-based tolling system is given in figure 1: The tolls are collected via OBUs installed in the heavy goods vehicles (HGVs) and transmitted via mobile data networks to the central system for processing. A separate part of the toll system is responsible for the enforcement and in the case of the German toll system an additional manual mode of tolling is implemented (not shown in the figure). The simulation model implements those technical systems and processes relevant for collecting tolls and for deploying updates to the OBUs: The vehicle fleet with its OBUs, the central systems required – i.e. a typical DMZ and the servers to receive toll data or provide updates – and the IP-based communication via mobile data networks.

Modeling the toll system we aimed to include all relevant processes at time scales of one second or longer, the technical systems in turn generate events with a higher temporal resolution in the model. However, the model of the tolling system needs to be accompanied by a model of the user behavior ("driving patterns" in figure 1) – the points in time when a given HGV is powered on or off, creates a toll event, looses or recovers its connection to the mobile data network.

The simulation model is applied in forecasting the dynamic behavior of the real-world tolling system – either the operations of the existing system (see section II-C) or for anticipated changes to the system (see sections II-A and II-B). Each application requires the simulation to predict the system behavior over several months – where the typical work-flow expects the simulation results to be available on the next business day. The existing simulation model is implemented to utilize a single CPU core and considerable effort has been expended to achieve an adequate performance [17–19]. However, over time the level of detail included in the simulation model and the number of OBUs in the real-world system increased. In addition simulation runs should deliver medium- and long-term predictions encompassing six months to one year.

### A. *Simulations in the Analyze and Design Phase*

In the analysis phase we use the simulation model to validate the system requirements. Addressing the large amount of requirements in typical software-intensive systems, requirements are defined at different levels of abstraction: The top level defines why the system is built and what the owning organization hopes to achieve. This type is termed as business or stakeholder requirements. Already at this level-of-abstraction the requirements need to be validated as soon as possible – is the requirement really necessary at the documented level? Seemingly inconsequential numerical targets can have profound effects on the technical solutions, e.g. by necessitating a high-availability architecture.

The translation of the requirements into an executable specification (or simulation model) allows exploring the effects of the requirements on the solution space early on and vice versa: The virtual prototype transports operational properties of the real-world system back to the solution space potentially modifying or restricting the requirements. An example of our approach is the comparison of a thin-client and a thick-client solution in the domain of electronic tolling [20].

Our focus in this phase is to avoid system operation faults based on wrong assumptions or conjectures of the new requirements. The most important aspects checked by us are:

- Excessing the capacity limits of the key subsystems
- Appearance of non-valid system states
- The worst case scenarios in case of eventual system failures.

The aspects we want to simulate are designed manually. The most frequent checks are based on the system behavior checking concerning various system parameter values.

In the design phase the system architecture becomes more detailed: The analysis model is linked with the used frameworks, libraries and other third party software components such as database or GUI. The executable specification helps in drafting accurate requirements and the simulation runs yield the resulting dynamic behavior prior to the implementation of the system. At the same time, the simulation model becomes in itself more specific by adding the necessary behavior and parameters to allow measuring its performance. Depending on the complexity and runtime the optimization can be delegated to an automatic optimization algorithm.

In the design phase we can validate the system behavior with improved functional granularity. During this phase the software development process invests in design documents – consistent, fine-grained and formal models start to show up in the documentation. In our approach we can take these models as a starting point and use model transformations to transform the architecture models into the appropriate simulation model.

### B. *Simulations in the Develop and Test Phase*

The implementation phases shift the focus to the system under construction. Here the requirements are supposed to remain fixed and only minor adjustments need to be returned to the requirements repository. The simulation model is an executable representation of the state-of-knowledge and is technically able to integrate a given component into the overall system – especially as long as the whole system is not yet available.

In that way simulations are part of the decision making process: Implementation variants can be explored and com-

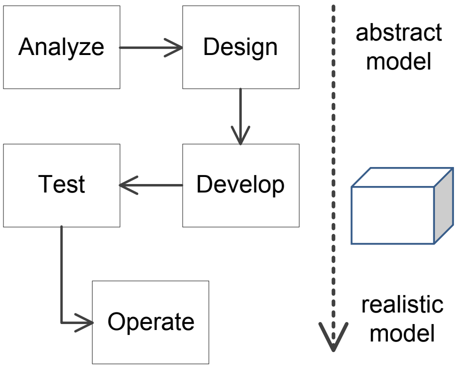


Fig. 2. The system development process (left) can be accompanied by a simulation model (right)

pared through simulations. The developed components are functionally integrated (transformed) and evaluated in the simulation environment. For an example one could look at the avionics industry [5, 8, 21].

In the test phase, the high-level requirements are used for acceptance tests of the whole system. Usually the development of the virtual prototype is considerably ahead of the real system. Therefore the soft- and hardware components are initially integrated into a whole working simulated system rather than the real world components: So called Software-in-the-Loop (SiL) and Hardware-in-the-Loop (HiL) tests. The simulation model provides the still missing components and allows to test dynamic coupling effects even when not all components of the real system are available. In other words we use the simulation environment as a stimulus for testing the individual software components – the simulation environment is a test driver for the developed software components. The test drivers are currently generated manually. Our goal is to generate them fully automatically via UML-enhancements to define the simulation test drivers during the software components design.

### C. Simulations supporting System Operations

In the case of the German toll system the main objective of the simulation model is to support and to safeguard the day-to-day operations of the automatic toll system. To that extent the simulation model includes the most important processes (receiving toll data from the OBUs and sending updates to the OBUs) in a realistic way at a scale of 1:1. The dynamic behavior – with a strong daily and weekly rhythm – the predictions of the simulation model can and must be verified against data from the real-world system [22, 23] and typically show a good correlation [24].

Simulation models of the kind discussed here address the operator of a software-intensive system-of-systems. Typically the operator faces two challenges simultaneously: The day-to-day operations and the necessary changes and updates within the technical systems. Of course, the handling of technical details can and typically will be outsourced to specialized providers. The one challenge that remains for the system operator to handle is the integration of all technical and organizational parts into a working whole [25].

As in the test-driven-development (TDD) case, testing is not the aim of simulation-driven-development (SDD) rather the "driven [...] focuses on how TDD leads analysis, design, and programming decisions" [26]. In that sense, SDD tries to put the design to the ultimate test-case – the real-world operational context. The simulation model – an executable specification of the existing real-world system – is the starting point to focus any software development on the operational consequences. These consequences might be of a purely technical nature, e.g. the system architecture and performance, or include non-functional requirements and business or financial aspects. In particular these challenges dominate environments that are rich in legacy systems. The on-going development of theses systems is largely faced with integration issues between systems [27]. SDD addresses integration of systems as a cross-cutting concern by providing the software developer (or requirements engineer) with an executable copy of the real-world system.

## III. Simulator architecture and benchmark model

To transition from a sequential DES simulation model to a parallel one we first look at the possibilities offered by the simulation and modeling tool in use (see section III-A) and different ways of parallelizing existing models as discussed in literature (section III-B). The section ends with brief remarks on the PDES performance in general – the next section introduces and discusses a particular benchmark model.

### A. Simulation and modeling tool, DES and PDES performance considerations

To model and simulate the structural and behavioral properties of the German toll system we selected and applied a system design tool for modeling and executing DES models [11]. The tool is specialized in integrated design of complex distributed systems and processes across different design levels. It offers a unique blend of performance [28] and customizability to manage extremely complex models within diverse usage environments. The simulation tool consists of several separated, mostly platform independent components, as a graphical user interface including a multi model editor, a simulation kernel, a library of standard model components, and a mission controller to run multiple simulations in parallel.

The simulation kernel used supports the execution of sequential DES and parallel DES (PDES) models. With DES and PDES the operation of a system is expressed as a discrete sequence of events in time. Each event occurs at a particular instant in time and marks a change of state in the system. Between consecutive events, no change in the system is assumed to occur. Thus the simulation can directly jump in time from one event to the next. All events are managed by a so called future event list (FEL).

### B. How to approach parallelizing DES models

PDES can substantially improve the performance and capacity of simulation, allowing the study of larger, more detailed

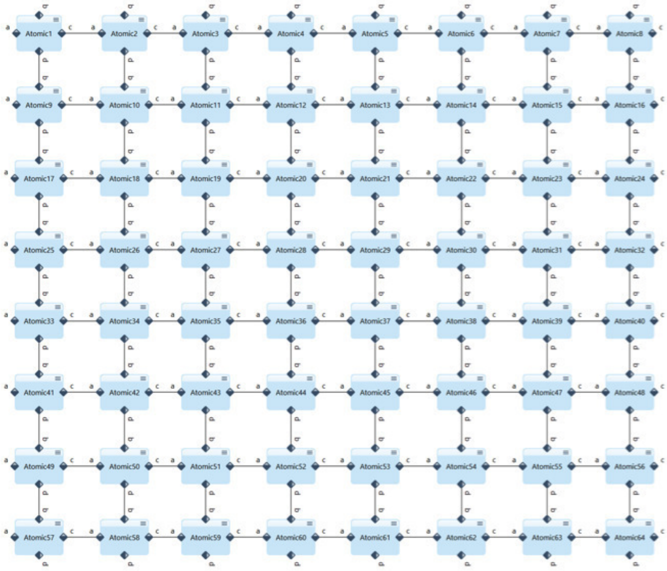


Fig. 3. Block diagram with a synthetic benchmark model using 8x8 network nodes.

models, in less time, and to be able to scale a problem if necessary. Thereby the performance and scalability is limited by communication latencies between the participating cores and nodes. Unfortunately, a prerequisite of parallelization is the decomposition of the model for processing on multiple processors or processor cores – it might become necessary to refactor existing serial mode models. This can be done in several ways [29]:

- Through the use of parallelizing compilers,
- In the form of replicated trials (run multiple serial simulations in parallel) or
- Distributed functions or
- Distributed events (with centralized FEL).
- Especially for optimizations a simple approach is the time-parallel domain decomposition (run multiple serial simulation in sequence) whereas
- the space-parallel domain decomposition (run multiple model parts in parallel) is used to accelerate a single simulation run.

The last enumerated decomposition approach, the space-parallel decomposition, is used by our simulation tool. Here, the simulation model is decomposed into sub-models or components. Each component is assigned to a process, where several processes may run on the same processor. This approach is applicable to any model and shows the greatest potential in offering scalable performance for complex models [30, 31]. Since the FEL is also decomposed into individual local FELs, it would never become the bottleneck. A higher degree of parallelism is expected because concurrent processing is encouraged.

In general, PDES approaches can be divided into two categories — conservative and optimistic — according to the way they handle the causality constraint of local FELs. Violating the causality constraint means that the future can affect the past leading to incorrect simulation results. In our case the simulation tool supports an optimistic synchronization mechanism, also known as time warp [32]. Causality errors are detected and fixed using rollback algorithms at the additional cost of necessitating rollback functions within the model.

### C. *PDES performance*

Optimistic synchronization mechanisms have inherently more overhead than conservative synchronization mechanisms. The overhead includes e.g. state saving, global virtual time calculation and rollback steps. The degree to which they may affect the simulation performance depends among other things on the granularity of the model to be simulated and the support from the hardware. Both approaches – conservative and optimistic – have their advantages depending on the specific application case. In terms of general purpose simulation the optimistic approach appears to be slightly in advantage [33].

## IV. PDES Benchmark model

In this section we introduce a synthetic network model consisting of a grid of simple networking units (section IV-A) and give the sequential and parallel simulation performance for different model sizes in section IV-B.

### A. *Description of the synthetic model*

To evaluate the potential of parallelization we developed a synthetic benchmark model using our modeling and simulation tool. The model consists of several communication network nodes. When a network node receives an event, it consumes this event, handles some synthetic workload and sends a new event to a randomly selected neighbor network node with random delay. The synthetic workload on every event is approximately 10 000 floating point operations and the random delay is exponentially distributed with a mean of 5. Initially

TABLE I
Computation time in msec

| | Threads | 64 Nodes | 128 Nodes | 256 Nodes |
|---|---|---|---|---|
| Serial | 1 | 8922 | 8953 | 8953 |
| Parallel | 2 | 5344 | 5109 | 4953 |
| | 3 | 3953 | 3719 | 3500 |
| | 4 | 3218 | 2985 | 2734 |
| | 5 | 2828 | 2532 | 2313 |
| | 6 | 2531 | 2235 | 2015 |
| | 7 | 2313 | 2016 | 1782 |
| | 8 | 2172 | 1875 | 1641 |

TABLE II
Speedup factor

| | Threads | 64 Nodes | 128 Nodes | 256 Nodes |
|---|---|---|---|---|
| Serial | 1 | 1 | 1 | 1 |
| Parallel | 2 | 1,67 | 1,75 | 1,81 |
| | 3 | 2,26 | 2,41 | 2,56 |
| | 4 | 2,77 | 3 | 3,27 |
| | 5 | 3,15 | 3,54 | 3,87 |
| | 6 | 3,53 | 4,01 | 4,44 |
| | 7 | 3,86 | 4,44 | 5,02 |
| | 8 | 4,11 | 4,77 | 5,46 |

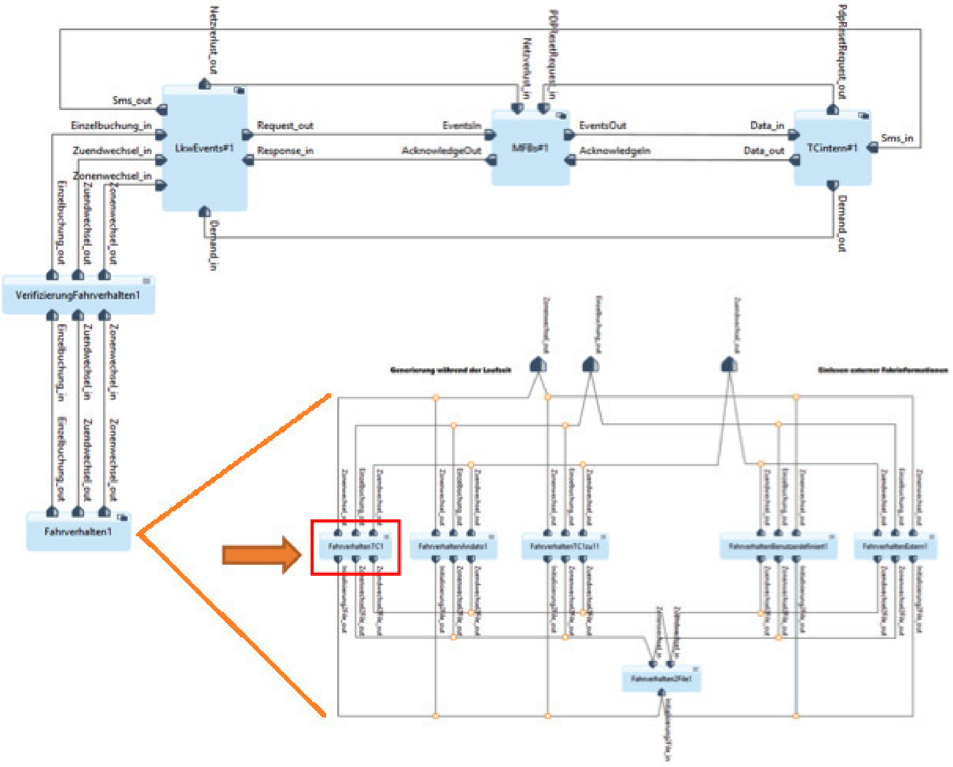


Fig. 4. Component responsible for generation of driving patterns

every second network node sends an event, which results in an event density of 50 percent. During a simulation run the total number of events in the system remains constant, due to the similar parametrization and behavior of all network nodes. Figure 3 shows the block diagram of the model with 8x8 network nodes.

### B. *Parallel and serial performance of synthetic model*

In order to evaluate the simulation performance we simulated three sizes of the network model (8x8, 8x16 and 16x16 network nodes) using 1 to 8 threads. The clustering of the model and the assignment of these clusters to threads was done automatically by the simulation kernel. All simulations were executed on a PC using a i7-3635QM CPU with 2.40GHz, 4 cores and 8 threads and the simulation kernel of MSArchitect® Enterprise in version 2.3 with 64Bit.

The results are summarized in tables I and II. The first table gives the computation time needed for the simulation runs and the second table lists the relative speedup factor for each parallel simulation compared to the serial simulation. In theory the optimal speedup factor could be equal to the number of threads resp. cores used. Practically this is not possible due to the communication and synchronization overhead between the threads and the fact that the CPU does not offer 8 threads of equal computational power. All in all a substantial performance increase of up to 5.46 with 8 threads can be observed – higher than the 4 CPU cores offered and lower than the 8 parallel threads offered by the CPU.

## V. Parallelization of the simulation model

The synthetic benchmark model presented in the previous section allowed a sizable speed-up through the use of a PDES simulation kernel. In this section we look at the challenges when an existing simulation model is the starting point of the parallelization effort (section V-A). In particular, some parts of the model are intrinsically serial – as required from the particular application domain. Section V-B discusses the process of selecting appropriate parts of the simulation model for parallelization.

### A. *Partial transformation as proof of concept*

After successfully evaluating the parallel simulation approach using a synthetic model and the promising performance potential we decided to shift from serial to parallel simulation for the model of the toll system. Since the parallelization is a time-consuming process, we pursue an incremental model transformation, starting with the performance critical parts. In a first step we looked to the application-level performance of our model of the German toll system to locate appropriate model components for parallelization. This was done using both the kernel logging capabilities of our simulation tool and an external profiling application [17, 18]. Kernel logging allows to count the number of calls of atomic models as well as the total number of samples (corresponding to a processor cycle). The external profiler allows measuring the space complexity (memory), the time complexity (duration,

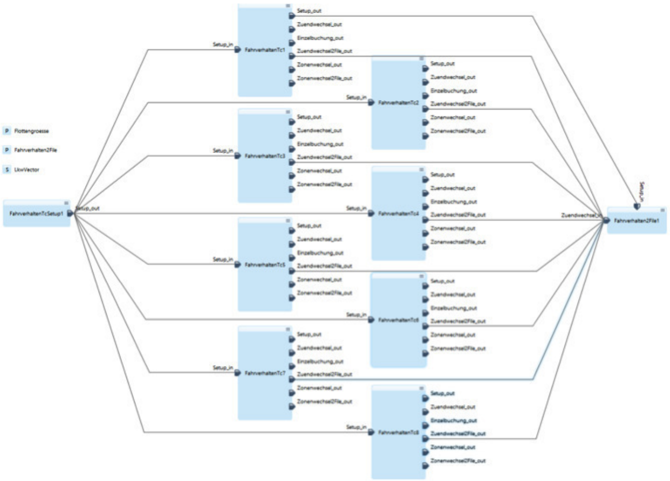


Fig. 5. Partial parallelization of the simulation model: Using 8 threads in generating the driving patterns.

CPU time), and the usage of particular instructions of a target program by collecting information on their execution.

To implement a partial parallelization we choose a setup using a similar hardware and software constellation as the synthetic model. The fleet size used in the simulation runs was set to 800 000 OBUs and a time frame of 20 weeks was considered. In the simulation runs the component responsible for generation of driving patterns was identified as performance critical (figure 4). This component consumes about 9.7% of the overall computation time in serial mode.

### B. Selecting parts of the model for parallelization

To allow the execution of a given simulation model in parallel simulation mode the underlying PDES kernel calls a different set of atomic interface functions than in the serial mode: Instead of the function `run()`, which is called every time an event is received, the functions `runForward()`, `runReverse()` and `runCommit()` need to be implemented. This is necessary in the case of optimistic synchronization since this mechanism allows the speculative execution of events that might require a rollback to ensure causality. In addition the model architecture needs to be adapted following three rules:

- Reduce the dependencies between model components to simplify clustering,
- Transfer external states to internal states or to port based communication,
- Restructure the internal architecture of model components to enable functional parallelization, while leaving the interfaces stable.

In our case the selected model component is responsible for the generation of driving patterns for the whole vehicle fleet. We refactored this component so that the internal structures were multiplied according the number of parallel simulation threads (in our case 8) – the driving patterns of different heavy-goods-vehicles are inherently independent in our model. In that way each internal structure is responsible for one-eighth of the vehicle fleet, each executed as a separate thread (see figure 5). Executing the driving patterns component separately from the overall model in parallel mode let to a computation time of 3 min 30 sec compared to 8 min 28 sec in serial mode. This means a speedup of factor 2.4, which is quite good compared to the theoretical speedup factor of the synthetic model of factor 5.46.

### C. Complete parallelization and backward compatibility

To parallelize the model of the German toll system completely the architecture of about 60% of the model components (composite blocks) need to be reworked according to the three architectural rules mentioned above. The remainder (mostly atomic blocks) need at least to be transferred to the set of parallel atomic interface functions. Both interfaces – for serial and parallel execution – can coexist in the simulation tool chosen for this work. This allows for an incremental porting of complex simulation models towards a fully parallelized model.

## VI. Summary

Simulation models offer the ability to transform the software development process by placing each step into the realistic operational context. One important pre-condition is that the simulation model is sufficiently fast to generate realistic behavior at a scale of 1:1. Staying with a realistic level-of-abstraction the application of some models is limited by the execution time. In our example the large number of objects at run-time – models of on-board-units – and the necessity to predict the dynamic behavior of the system for a considerable length of time limit the use of simulation runs in the day-to-day decision making process.

In this paper we looked into increasing the simulation performance through parallelization either automatically or at the level of the simulation model. To further the discussion we implemented a synthetic network model with a configurable grid of identical components as a benchmark for the automatic parallelization achieved by the simulation kernel. In this case the speed-up compared favorably with the performance offered by the CPU chosen to run the benchmark.

In the real-world example the speed-up is more difficult to achieve – large parts of the model are either difficult to parallelize or contribute negligibly to the overall computation time. Profiling the simulation model we identified one part – the model of the user behavior – as both particularly time consuming and compact (as expressed by the size of this particular part of the model). Parallelizing only a part of the model limits the speed-up: The remaining model parts run in serial mode and we achieve only a speed-up of 2.4 using 8 threads. However, this speed-up is the result of re-factoring only 5% of the simulation model.

Future work is needed to automate the parallelization of existing models and to steer the modeling engineer to the most promising parts of a given simulation model.

## References

[1] Philip W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. doi: 10.1142/9789812385123_others01.

[2] L. Lamport. Distribution, May 1987. URL http://research.microsoft.com/en-us/um/people/lamport/pubs/distributed-system.txt. [accessed 19-Mar-2015].

[3] Dale Griffin and Amos Tversky. The weighing of evidence and the determinants of confidence. *Cognitive psychology*, 24(3):411–435, 1992. doi: 00l0-0285/92$9.00.

[4] Algirdas Avizienis, J-C Laprie, Brian Randell, and Carl Landwehr. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing*, pages 11–33, 2004. doi: 10.1109/TDSC.2004.2.

[5] Tommy Baumann. *Automatisierung der frühen Entwurfsphasen verteilter Systeme*. Südwestdeutscher Verlag für Hochschulschriften, Saarbrücken, Germany, 2009. ISBN 978-3-8381-1266-4.

[6] Tommy Baumann. Simulation-driven design of distributed systems. *SAE Technical Paper*, 2011. doi: 10.4271/2011-01-0458.

[7] E. A. Lee and D. G. Messerschmitt. Static scheduling of synchronous data flow programs for digital signal processing. *IEEE Transactions on Computers*, C-36(1):24–35, 1987. doi: 10.1109/TC.1987.5009446.

[8] N. Fischer and H. Salzwedel. Validating avionics conceptual architectures with executable specifications. *Journal of Systemics, Cybernetics & Informatics*, 10(4), 2012.

[9] Bernd Pfitzinger, Tommy Baumann, and Thomas Jestädt. Network resource usage of the German toll system: Lessons from a realistic simulation model. *46th Hawaii International Conference on System Sciences (HICSS)*, pages 5115–5122, 2013. doi: 10.1109/HICSS.2013.415.

[10] Richard M. Fujimoto. *Parallel and distributed simulation systems*, volume 300. Wiley-Interscience New York, 2000. ISBN 978-0471183839.

[11] Msarchitect. URL http://www.andato.com/. [accessed 10-Dec-2012].

[12] Julia Numrich, Sascha Ruja, and Stefan Voß. Global navigation satellite system based tolling: state-of-the-art. *NETNOMICS: Economic Research and Electronic Networking*, 13(2):93–123, 2012. doi: 10.1007/s11066-013-9073-9.

[13] Andrew T. W. Pickford and Philip T. Blythe. *Road user charging and electronic toll collection*. Artech House London, 2006. ISBN 978-1-58053-858-9.

[14] Bundesministerium der Finanzen. Haushaltsabschluss 2011, Feb. 2012. ISSN 1618-291X. URL www.bundesfinanzministerium.de/Content/DE/Monatsberichte/Publikationen_Migration/2012/02/inhalt/Monatsbericht-Februar-2012.pdf?__blob=publicationFile&v=3. [accessed 09-May-2012].

[15] Bundesministerium der Finanzen. Sollbericht 2013. *Monatsbericht des BMF*, (2):6–22, Feb. 2013. ISSN 1618-291X. URL http://www.bundesfinanzministerium.de/Content/DE/Monatsberichte/2013/02/Downloads/monatsbericht_2013_02_deutsch.pdf?__blob=publicationFile&v=4. [accessed 20-Mar-2013].

[16] Bundesministerium der Finanzen. Haushaltsabschluss 2013, Jan. 2014. ISSN 1618-291X. URL http://www.bundesfinanzministerium.de/Content/DE/Monatsberichte/2014/01/Downloads/monatsbericht_2014_01_deutsch.pdf?__blob=publicationFile&v=6. [accessed 26-Nov-2014].

[17] Tommy Baumann, Bernd Pfitzinger, and Thomas Jestädt. Simulation driven design of the German toll system – evaluation and enhancement of simulation performance. In *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 901–909. IEEE, 2012. ISBN 978-1-4673-0708-6.

[18] Tommy Baumann, Bernd Pfitzinger, and Thomas Jestädt. Simulation driven design of the German toll system – profiling simulation performance. In *2013 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 923–926. IEEE, 2013. ISBN 978-1-4673-4471-5.

[19] Tommy Baumann, Bernd Pfitzinger, and Thomas Jestädt. Simulation driven development of the German toll system – simulation performance at the kernel and application level. In *Advances in Business ICT*, volume 257, pages 1–25. Springer International Publishing, 2014. doi: 10.1007/978-3-319-03677-9.

[20] Bernd Pfitzinger, Tommy Baumann, Dragan Macos, and Thomas Jestädt. Using simulations to study the efficiency of update control protocols. *47th Hawaii International Conference on System Sciences (HICSS)*, pages 5154–5161, 2014. doi: 10.1109/HICSS.2014.634.

[21] Pascal Traverse, Isabelle Lacaze, and Jean Souyris. Airbus fly-by-wire: A total approach to dependability. In Renè Jacquart, editor, *Building the Information Society*, volume 156 of *IFIP International Federation for Information Processing*, pages 191–212. Springer US, 2004. ISBN 978-1-4020-8156-9. doi: 10.1007/978-1-4020-8157-6_18.

[22] Robert G. Sargent. Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation*, pages 130–143. Winter Simulation Conference, 2005.

[23] R.G. Sargent. Verification and validation of simulation models. In *Proceedings of the 2010 Winter Simulation Conference (WSC)*, pages 166–183, Dec 2010. doi: 10.1109/WSC.2010.5679166.

[24] Bernd Pfitzinger, Dragan Macos, and Thomas Jestädt. Exploring the heavy goods vehicle fleet behaviour through simulations: Notes from the German toll system. *IET Intelligent Transport Systems*, Aug 2014. ISSN 1751-956X. doi: 10.1049/iet-its.2013.0175.

[25] Michael Hobday, Andrew Davies, and Andrea Prencipe. Systems integration: a core capability of the modern corporation. *Industrial and corporate change*, 14(6):1109–1143, 2005. doi: 10.1093/icc/dth080.

[26] David Janzen and Hossein Saiedian. Test-driven development: Concepts, taxonomy, and future direction. *Computer*, 38:43–50, Sep 2005. ISSN 0018-9162. doi: 10.1109/MC.2005.314.

[27] Azad M. Madni and Michael Sievers. Systems integration: Key perspectives, experiences, and challenges. *Systems Engineering*, 17(1):37–51, 2014. ISSN 1520-6858. doi: 10.1002/sys.21249.

[28] A. Pacholik, T. Baumann, W. Fengler, and D. Grüner. Discrete event simulation performance – benchmarking simulators. In *International Simulation Multi-Conference (SummerSim)*, Genoa, Italy, 2012.

[29] Voon-Yee Vee and Wen-Jing Hsu. Parallel discrete event simulation: A survey.

[30] R. Righter and J.C. Walrand. Distributed simulation of discrete event systems. *Proceedings of the IEEE*, 77(1):99–113, Jan 1989. ISSN 0018-9219. doi: 10.1109/5.21073.

[31] A. J. Wing. Advances in parallel algorithms. chapter Discrete Event Simulation in Parallel, pages 179–226. John Wiley & Sons, Inc., New York, NY, USA, 1992. ISBN 0-470-21907-6.

[32] David Jefferson and Henry A Sowizral. Fast concurrent simulation using the time warp mechanism. 1982.

[33] Richard M. Fujimoto. Parallel discrete event simulation. *Commun. ACM*, 33(10):30–53, October 1990. ISSN 0001-0782. doi: 10.1145/84537.84545.

DOI: 10.15439/2015F161
ACSIS, Vol. 5

# Selected methods of artificial intelligence for Internet of Things conception

Aneta Poniszewska-Maranda
Lodz University of Technology
ul.Wolczanska 215, 90-924 Lodz, Poland
Email: aneta.poniszewska-maranda@p.lodz.pl

Daniel Kaczmarek
Lodz University of Technology
ul.Wolczanska 215, 90-924 Lodz, Poland
Email: dkdaniel@vp.pl

***Abstract*—The concept of Internet of Things appeared several years ago and in that time has evolved into one of pillars of the new technologies sector. The next step is to add the artificial intelligence to Internet of Things systems. Artificial intelligence is increasingly used in everyday life. It is a concept of a wide range and applies in practice in many fields of science. The aim of presented paper was to investigate the usefulness of selected artificial intelligence methods in the concept of Internet of Things. To investigate this purpose, exemplary system was built and it uses the artificial neural networks.**

## I. Introduction

THE Internet is a powerful tool used in all kinds of the information systems. The network is available almost anywhere, at home, at work, also on mobile devices (phones, watches). People start to think to connect the Internet to almost all devices of everyday use, so they can communicate with each other by taking simple decisions for people and helping them in their life. Such idea is called the *Internet of Things (IoT)*. It is estimated that currently about 15 billion devices are connected to the Internet, but this number is still less than 1% of things that in fact could be connected to the network [10].

The next step is to add the artificial intelligence to Internet of Things systems. Artificial intelligence is increasingly used in everyday life. It is a concept of a wide range and applies in practice to many fields of science. It is used in applications such as prompting videos to watch, having regard to the history of the watch (service *netflix*) or recognize people on the recordings of the monitoring. Its great advantage is the elements related to machine learning, through which different methods of artificial intelligence are able to interpret a lot of data and present some of their summary. This is definitely a big amenity for a man who no longer has to statically analyse all the data coming from the specified system, for example view a recording of the monitoring in the context of searching for a particular person.

The presented paper examines whether the chosen methods of artificial intelligence are suitable for use in the concept of *Internet of Things*. The main assumption is to use the mobile device (mobile phone) as a smart object in the Internet of Things. The system based on the concept of *Internet of Things* was designed and built to examine this issue and this system was then implemented with different methods of artificial intelligence.

The presented paper is structured as follows: section 2 presents the issues of Internet of Things concept with its definitions, section 3 gives the outline of the architecture of IoT, presenting the three-layer and five-layer architectures, while section 4 deals with the aspect of communication in IoT concept. Section 5 describes the use of selected methods of artificial intelligence in the conception of IoT and section 6 presents the created exemplary IoT system using such methods.

## II. Concept of Internet of Things

The concept of the Internet of Things appeared several years ago and in that time has evolved into one of the pillars of new technologies sector. There is no clear definition of this concept in the literature. In many cases, the definitions are complementary, creating a more accurate description of the problem.

The Internet of Things is a vision, in which objects become part of the Internet, where every object is uniquely identifiable and accessible on the Web. These objects may directly or indirectly collect, process or exchange data via data communications network. This concept can be described by a simplified equation [1]:

$$\textit{physical objects} + \textit{sensors and microprocessors} = \textit{IoT}$$

One of the first attempts to define this concept was pretty simple concept. The Internet of Things is all objects in everyday life, which are equipped with wireless identifiers and so that they can communicate with each other and be managed by a computer [3].

The shortest definition, which may describe the concept of the Internet of Things shows that it is clearly identification of objects and their representation in the structure of the Internet. This is the most general definition that may expresses the most important elements concerning the issue of IoT [4].

By analysing this information, it can be concluded that the IoT systems have a lot of common issues with multi-agent systems. However, it is worth saying that these two concepts are not identical and have some subtle differences. One of the important difference is a fact that all objects in the concept of the Internet of Things should be able to communicate with the environment, be aware of what they are and be clearly identifiable in their environment [5].

978-83-60810-66-8/$25.00©2015, IEEE

On the basis of all above definitions of the issues of IoT and taking into account its important aspects, the concise and yet comprehensive definition of this concept was created. Thus,

*Internet of Things is a concept, where clearly identifiable and smart objects can communicate with each other in a defined environment to make autonomous decisions by analysing and processing the data collected from the environment.*

The environment can be the Internet or only a portion (e.g. local area network and the devices used only at home). Additionally, there may be mentioned also that the interconnected devices provide the user with various number of applications and services which enable it to communicate with them. To take the autonomous decisions, except to analyse the collected information, objects may use generally understood knowledge, that is, for example, elements related to the habit of the user in some aspect, so that the object will be able to make the better decisions.

## III. Architecture of Internet of Thing

Architecture and technology for smart objects, included into Internet in accordance with the concept of Internet of Things must be extremely flexible. They have to provide the infrastructure, taking into account the heterogeneity of devices and the need for ubiquitous communication, which must be continuous. Just as in the definition of the concept of IoT, specifying its architecture can not clearly define how it should look like. There are many general proposals for IoT infrastructure, but not all of them are sufficiently flexible.

To start talking about specific aspects concerning system architecture based on the IoT, we need first to mention the main pillars of this concept, which will be the basis for the design of adequate infrastructure, namely (Fig. 1):

- anytime,
- anyplace,
- anything.

Taking into account the main pillars and intelligent objects features posted on the Internet, the main aspects that the IoT architecture must meet can be presented. This means for example, that every item included in the Internet of Things (considering its first feature) has a unique identifier in the particular environment in which it occurs and it is available anytime, anywhere and with anything (for all other devices in the structure). Similarly, we can develop two more features of intelligent objects in conjunction with the main pillars of the IoT. This analysis makes us aware that the provision of adequate infrastructure in accordance with the concept of the Internet of Things is not a trivial issue [1], [2], [6].

Implementation of systems based on the concept of Internet of Things requires a comprehensive look at the infrastructure, from the lowest layers of collecting data from the sensors to the highest layers presenting the user different statements of analysed data.

### A. 3-layer architecture

The basic architecture which can be successfully used in such systems is called 3-layer architecture of IoT. It consists of the following layers, starting from the lowest [7], [8]:

- perception layer,
- transport layer (network layer),
- application layer.

*Perception layer* is the first layer in a three-tier architecture of IoT. It is responsible for collecting the data from the real world. Its role is to combine the real world with the virtual world in the context of collection and pre-processing of information. It includes various types of sensors or electronic tags. Most of the modern smartphones are equipped with various sensors, such as ambient light sensor, accelerometer, gyroscope and proximity sensor. This means that the mobile phone can be successfully used in the smart object of IoT, which additionally can pre-process the collected data so that they will be readable for the rest of the system [9], [7].

The second layer, *transport layer (network)*, of this architecture provides the processing of data from sensors, some local storage and forwarding them to the application layer. This layer can be compared to a neural network located in the brain. Since it depends which way and where (to which device) the information will be provided. One of its main tasks is to ensure the effective reach of information and taking care of continuous operations. The network layer is based on modern communication technologies, wired and wireless.

The last layer of this architecture, *application layer*, is the most extensive. It ensures the delivery of services and applications for the end users. It may seem that the important element in this layer is to ensure a user-friendly interaction with the interface. However, this is a small part of the responsibility of the layer. Its fundamental role is to provide a platform for data collection and subsequent analysis. It is understood as its appropriate interpretation, processing and establishing the relationships.

### B. 5-layer architecture

The 3-layer architecture is the fundamental architecture of the Internet of Things. It describes all the necessary infrastructure elements used in the IoT, however, by the fact that there are relatively few layers, each of which is responsible for many aspects. In consequence, it should be more extract components attributable to the particular layer. Such elements include a 5-layer architecture. Its use increases the flexibility of the entire infrastructure.

The 5-layer architecture of IoT introduces two additional layers between the existing layers of 3-layer architecture. Starting from the lowest layer, this architecture has the following (Fig. 2) [9]:

- perception layer,
- access to gateway layer,
- network layer,
- middleware layer,
- application layer.

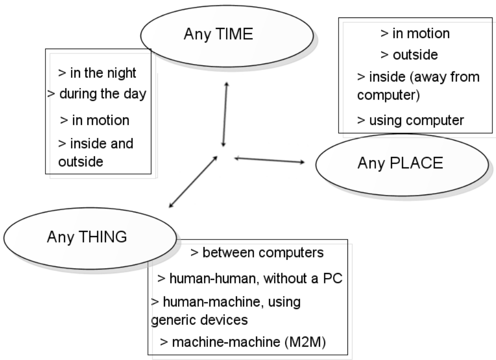


Fig. 1. Schema presenting the main pillars of IoT concept for its communication

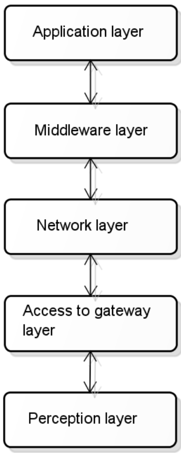


Fig. 2. Schema presenting the 5-layer architecture of IoT

The first layer, *perception layer*, just is primarily responsible for the collection of data from the environment. Another, a new *access to gateway layer* supports the *communication network layer*. Both these two layers are responsible for management of data in exchange infrastructure. The main idea of adding a new layer is to transfer the data directly between the intelligent objects.

With this approach, we can definitely increase the flexibility of the system due to the ability to isolate the communication between devices and services. This second layer is a communication bus between objects and network gateway.

Another newly-added layer is called *middleware layer*. The main purpose of this layer is to increase the flexibility of the interface between the hardware and the software. Information coming from network layer is transmitted to the middleware layer, which is a certain abstraction from both the upper and lower layers. It can be compared to API in the context of the issue of appropriate interfaces for application layer in data management. This approach makes it easier to make the modifications in the whole computer system [9].

The presented architectures form a whole and their mutual application (including joining) may successfully allow to build the complex systems based on the IoT.

## IV. Communication between Internet of Things elements

Today, the concept of the Internet of Things is a component of many different technologies. From a logical point of view, the systems based on the issue of IoT can be seen as a set of smart devices that work together to achieve a set goal. From a technical point of view, such system may consist of several modules, each of which can be implemented in different technologies, uses different architecture and thus uses different items related to the processing of data and even the internal communication [1], [11].

The concept of Internet of Things can be divided into three main elements, relating to the data that is transmitted in the system:

- data capture,
- aggregation of data,
- analysis and data processing.

The first element in the general concept of topology of IoT is *data capturing*. Sensors can be treated as separate components or as part of IoT devices. One device can have more than one sensor. Data obtained from the environment and pre-processed by the device can be exchanged between them in order to achieve a common goal. At this stage, the first communication occurs within the same IoT devices. It can be achieved by wireless or wired technologies.

Once all the required elements at the device level are achieved (data collection and preliminary processing), the *aggregation of data* starts in the IoT and the data is sent to the gateway, where later via the Internet it is delivered to different data centres. Then data is properly *analyzed, processed and presented* to the user with the use of appropriate applications. This processing may take place in specialized servers. It should be added that the prepared data is usually sent to the device again, so it can respond accordingly. Of course, the user can influence the IoT devices in the context of their actions (configurations).

The concept of the Internet of Things has great potential and can make life easier in many aspects. However, the creation of very large IoT systems may encounter many problems, which may be called the challenges which need to be addressed. The main challenge is the heterogeneity of devices. Since IoT vision involves connecting almost all devices to the Web, it has to do with the large variation mainly due to the computational and communication capabilities. In addition, this involves a huge amount of data collected from the sensors. We can talk here about the big data, that is for large and diverse data sets, where processing and analysis are difficult, but valuable, because they can lead to acquiring the new knowledge. The main element that can provide support for this challenge is well-chosen architecture and protocols [9].

Equally important challenges are still the elements associated with *location*, *self-configuration* and *data management*. As each device in IoT is uniquely identifiable, it is possible to locate it and its location will be used as a additional functionality. However, there is a problem with the confidentiality. Due to the number and complexity of systems, an important issue is the self-configuration, to relieve a man. It is also a challenge, since a multitude of configurations can lead to the fact that people will not understand how some elements of the system have to behaviour in the context of the property, which may prove to be as dangerous as the lack of data security. The exchange and analysis of data are important in IoT. Therefore, in addition to their safety, it is important to ensure the proper data models and description of their contents, to use them as a sure knowledge.

## V. Internet of Things systems using artificial intelligence

Systems built using the concept of Internet of Things are based not only on the simple sensors that transmit information to the systems, that operate primarily on the basis of statistics and simple mathematical calculations. Such systems are increasingly complex and can make the decisions in bigger number of aspects. It is easy to imagine a system that switches the heating only based on the ambient temperature. However, it could also regulate the temperature in terms of the number of people present in the room, the habits of certain users (individuality), the specific rooms and time of a day. Therefore, to give some intelligence to these systems is an important issue, but rather complex.

Previously mentioned example with temperature and habits of users shows that IoT systems should learn the habits and adapt to them (teaching) in this case. Such elements are not achieved by ordinary statistics or simple equations. In this case the system needs more sophisticated tools, such as artificial intelligence methods.

The idea of the use of artificial intelligence in the Internet of Things is associated with another issue. It is the fact of independence of the machines in the context of their supervision. Application of AI methods can affect in a positive way to save the time. It is important not to lose a control over the device completely, but equally important is the lack of monotonous supervision of these applications from the point of use. It is better to be able to communicate with the system in a way natural for humans than for machines – based on the example with temperature – while overheating of the room, it's better to make the interaction with the device using command "is too hot" than reduce the temperature of 0.432 Celsius degree.

The main element associated with the operation of IoT system with the artificial intelligence is its location in the architecture. An important aspect is the performance and the appropriate amount of place for data, which is a knowledge of the system, so the AI methods can not be placed at each level. Figure 3 shows the general idea of placing the artificial intelligence methods in the context of major IoT architectures.

The most natural places for AI methods are all kinds of servers because of their computing power. Such location has a positive effect on another aspect too. It is about the fact of reuse. The place of operation of different artificial intelligence methods can be compared to the human brain. The fact that all the knowledge and the associated inference and learning are placed in the server rooms, it is possible to use it in a larger perspective.

Figure 4 presents our idea of general scheme of information flow in the IoT systems, using artificial intelligence. There are three main elements in the flow of information:

- Preliminary communication – data sent from the real world for a variety of systems.
- Context communication – data processed by systems that already have the appropriate context and make the IoT systems and devices can respond accordingly (take appropriate decisions).
- Internal communication – understood as an additional channel of communication between intelligent objects.

The first stage can be called a *preliminary communication*. At the beginning, in accordance with the concept of Internet of Things, the data from the environment (real world) is collected by the IoT devices. They can be the external sensors, as well as those built into the device. Then, data is preprocessed to be clear for the rest of the system. Depending on the particular case, the data is further transmitted over the Internet to the main IoT systems or between other devices to gather all relevant information from the real world (internal communication elements). In the context of mobile devices, it is important to assure the temporary storage of information in things during the lack of access to the Web. When all the data in the context of a particular cycle, is already in the main IoT systems, respectively, it is prepared for external artificial intelligence systems. This process may involve selecting an appropriate specific information to get to a particular method of artificial intelligence (for example the ambient temperature and the number of people as inputs to the neural network in the system controlling the temperature).

The next communication step is a *context communication*. Its main purpose is to provide the concrete system answers to the IoT devices and to the subsystems that are designed to take concrete actions in the context of the relevant decisions. When the artificial intelligence methods have exited and give a reply, it should be properly interpreted. This task should address the

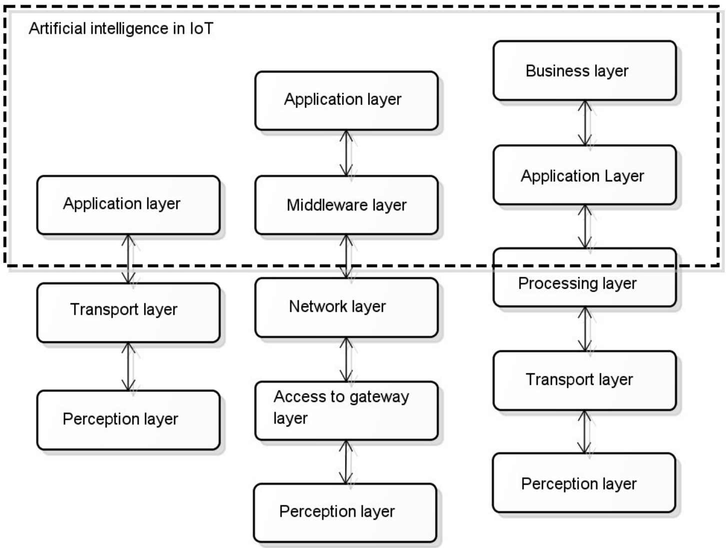


Fig. 3. Placing the artificial intelligence in the background of IoT architectures

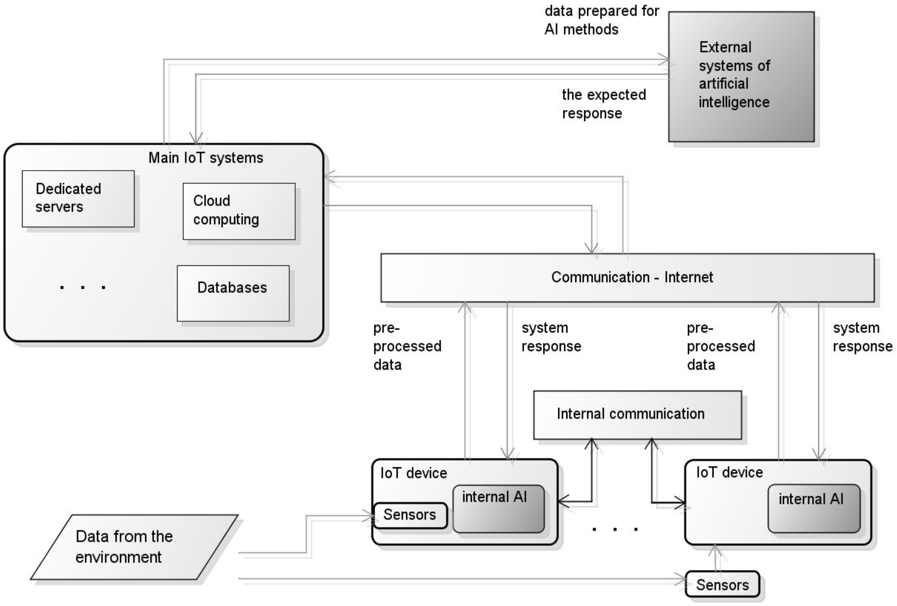


Fig. 4. General scheme of information flow in IoT systems with the use of artificial intelligence methods

main IoT systems. Very often the AI answer include digits that do not make much sense without the proper context. Therefore, an important step is to link them with the knowledge located in the system to be able to conclude the overall response of the system to a particular problem. In this way the system can properly react, so take the decision. This idea can be seen as smart making of autonomous decision.

Properly processed data, which are located in the main IoT systems should usually be further sent to certain IoT devices so that they can properly respond. A situation where the processed data must be sent to other IoT devices than those which are derived from the original data can be imagined. The monitoring system can be an example. Devices with cameras may send data to the main IoT systems, that after detecting of a specific threat (AI systems use object recognition) can send data to the other Internet of Things devices which may in some way respond to such a threat.

The last extra communication is called *internal communication*. It can be placed both at the initial communication, as well as at the contextual one in the proposed model. In the first case it can be used to gather enough data from multiple devices and simultaneously to send them to the upper layers of the system. In the second case, having a concrete answer from the system, it can be used the multiple devices work simultaneously together. This approach is combined with additional AI elements contained in the devices – it can enter into intelligent objects more aspects of their autonomy.

## VI. IoT system with the use of artificial intelligence

To investigate the actual usefulness of selected methods of artificial intelligence in the concept of the Internet of Things, the information system *smart-IoT* was created. Its main task is to test the time to return home and set the alarms at the right time to put the event. The proposed system consists of three main elements:

- mobile devices – smart objects,
- central server – acts as an IoT management system,
- micro-services – include elements of AI.

The system is based on the 4-layer architecture:

- Perception layer – covers only mobile devices. Its main task is to collect the data from the environment using sensors embedded in mobile phones (GPS, accelerometer, gyroscope) and pre-treatment of this data.
- Network layer – is responsible for transferring data between all the elements of the system using LAN and Wi-Fi communications channels.
- Processing layer – an essential element of the whole system. This layer is responsible for the processing of information in a central server with the appropriate use of micro-services, so the methods of artificial intelligence. Its primary aim is to convert the input data and generate a specific answer for that data in the context of the service.
- Application layer – is responsible for providing data to the end user. Its fundamental role is to ensure the validity of data (to be updated with information from the decision-making module).

Intelligent objects in the proposed system are mobile devices (specifically mobile phones). Their main task is to collect the relevant data from the real world by using the built-in sensors. The next stage is the initial processing of the data. All of these tasks are performed by using the dedicated mobile application.

In accordance with the concept of Internet of Things, all devices must be uniquely identifiable throughout the system. For this reason, the *smart-IoT* system proposed the use of unambiguous numeric identifiers. Each mobile device has a given unique identifier, which is recorded in the central management server. Such an approach also allows to limit the access of unauthorized devices, because the system will support only those for which the information is stored in the database.

The main components of IoT, understood as the main systems of the Internet of Things, are represented by a central server, acting as a manager. It consists of 4 major components:

- module of services identification,
- module of artificial intelligence management,
- decision-making module,
- database module.

## VII. Conclusions

The Internet of Things is a technological revolution that represents the future of computing and communication. This concept is characterized by heterogeneous technologies and devices and assumes that all devices will be connected to the Internet. The next step is to add the artificial intelligence to IoT systems. Thanks to this, devices become intelligent and can make autonomous decisions. These smart devices have the ability to interact with humans and other smart devices.

These devices should have a certain autonomy in the context of decision-making process. In building of IoT systems an important element is its architecture and its scalability and flexibility. Their key action aspect is the exchange and analysis of data. Joining of this type of systems with artificial intelligence is not a trivial task. AI methods typically use a lot of processing power, therefore, using them directly in devices often becomes impossible. They are usually placed on external servers, so a user can use them in the context of multiple devices at the same time.

The presented paper investigated the usefulness of artificial intelligence in the concept of Internet of Things. To do this the exemplary system was built and it uses the artificial neural networks. This system uses mobile devices as the smart objects. Neural networks have been taught by back-propagation algorithm. Experiments with neural networks were carried out by using the main services of the system (such as auto set alarms for a specific event and estimating the time to return home). These experiments show that artificial intelligence methods are suitable for use in concept of Internet of Things.

## References

[1] A. McEwen and H. Cassimally, "Designing the Internet of Things", Wiley, 2014,

[2] F. da Costa, "Rethinking the Internet of Things. A Scalable Approach to Connecting Everything", Apress open, California, 2013

[3] An Introduction to the Internet of Things (IoT), http://www.lopezresearch.com/

[4] A. Arsénio and H. Serra and R. Francisco and F. Nabais and J. Andrade and E. Serranol, "Internet of Intelligent Things: Bringing Artificial Intelligence into Things and Communication Networks", Springer Science, 2014

[5] M. Ruggieri and H. Nikookar and O. Vermesan and P. Friess, "Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems", River Publishers, 2013

[6] D. Uckelmann and M. Harrison and F. Michahelles, "Architecting the Internet of Things", Springer, 2011

[7] M. Wu and T.-L. Lu and F.-Y. Ling and L. Sun and H.-Y. Du, "Research on the architecture of Internet of things", 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010

[8] N. Lin and W. Shi, "The Research on Internet of Things Application Architecture Based on Web", IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA), 2014

[9] C-W. Tsai and C.-F. Lai and A. V. Vasilakos, "Future Internet of Things: open issues and challenge", Springer Science, New York, 2014

[10] Cisco, http://www.cisco.com/web/about/ac79/docs/innov/IoE.pdf

[11] S. Sicari and A. Rizzardi and L. A. Grieco and A. Coen-Porisini, "Security, privacy and trust in Internet of Things: The road ahead", Computer Networks, 2014

[12] P. Lynggaard, "Artificial intelligence and Internet of Things in a "smart home" context: A Distributed System Architecture", PhD dissertation, Aalborg University Copenhagen, 2013

DOI: 10.15439/2015F230
ACSIS, Vol. 5

# Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction

Michał Skuza

Email: michalskuza@hotmail.com

Andrzej Romanowski

androm@kis.p.lodz.pl

Lodz University of Technology, Institute of Applied Computer Science, Poland.

*Abstract*— **This paper covers design, implementation and evaluation of a system that may be used to predict future stock prices basing on analysis of data from social media services. The authors took advantage of large datasets available from Twitter micro blogging platform and widely available stock market records. Data was collected during three months and processed for further analysis. Machine learning was employed to conduct sentiment classification of data coming from social networks in order to estimate future stock prices. Calculations were performed in distributed environment according to Map Reduce programming model. Evaluation and discussion of results of predictions for different time intervals and input datasets proved efficiency of chosen approach is discussed here.**

***Keywords***: Sentiment Analysis, Big Data Processing, Social Networks Analysis, Stock Market Prediction.

## I. Introduction & Rationale

It is believed that information is the source of power. Recent years have shown not only an explosion of data, but also widespread attempts to analyse it for practical reasons. Computer systems operate on data measured in terabytes or even petabytes and both users and computer systems at rapid pace constantly generate the data. Scientists and computer engineers have created special term "big data" to name this trend. Main features of big data are volume, variety and velocity. Volume stands for large sizes, which cannot be easily processed with traditional database systems and single machines. Velocity means that data is constantly created at a fast rate and variety corresponds to different forms such as text, images and videos.

There are several reasons of a rise of big data. One of them is the increasing number of mobile devices such as smartphones, tablets and computer laptops all connected to the Internet. It allows millions of people to use web applications and services that create massive amounts of logs of activity, which in turn are gathered and processed by companies. Another reason is that computer systems started to be used in many sectors of the economy from governments and local authorities to health care to financial sector. The analyses of information that is a by-product of different business activities by companies can lead to better understanding the needs of their customers and prediction future trends. It was previously reported in several research papers that precise analysis of trends could be used to predict financial markets [1]

Big size of data and the fact it is generally not well structured result in situation that conventional database systems and analysis tools are not efficient enough to handle it. In order to tackle this problem several new techniques ranging from in-memory databases to new computing paradigms were created.

Besides big size, the analysis and interpretation are of main concern and application for big data perspective stakeholders. Analysis of data, also known as data mining, can be performed with different techniques such as machine learning, artificial intelligence and statistics. And again it is important to take into consideration the size of data to be processed that in turn determines if a given existing algorithm or approach is applicable.

### A. Big data

There are several definitions what Big data is, one of them is following: "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse." [2] This definition emphasizes key aspects of big data that are volume, velocity and variety [3]. According to IBM reports [4] everyday "2.5 quintillion bytes of data" is created. These figures are increasing each year. This is due to previously described ubiquitous access to the Internet and growing number of devices. Data is created and delivered from various systems operating in real-time. For example social media platforms aggregate constantly information about user activities and interactions e.g. one of most popular social sites Facebook has over 618 million daily active users [5]. Output rate of the system can be also important when nearly real-time analyses are needed. Such an on-the-fly analysis is required in recommendations systems when the user's input affects content provided by web site; a good examples are online retail platforms such as Amazon.com. This aspect requires various ways of storing the data to maximize speed and sometimes using column-

978-83-60810-66-8/$25.00©2015, IEEE

oriented database or one of schema-less systems (NoSQL) can do the job, since big data is rarely well structured

But big data is not only challenging but primarily creates opportunities. They are, among the others: creating transparency, optimization and improving performance, generation of additional profits and nothing else than discovering new ideas, services and products.

### *B. Social media*

One of the trends leading to rise of big data is Web 2.0. It is a major shift from static websites to interactive ones with user-generated content (UGC). Popularization of Web 2.0 resulted in many services such as blogging, podcasting, social networking and bookmarking. Users can create and share information within open or closed communities and by that contributes to volumes of big data.

Web 2.0 led to creation of social media that now are means of creating, contributing and exchanging information with others within communities by electronic media. Social media can be also summarized as "built on three key elements: content, communities and Web 2.0" [6]. Each of those elements is a key factor and is necessary for social media. One of the most important factors boosting social media is increasing number of always Internet-connected mobile devices such as smartphones and tablets.

Twitter is a micro blogging platform, which combines features of blogs and social networks services. Twitter was established in 2006 and experienced rapid growth of users in the first years of operations. Currently it has over 500 million registered users and over 200 million active monthly users [7]. Registered users can post and read messages called "tweets"; each up to 140 Unicode characters long – originated from SMS carrier limit. Unregistered users can only view tweets. Users can establish only follow or be-followed relationships. A person who subscribes to other user is referred as "follower" and receives real-time updates from that person. However users do not have to add people who are their followers. Twitter can be accessed from various services such as official Twitter web page, mobile applications from third parties and SMS service. As Twitter is an extremely widespread service, especially in US and as the data structure is compact so it forces users to post short comments authors of this paper believe this is a good source of information in the sense of snapshots of moods and feelings as well as for up-to-date events and current situation commenting. Moreover, Twitter is a common PR communication tool for politicians and other VIPs shaping, or having impact on the culture and society of large communities of people. Therefore Twitter was chosen for experimental data source for this work on predicting stock market.

## II. Predicting Future Stock Prices

### *A. Experimental System Design and Implementation*

Main goal of this section is to describe implementation of a system predicting future stock prices basing on opinion detection of messages from Twitter micro blogging platform. Unlike the authors of [12] we chose Apple Inc. – a well known consumer electronics company – a producer of Mac computers, iPod, iPad, iPhone products and provider of related software platforms and online services just to name a few.

System design is presented on Figure 1 and it consists of four components: Retrieving Twitter data, pre-processing and saving to database (1), stock data retrieval (2), model building (3) and predicting future stock prices (4). Each component is described later in this text.

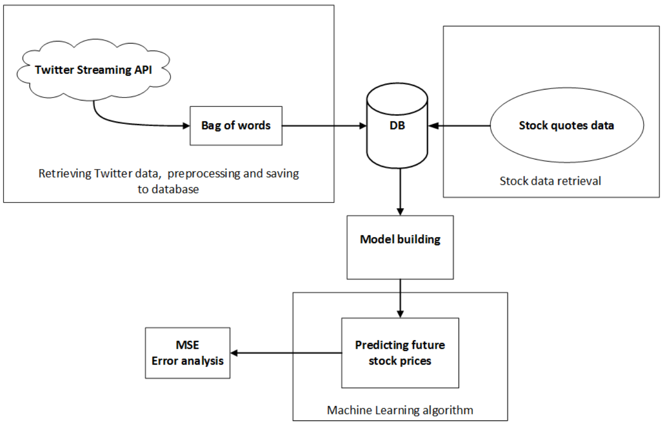

Twitter Streaming API
Bag of words
DB
Stock quotes data
Retrieving Twitter data, preprocessing and saving to database
Stock data retrieval
Model building
MSE Error analysis
Predicting future stock prices
Machine Learning algorithm

Figure 1: Design of the system

1. Retrieving Twitter data, pre-processing and saving to database.

This component is responsible for retrieving, pre-processing data and preparing training set. There are two labelling methods used for building training set: manual and automatic.

2. Stock data retrieval

Stock data is gathered on a per minute basis. Afterwards it is used for estimating future prices. Estimation is based on classification of tweets (using sentiment analysis) and comparing with actual value by using Mean Squared Error (MSE) measure.

3. Model building.

This component is responsible for training a binary classifiers used for sentiment detection.

4. Predicting future stock prices

This component combines results of sentiment detection of tweets with past intraday stock data to estimate future stock values.

### *B. Twitter data acquisition and pre-processing*

Twitter messages are retrieved in real time using Twitter Streaming API. Streaming API allows retrieving tweets in quasi-real time (server delays have to be taken into consideration). There are no strict rate limit restrictions, however only a portion of requested tweets is delivered. Streaming API requires a persistent HTTP connection and authentication. While the connection is kept alive, messages are posted to the client. Streaming API offers possibility of filtering tweets according to several categories such as location, language, hashtags or words in tweets. One disadvantage of using Streaming API is that it is impossible to retrieve tweets from the past this way.

Tweets were collected over 3 months period from 2nd January 2013 to 31st March 2013. It was specified in the query that tweets have to contain name of the company or hashtag of that name. For example in case of tweets about Facebook Inc. following words were used in query ’Apple’, ’#Apple’, ’AAPL’ (stock symbol of the company) and ’#AAPL’. Tweets were retrieved mostly for Apple Inc. (traded as ’AAPL’) in order to ensure that datasets would be sufficiently large for classifications. Retrieved data contains large amounts of noise and it is not directly suitable for building classification model and then for sentiment detection. In order to clean twitter messages a program in Python programming language was written. During processing data procedure following steps were taken. Language detection information about language of the tweet is not always correct. Only tweets in English are used in this research work. Duplicate removal - Twitter allows to repost messages. Reposted messages are called retweets. From 15% to 35% of posts in datasets were retweets. Reposted messages are redundant for classification and were deleted. After pre-processing each message was saved as bag of words model – a standard technique of simplified information representation used in information retrieval.

### *C. Sentiment Analysis*

Unlike classical methods for forecasting macroeconomic quantities [13,15,16] prediction of future stock prices is performed here by combining results of sentiment classification of tweets and stock prices from a past interval. Sentiment analysis [8,14] - also known as opinion mining refers to a process of extracting information about subjectivity from a textual input. In order to achieve this it combines techniques from natural language processing and textual analysis. Capabilities of sentiment mining allow determining whether given textual input is objective or subjective. Polarity mining is a part of sentiment in which input is classified either as positive or negative.

In order to perform a sentiment analysis classification a model has to be constructed by providing training and test datasets. One way of preparing these datasets is to perform automatic sentiment detection of messages. This approach was used in several works such as [9]. Another possibility of creating training and test data is to manually determine sentiment of messages, which means it is a standard, supervised learning approach. Taking into consideration large volumes of data to be classified and the fact they are textual, Naïve Bayes method was chosen due to its fast training process even with large volumes of training data and the fact that is it is incremental. Considered large volumes of data resulted also in decision to apply a map reduce version of Naïve Bayes algorithm. In order to perform sentiment analysis on prepared bags of words a model has to be constructed by providing training and test datasets for classification. These datasets were created using two different methods. One was applying an automatic sentiment detection of messages. It was achieved by employing SentiWordNet [10] which is a publicly available resource aimed to support performing sentiment and opinion classifications. The other method was a manual labelling of sentiment of tweets. Each message was marked as positive, negative or neutral. There were two training datasets. First one consisted of containing of 800 hundred tweets. The other dataset consisted of 2.5 million messages. Only 90% of each dataset was used directly as a training set the other 10% was used for testing.

As a result of two classifiers were obtained using manually labelled dataset. First classifier determines subjectivity of tweets. Then polarity classifier classifies subjective tweets, i.e. using only positive and negative and omitting neutral ones. In order to use classification result for stock prediction term: ’sentiment value’ (denoted as a **ε**) was introduced - it is a logarithm at base 10 of a ratio of positive to negative tweets (Eq. 1).

$$\varepsilon = \log_{10} \frac{number\_of\_positive\_tweets}{number\_of\_negative\_tweets} \quad (1)$$

If **ε** is positive then it is expected that a stock price is going to rise. In case of negative **ε** it indicates probable price drop. In order to estimate price of stock, classification results are combined with a linear regression of past prices where one weight is a sentiment value. Predictions for a specific time point are based on analysis of tweets and stock prices from a past interval - Eq. 2 shows the formula for the relationships of past value of stock taken into analysis.

$$y_i = \alpha + (\beta + \varepsilon_i)x_i \quad (2)$$

where: $y_i$ is a past value of a stock at a given time of $x_i$, $x_i$ is time variable, $\varepsilon_i$ is a sentiment value calculated for a given time of $x_i$, $i = 1, \ldots, n$ , $\beta$ is a linear regression coefficient defined as (Eq. 3)

$$\beta = \frac{\Sigma x_i y_i - \frac{1}{n}\Sigma x_i \Sigma y_i}{\Sigma x_i^2 - \left(\frac{1}{n}\Sigma x_i\right)^2} \quad (3)$$

and $\alpha$ coefficient is given by (Eq. 4):

$$\alpha = \overline{p} - \beta\overline{t} \quad (4)$$

where $\overline{p}$ and $\overline{t}$ are mean values of price of stock over a period of $t$.

## III. Results and Analysis

Predictions were prepared using two datasets for several different time intervals, i.e. time differences between the moment of preparing the prediction and the time point for which the forecast was prepared. Predictions were conducted for one hour, half an hour, 15 minutes and 5 minutes ahead of the moment being forecasted. Two tweet datasets were used: one with messages containing company stock symbol 'AAPL', and the other dataset included only tweets containing name of the company, i.e. 'Apple'. Training datasets consisted of 3 million tweets with stock symbols and 15 million tweets with company name accordingly. Tweets used for predictions were retrieved from 2nd to 12th of April 2013. Approximately 300 000 tweets were downloaded during New York Stock Exchange trading hours each day via Twitter Streaming API.

Experiments were conducted using two models of classifiers, first was built using manually labelled, dataset-based trained classifier and the other was trained with automatically labelled tweet training datasets. Experiments were conducted in the following manner. For each of the following prediction time intervals: 1 hour, 30 min and 15 min all four models (permutations of manual or auto-classifiers coupled with AAPL or Apple keywords) were used. For 5 minutes prediction small number of tweets with stock symbol (AAPL) per time interval resulted in limiting predictions only to models trained with messages with company name (Apple).

Time axis shown on Fig 2 and following figures shows NASDAQ trading hours converted to CEST 1 time zone. Results of predictions are also compared to actual stock prices using Mean Square Error (MSE) measure that are presented in table 1. Sample predictions are presented for 1 hour, 30 minutes, 15 minutes and 5 minutes, while following figures are presented only for 1 hour, 30 min and 5 min.

TABLE 1: MEAN SQUARE ERROR VALUES OF PREDICTED AND ACTUAL STOCK PRICES.

| MSE | 1 Hour | 30 Min | 15 Min | 5 Min |
|---|---|---|---|---|
| Manual & 'AAPL' | 1.5373 | 0.6325 | 0.3425 | - |
| Auto & 'AAPL' | 0.947 | 0.3698 | 0.2814 | - |
| Manual & 'Apple' | 1.9287 | 1.5152 | 0.9052 | 0.5764 |
| Auto & 'Apple' | 1.8475 | 1.4549 | 0.8325 | 0.3784 |

### One-hour predictions

A first objective was to test a performance of predictions for 1-hour intervals. Example results of 1-hour interval predictions are presented on Fig. 2.

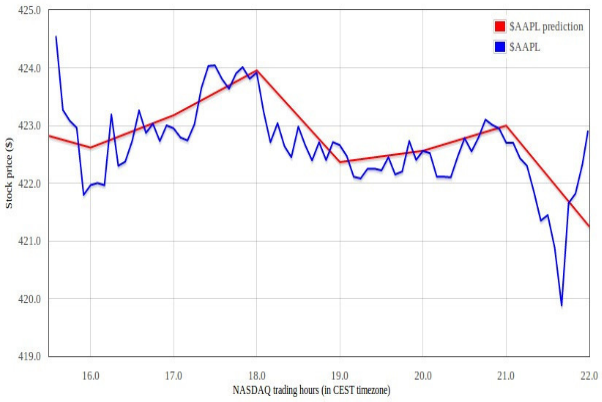


Figure 2: One-hour prediction. Manually labelled 'AAPL' training dataset.

Blue line (fluctuated) corresponds to actual stock prices and red line (steadily changing) shows predicted values. Predictions in this time intervals would not provide accurate result but they can be used to evaluate if the method correctly estimates trends. Predictions for the same day are presented below using 4 different models of classifier.

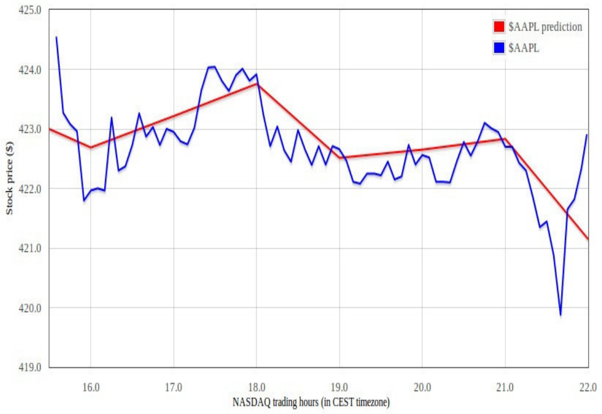


Figure 3: One-hour prediction. Automatically labelled 'AAPL' training dataset.

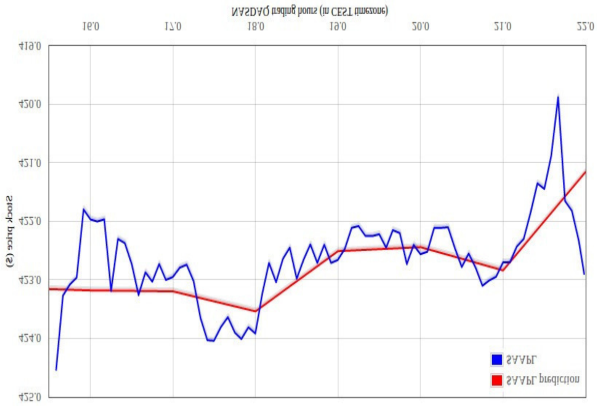


Figure 4: One-hour prediction. Manually labelled 'Apple' training dataset.

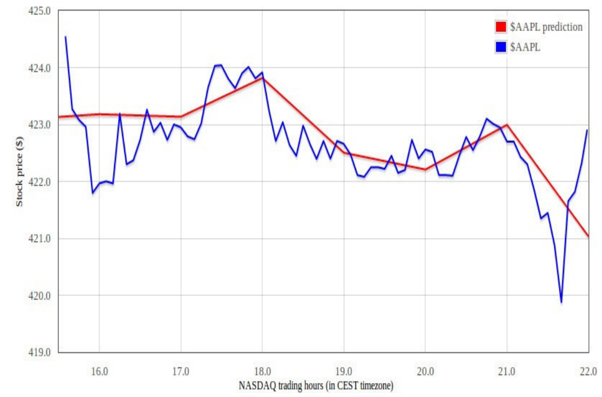


Figure 5: One-hour prediction. Automatically labelled 'Apple' training dataset.

As it can be observed on Fig. 2-5 predictions of each model are similar. They correctly forecast trends however due to big time interval, i.e. 1 hour it is not possible to determine whether models can predict sudden price movements. Furthermore when comparing results for two datasets, predictions using models trained with tweets with stock symbol perform better that those trained with tweets containing company name.

### 30 minutes predictions

This subsection describes 30 minutes predictions. It is expected for these predictions not only to forecast trends but sudden price movements as well; this expectation is due to the fact of smaller time interval between time of prediction and forecasted moment. Results are shown on Fig. 6-7 graphs. First model (‘manual’ for AAPL keyword) predictions are less accurate in comparison with second model (auto/AAPL). Yet, significant difference between predicted and actual prices from 17 to 21 in both first models is still there (for further analysis no figures for AAPL is shown from this point on since APPLE gives better performance). Models using dataset with actual company name (plots at Fig. 6 and Fig. 7) perform much better than two first ones. Predictions of prices follow actual ones. However in all cases price forecasting is less accurate when there are several dynamic changes of price movement trend. It is especially visible in all figures for periods from 18 to 20 hours that predictions do not show correlations with actual prices. It may result, among the others, from too long time intervals in comparison to rapid price movements.

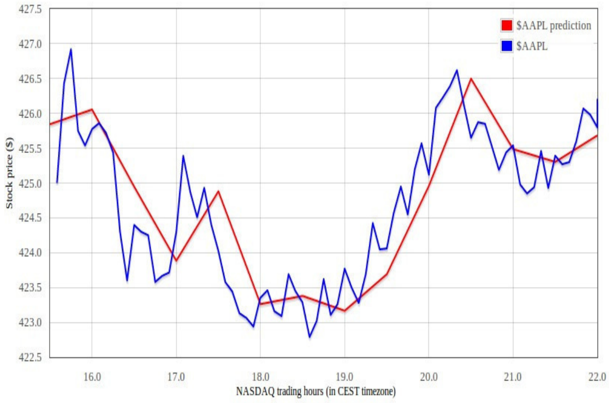


Figure 6: 30 minutes prediction. Manually labelled ’Apple’ training dataset.

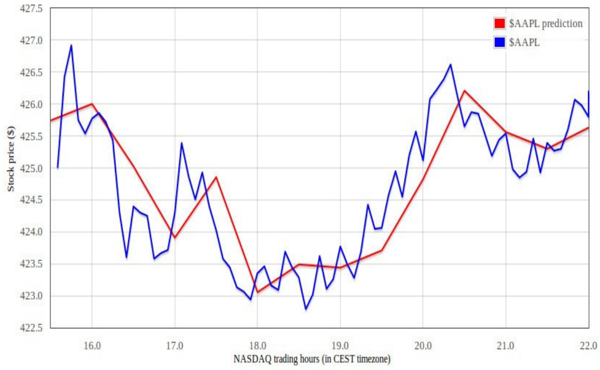


Figure 7: 30 minutes prediction. Automatically labelled ’Apple’ training dataset.

For 15 minutes dynamic price movements are somehow reflected in prediction, although it is not any significant indication in a sense of preserving real nature and amplitude of those fluctuations. Rapid price movements are not easily indicated due to chosen time interval; still too long for better accuracy. Only for this 15 min interval classification based on tweets with stock symbol yield better results. Furthermore it is important to note that using automatically trained training with bigger number of records strongly affects result of prediction.

### 5 minutes predictions

Last experiment was to perform 5 minutes predictions. Due to short time interval it was expected that prediction would the most accurate. In this part only dataset build with messages with actual company name was used. This is because the number of messages from datasets with stock symbol per time interval was very small and the results of predictions were not reliable.  Results are shown on fig. 8-9.

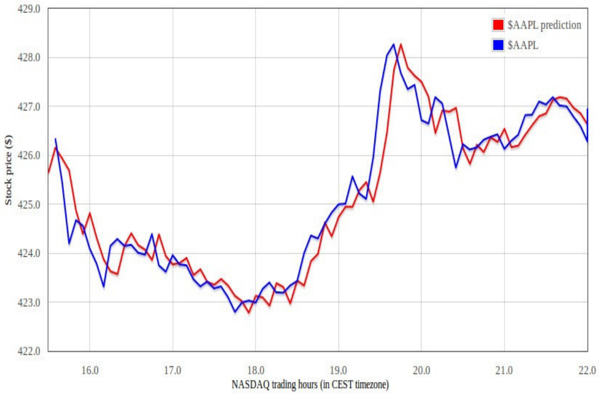


Figure 8: 5 minutes prediction. Manually labelled ’Apple’ training dataset.

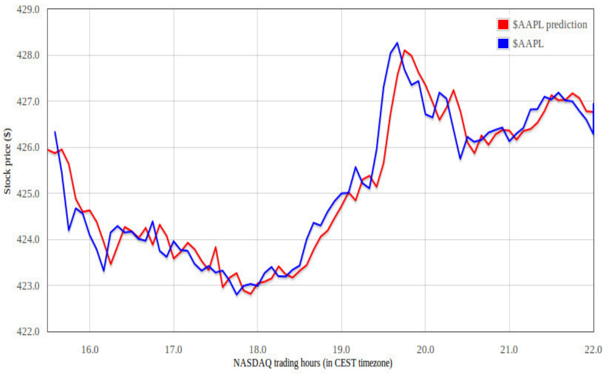


Figure 9: 5 minutes prediction. Automatically labelled ’Apple’ training dataset.

## IV. DISCUSSION OF RESULTS

As it can be observed from presented results, predictions of stock prices depend strongly on choice of training dataset, their preparation methods and number of appearing messages per time interval. Predictions conducted with models trained with datasets with messages containing company stock symbol performs better. It can be explained by the fact that these messages refer to stock market. Tweets with company name may just transfer information, which does not affect financial results. Another important factor is a choice of preparation of training set. Two methods were used. One of the methods was a manual labeling sentiment value of messages. This method allows to more accurately label training data but is not effective for creating large training sets. The other method was applying SentiWordNet, which is a lexical resource for sentiment opinion mining. It enabled to create bigger training datasets, which resulted in building more accurate models. Last factor that is important for prediction is number of appearing messages per time interval. Although model trained with datasets with company name were not accurate in comparison to the other datasets, there is bigger number of tweets per time interval. It allowed performing prediction for shorter time intervals, which were not possible for dataset with messages containing company stock symbol. Described methods can be also used with other stock predictions procedures in order to maintain higher accuracy. It is also important to note that stock prediction methods are not able to predict sudden events called 'black swans' [11].

## V. Conclusions

This paper discusses a possibility of making prediction of stock market basing on classification of data coming Twitter micro blogging platform. Results of prediction, which were presented in previous section show that there is correlations between information in social services and stock market. There are several factors that affect accuracy of stock predictions. First of all choice of datasets is very important. In the paper two types of datasets were used one with name of the company and the other with stock symbol. Predictions were made for Apple Inc. in order to ensure that sufficiently large datasets would be retrieved. There were large differences in size between these two sets. This lead to situation that it was not possible to perform 5 minutes predictions basing on tweets with stock symbol due to too few messages. Additionally although dataset with company name was bigger it may not be accurate for predictions. This due to the fact that company name can be used as a household name and the messages do not refer to stock market. In case of tweets with stock symbol there is bigger probability that people who posted are relating to stock prices. Predictions can be improved by Adding analysis of metadata such as exact location of a person while posting message, number of retweets, number of followers etc. This information may be used to determine which users are more influential and creating a model of interactions between users. Number of messages posted by a user and its frequency may be used to discard spammers and automated Twitter accounts. It is also possible to employ different sources of information. Although Twitter is very popular and offers nearly time communications there exist other sources information such different social networks, blogs, articles in online newspapers. Adding analysis of other may contribute to more accurate predictions.

## References

[1] Z. Da, J. Engelberg, P. Gao: *In Search of Attention*, The Journal of Finance Volume 66, Issue 5, pages 1461–1499, October 2011, doi: 10.1111/j.1540-6261.2011.01679.x

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A.H. Byers. Big data: The next frontier for innovation, competition, and productivity, McKinsey, May 2011.

[3] Edd Dumbill. What is big data? : an introduction to the big data landscape. http://radar.oreilly.com/2012/01/what-is-big-data.html, 2012.

[4] P. Zikopoulos, C.Eaton, D. DeRoos, T. Deutch and G. Lapis, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media, 2011

[5] M. Zajicek. Web 2.0: hype or happiness? In Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A), W4A '07, pages 35–39, New York, NY, USA, 2007. ACM. doi: 10.1145/1243441.1243453

[6] T. Ahlqvist and Valtion teknillinen tutkimuskeskus. Social media roadmaps: exploring the futures triggered by social media. VTT tiedotteita. VTT, 2008.

[7] Twitter Statistics. http://www.statisticbrain.com/twitter-statistics/, 2013. [Online; accessed 2-January-2013].

[8] B. Pang and L. Lee. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2):1–135, January 2008, doi: 10.1561/1500000011

[9] Y-W Seo, J.A. Giampapa, and K. Sycara. Text classification for intelligent portfolio management. Technical Report CMU-RI-TR 02-14, Robotics Institute, Pittsburgh, PA, May 2002.

[10] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06, pages 417–422, 2006. In In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06, pages 417–422, 2006, doi: 10.1155/2015/715730

[11] N.N. Taleb, Common Errors in the Interpretation of the Ideas of The Black Swan and Associated Papers (October 18, 2009)

[12] M. Paluch, L. Jackowska-Strumillo: The influence of using fractal analysis in hybrid MLP model for short-term forecast of closing prices on Warsaw Stock Exchange. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 2, pages 111–118 (2014) doi: 10.15439/2014F358

[13] ·M. Marcellino, J. H. Stock, M.W. Watson, A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series, Journal of Econometrics Volume 135, Issues 1–2, November–December 2006, Pages 499–526 doi:10.1016/j.jeconom.2005.07.020

[14] Asur, S., Huberman, B.A., Predicting the Future with Social Media IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010, pp 492 - 499 doi:10.1109/WI-IAT.2010.63

[15] K-J. Kim, I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, Expert Systems with Applications Volume 19, Issue 2, 2000, Pages 125–132 doi:10.1016/S0957-4174(00)00027-0

[16] E. J. Ruiz, V. Hristidis, C. Castillo, and A. Gionis, "Correlating Financial Time Series with Micro-Blogging activity," WSDM 2012. Doi: 10.1145/2124295.2124358

DOI: 10.15439/2015F51
ACSIS, Vol. 5

# Towards explaining publishing activity in Facebook

Jerzy Surma
Department of Computer Science and Digital Economy
Warsaw School of Economics
Warsaw, Poland
E-mail: jerzy.surma@sgh.waw.pl

Luvai Motiwalla
Operations and Information Systems Department
University of Massachusetts
Lowell, USA
E-mail: luvai_motivalla@uml.edu

***Abstract*—A growing number of users, seeking to influence their friends, are adopting social network sites, like Facebook to increase their social presence. In the relatively few scholarly studies that consider how writing comments on social network sites influences their friends, the focus has been on marketing or promotion. Little is known about the impact of user writing versus reading on social media. In our paper, we address this gap by studying how the writing comments influences the reactive behavior (likes, comments) of their friends. Our research sheds more light on what motivates users on social networks sites to write and publish on their page. Specifically, we study the user-to-user interactions on Facebook. We found that the number of messages the user broadcasts is dependent on the feedback he or she received from friends.**

## I. Introduction

LARGE volumes of status update messages are posted by Facebook (FB) users on their pages to inform their friends and followers about their current activity like eating in restaurant, watching movie or others [1]. They are an increasingly popular form of communication that allow users to interact with their followers who can react with a comment or a ‘‘like’’. Status updates enable quick and effortless one-to-many communication [2]. On average, there are 60 million status updates per day FB making it by far most popular social network site (SNS) and a logical place to study user influence behaviors and consequences of the social processes associated with SNS usage [3]. The popularity and novelty of status updates makes it a very interesting topic for a variety of empirical studies [4].

This paper focuses on understanding users’ write or posting activity, which is crucial for the social network site’s existence. *What reactions are influencing users to publish?* Our research finding shows that feedback given by friends and followers has a critical influence on the writing activity. Data obtained from social networks sites are undoubtedly an important source of knowledge about user behaviors and constitute an important source of information for developing a new business models. Facebook thanks to Graph API protocol, provides access to information on the behaviors through consistent view of the social graph with a representation of objects in the graph (e.g., people, photos, events, and pages) and the connections between them (e.g., friend relationships, shared content, and photo tags). Certainly, every user must provide permission to allow access to their data. Social network sites reflect how people act in their social environment. An interesting aspect of living in a society is how an individual’s behavior changes under the influence of other people. Social network sites seem to be a natural environment for influencing behaviors captured by SNSs like FB, become a good source for empirical research [3]. This reactive behavior is a powerful organizing standard that governs the social structures and weaves interpersonal webs of connection [5]. Several other studies on social network behavior have found presence of influence in groups with similar age or ethnicity [6].

Status updates are short messages that are posted to the personalized welcome page (News Feed) of all FB friends of the user as well as on the user’s own profile page. Most social networking sites, like FB, Google+, and MySpace utilize some form of status updates, and in some cases, like on Twitter, they serve as the main function. These posts are restricted in length (e.g., 420 characters on FB) and recipients can comment on them or indicate that they ‘‘like’’ them. Status updates enable effortless and fast one-to-many communication

This research could be utilized for understanding user behaviors that are useful for market analysis. However, user information must be analyzed with compliance of privacy laws to avoid any individual privacy infringements. Adequate care to anonymize data before publishing the results would be helpful.

In the next section, we provide a brief overview on the related works in the scope of the behavioral research on social networks, with the examples from FB-related studies. Additionally, we have reviewed select studies on user behaviors in social media. In the third section, we discuss our methods of empirical data gathering and provide background information on the research group and variables included in our pilot study. The fourth section presented results yielded from our analysis of empirical data. The final section discusses our conclusions and elaborates on future research opportunities from this work.

## II. Related Work

Companies today have their social media departments gain insights into the behaviors of their customers through social networks sites like FB, Twitter and others for understanding user behavior on social networks. In addition, social networks sites are ideal grounds for observing influential behaviors that encourage more interactions and affect their followers’ opinion [7]. In fact, [8] suggest SNS can be used for rating the relevance of information by using SNS conventions such as

978-83-60810-66-8/$25.00©2015, IEEE

# 13th Conference on Advanced Information Technologies for Management

WE ARE pleased to invite you to participate in the 11th edition of Conference on "Advanced Information Technologies for Management AITM'15". The main purpose of the conference is to provide a forum for researchers and practitioners to present and discuss the current issues of IT in business applications. There will be also the opportunity to demonstrate by the software houses and firms their solutions as well as achievements in management information systems.

## Topics

The topics of interest include but are not limited to:

- Concepts and methods of business informatics
- Business Process Management and Management Systems (BPM and BPMS)
- Management Information Systems (MIS)
- Enterprise information systems (ERP, CRM, SCM, etc.)
- Business Intelligence methods and tools
- Strategies and methodologies of IT implementation
- IT projects & IT projects management
- IT governance, efficiency and effectiveness
- Decision Support Systems and data mining
- Intelligence and mobile IT
- Cloud computing, SOA, Web services
- Agent-based systems
- Business-oriented ontologies, topic maps
- Knowledge-based and intelligent systems in management

## Event Chairs

**Dudycz, Helena,** Wrocław University of Economics, Poland

**Dyczkowski, Mirosław,** Wrocław University of Economics, Poland

**Korczak, Jerzy,** Wrocław University of Economics, Poland

## Program Committee

**Abramowicz, Witold,** Poznan University of Economics, Poland

**Ahlemann, Frederik,** University of Duisburg-Essen, Germany

**Andres, Frederic,** National Institute of Informatics, Tokyo, Japan

**Brown, Kenneth,** Communigram SA, France

**Chmielarz, Witold,** University of Warsaw, Poland

**Cortesi, Agostino,** Università Ca' Foscari, Venezia, Italy

**Czarnacka-Chrobot, Beata,** Warsaw School of Economics, Poland

**De, Suparna,** University of Surrey, Guildford, United Kingdom

**Dufourd, Jean-François,** University of Strasbourg, France

**Franczyk, Bogdan,** University of Leipzig, Germany

**Januszewski, Arkadiusz,** UTP University of Science and Technology in Bydgoszcz, Poland

**Kannan, Rajkumar,** Bishop Heber College (Autonomous), Tiruchirappalli, India

**Kersten, Grzegorz,** Concordia University, Montreal, Poland

**Kowalczyk, Ryszard,** Swinburne University of Technology, Melbourne, Victoria, Australia

**Kozak, Karol,** Fraunhofer and Uniklinikum Dresden

**Leyh, Christian,** Technische Universität Dresden, Chair of Information Systems, esp. IS in Manufacturing and Commerce, Germany

**Ligęza, Antoni,** AGH University of Science and Technology, Poland

**Ludwig, André,** University of Leipzig, Germany

**Magoni, Damien,** University of Bordeaux – LaBRI, France

**Michalak, Krzysztof,** Wroclaw University of Economics, Poland

**Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland

**Pankowska, Malgorzata,** University of Economics in Katowice, Poland

**Pawełoszek, Ilona,** Częstochowa Univeristy of Technology

**Quirin, Arnaud,** University of Vigo

**Rot, Artur,** Wroclaw University of Economics, Poland

**Rudek, Radosław,** Wrocław University of Economics

**Stanek, Stanislaw,** General Tadeusz Kosciuszko Military Academy of Land Forces in Wroclaw, Poland

**Surma, Jerzy,** Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States

**Teufel, Stephanie,** University of Fribourg, Switzerland

**Tsang, Edward,** University of Essex, United Kingdom

**Wolski, Waldemar,** Uniwersytet Szczeciński

**Zanni-Merk, Cecilia,** Universite de Strasbourg, France

**Ziemba, Ewa,** University of Economics in Katowice, Poland

DOI: 10.15439/2015F338
ACSIS, Vol. 5

# Buying stock market winners on Warsaw Stock Exchange - quantitative backtests of a short term trend following strategy

Aleksander Fafuła
Wrocław University of Economics
ul. Komandorska 118/120 53-345 Wrocław
Email: aleksander@fafula.com

Krzysztof Drelczuk
Wrocław University of Economics
ul. Komandorska 118/120 53-345 Wrocław
Email: krzysztof.drelczuk@gmail.com

**This paper focuses on one of the most popular issues in the Polish finance – is the 'buying stock market winners' profitable on the Warsaw Stock Exchange? This study tested whether Ichimoku trend following strategy performed better than simple buy & hold benchmark. For automated backtests WIG30 index components in the period 2012-12-28 to 2015-05-06 were used. The empirical results suggest that buying recent "winners" is very ineffective. These preliminary findings may imply contrarian nature of the short-term Polish financial market.**

## I. Introduction

TRADERS take many investing approaches. One of these tactics is to buy well-performing companies with expectations the performance will continue. The decision can be made using either, among others: momentum, or trend following strategies. These strategies may appear similar, but in reality they use different assumptions.

Trend following is an investment or trading strategy that tries to take advantage of long, medium or short-term price movements that seem to play out in various markets. Traders who employ a trend following strategy do not aim to predict specific price levels; they simply jump on the trend (when they perceive that a trend has established with their own peculiar reasons or rules) and ride it. A market "trend" is a tendency of a financial market price to move in a particular direction over time. If there is a turn contrary to the trend, they exit and wait until the turn establishes itself as a trend in the opposite direction. In cases their rules signal an exit, the trader closes long positions and re-enters when the trend is re-established [1, 2].

Momentum strategies rely on the assumption that prices respond (at least in part) to the strength of their supply and demand inputs. Momentum takes many forms: the earnings reports for publicly traded companies, the relationship between buyers and sellers in the market, the typical rate of historical price rises and falls, etc. In a sense, momentum trading can be paradoxically concerned with the fundamentals of technical analysis [3].

There are many trend following strategies [4]. For instance "Cross Exponential Moving Average (EMA)" enters the market when candle closes above 5-period EMA and exits when candle closes below 20-period EMA. "Simple Moving Average (SMA) and Moving Average Convergence Divergence (MACD)" on the other hand, enter the market when price crosses 20-period (this value can be adjusted) EMA and MACD oscillator is positive. Most strategies are based on fairly simple indicators operating on short-term historical prices. Due to that feature of trend-following strategies, authors have chosen Ichimoku method. This method sometimes is called trading system due to its complexity comparing to standard trend following techniques. Ichimoku is also a moving average-based trend identification system, but it contains more data points than standard candlestick charts and provides a clearer picture of potential price action. Authors have chosen this technique not because it is better than the others, not because it performs better or worse but because it takes into consideration much more aspects of price than other strategies. This feature of this technique is crucial for quantitative backtests performed in this paper.

This paper re-examines the profitability of a short-term trend following strategy, which aims to buy stocks that have performed well in the past. These backtests do not include short-sale trades. It is not the author's intention to prove or disprove efficiency of Ichimoku technique. The main question to be answered is: does buying recent winners from the WIG30 index pay off.

The question of profitability of the trend following strategy on the Warsaw Stock Exchange (WSE) is important because existing evidence provides mixed results. Numerous scientists documented abnormal profits of the momentum or trend following strategies. Rouwenhorst [5] finds it on twelve European markets, and [6] for some stock emerging markets. Hameed and Yuanto [7] positively identify six Asian markets. Schiereck, DeBondt and Weber [8] find momentum profitable for intermediate-term German market. In contrast to these findings there are numerous examples of contrarian markets. For example Jagadeesh [9] and Lehmann [10] find reversals in short-term horizons. DeBondt and Thaler [11, 12] report long-term price reversals. Chang, McLeavey and Rhee [13] document short-term contrarian nature of Japan. Hameed and Ting [14] have similar findings about Malaysia. Finally Kang, Liu, and Ni [15] report the overreaction to firm-specific information as the single most important source of the short-term contrarian profit in China stock market. The state of short-term trend following on Polish WSE lacks similar conclusions.

978-83-60810-66-8/$25.00©2015, IEEE

## II. METHOD

For the purpose of the backtests an automated trading-agent was implemented. Basically the agent buys and sells stocks according to hard-coded Ichimoku rules. The testing framework is a part of the A-Trader system [16]. Presented work is an extension and follow-up of previous experiments conducted using A-Trader. Technically, the presented system is a multi-agent solution that supports the analysis of the time series of high frequency, such as trading instruments. The main features are its openness for integration, development of new system functionality and ensuring adequate communication between the various agents. The agents can act as data providers, indicators or final decision makers. The service orientated architecture and cloud computing solves the problem of computing power. This could be an issue for higher than daily trading frequencies.

The Ichimoku trading-agent consists of 6 elements (as in Ichimoku trading system [17]).

Tenkan-sen calculation: (highest high + lowest low)/2 for the last 9 periods. It is primarily used as a signal line and a minor support/resistance line. The Tenkan Sen is an indicator of the market trend. If the red line is moving up or down, it indicates that the market is trending. If it moves horizontally, it signals that the market is ranging. The Tenkan-sen Line is computed below.

$$TS = \frac{max(s_9) - min(s_9)}{2}$$

$s_9$ - time series created from last 9 periods.

Kijun-sen calculation: (highest high + lowest low)/2 for the past 26 periods. This is a confirmation line, a support/resistance line, and can also be used as a trailing stop line. The Kijun Sen acts as an indicator of future price movement. If the price is higher than the blue line, it could continue to climb higher. If the price is below the blue line, it could keep dropping. The Kijun-sen Line is computed below.

$$KS = \frac{max(s_{26}) + min(s_{26})}{2}$$

$s_{26}$ - time series created from last 26 periods.

Senkou span A calculation: (Tenkan-sen + kijun-sen)/2 plotted 26 periods ahead. Also called leading span 1, this line forms one edge of the kumo, or cloud if the price is above the Senkou span, the top line serves as the first support level while the bottom line serves as the second support level. If the price is below the Senkou span, the bottom line forms the first resistance level while the top line is the second resistance level. Span is computed below.

$$SSA = \frac{TS + KS}{2}$$

TS - tenkan-sen line,
KS - tenkan-sen line.

Senkou span B calculation: (highest high + lowest low)/2 calculated over the past 52 time periods and plotted 26

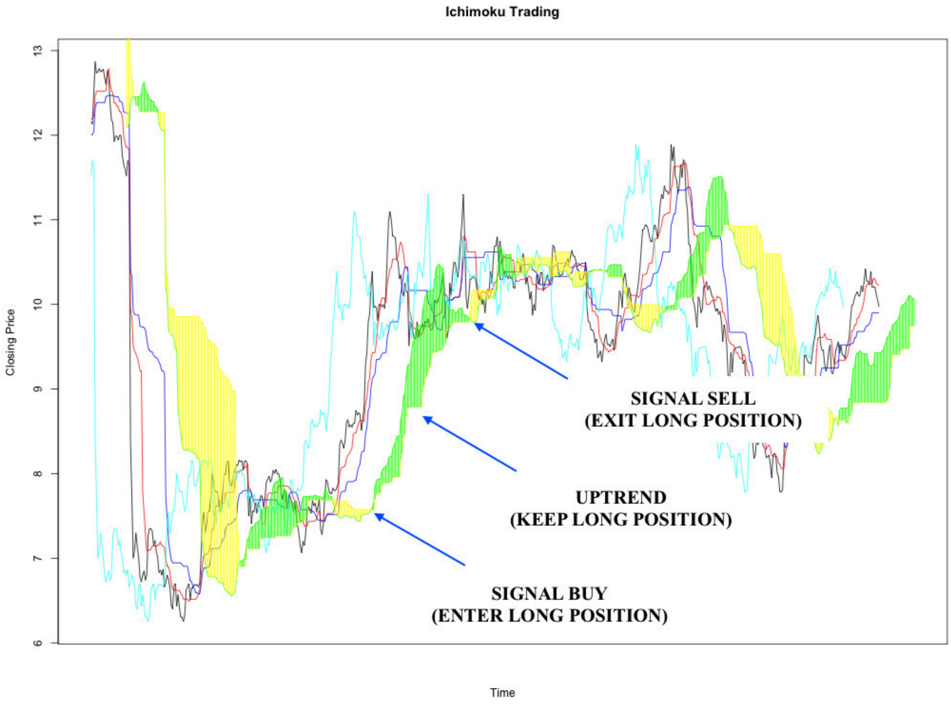


Fig. 1 Ichimoku Trading - entering and exiting long positions (symbol ORANGEPL).

periods ahead. Also called leading span 2, this line forms the other edge of the kumo. Span is computed below.

$$SSB = \frac{max(s_{52}) + min(s_{52})}{2}$$

$s_{52}$ - time series created from last 26 periods.

Kumo cloud is the space between senkou span A and B. The cloud edges identify current and potential future support and resistance points. The Kumo cloud changes in shape and height based on price changes. The Kumo height represents volatility as larger price movements form thicker clouds, which creates a stronger support and resistance. As thinner clouds offer only weak support and resistance, prices can and tend to break through. Generally, markets are bullish when Senkou Span A is above Senkou Span B and vice versa. Traders often look for Kumo Twists in future clouds, where Senkou Span A and B exchange positions, a signal of potential trend reversals. In addition to thickness, the strength of the cloud can also be ascertained by its angle; upwards for bullish and downwards for bearish. Any clouds behind price are also known as Kumo Shadows.

Chikou line calculation: today's closing price projected back 26 days on the chart. Also called the lagging span it is used as a support/resistance aid. If the Chikou Span or the green line crosses the price in the bottom-up direction, that is a buy signal. If the green line crosses the price from the top-down, that is a sell signal. Visually the rules implemented in automated-strategy are presented on the figure 1.

Figure 1 shows all Ichimoku elements with entry and exit positions. The colors on the figure corresponds to listed below Ichimoku elements:

- red Tenka-sen line,
- blue Kijun-sen line,
- green Senkou span A,
- yellow Senkou span B,
- cyan Chikou line.

It is important to avoid biases related to backtesting buy-only strategies on uptrend markets only. Therefore the backtests were conducted during diversified market periods. Fig. 2 shows corresponding WIG30 period (daily observations).

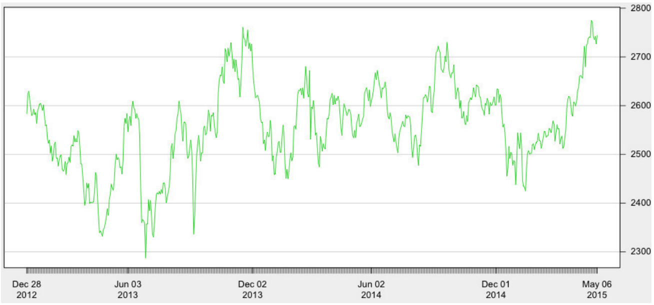


Fig. 2 WIG30, in backtested period, presenting non-monotonous trend.

During the backtested period the daily WIG30 returns had moderate skewness, with mean centered almost around zero. The weekly and monthly WIG30 returns were also slightly positive, with no significant variance. The summary of basic statistics is presented in table 1.

The backtests were performed in A-Trader with extensions written in R programming language. The reason of choosing R is because it is well tested statistical programming framework with wide variety of libraries and large community. This approach makes the backtests less vulnerable to errors. The backtesting architecture, presented in detail in [16], is shown at the fig. 3.

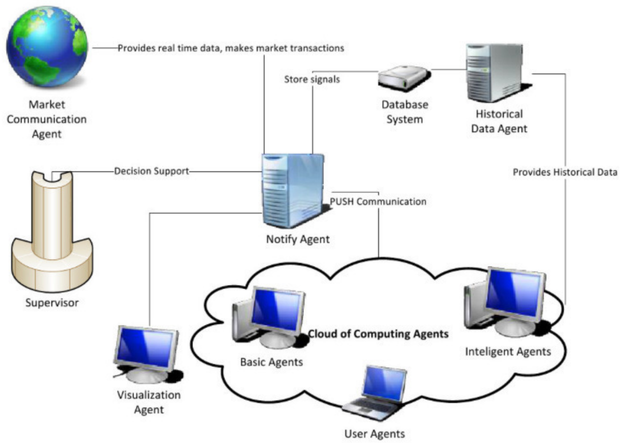


Fig. 3 A-Trader architecture.

TABLE I.
STATISTICS OF WIG30 RETURNS - VARIOUS TIME RESOLUTIONS

| Returns | # Obs | Minimum | Quartile 1 | Median | Mean | Quartile 3 | Maximum | Stdev |
|---|---|---|---|---|---|---|---|---|
| Daily | 581 | -0.0527 | -0.0054 | 0.0030 | 0.002 | 0.0057 | 0.0311 | 0.0097 |
| Weekly | 123 | -0.0865 | -0.0122 | 0.0019 | 0.007 | 0.0126 | 0.0578 | 0.0217 |
| Monthly | 29 | -0.0877 | -0.0189 | 0.0031 | 0.0028 | 0.0328 | 0.0818 | 0.0386 |

The following agents and components are distinguished in the A-Trader architecture:

- Notify Agent (NA),
- Historical Data Agent (HDA),
- Cloud of Computing Agents (CCA),
- Market Communication Agent (MCA),
- User Communication Agent (UCA),
- Supervisor (S),
- Database System (DS).

The data was acquired from the brokerage department of BOS Bank (the Polish Bank Ochrony Środowiska). Data was aggregated to 1-day periods. As a middle storage layer, a HDFS distributed file system was used. Such approach enables huge improvements in computation time and allows the data to perform millions of simulation in real time. These extensions of the A-Trader built framework are easily customizable and can be used in a variety of tests with minimum programming work.

## III. RESULTS

For the purpose of backtesting we took components of WIG 30 index (as of 2015.05.07). The list of backtested companies included: ALIOR, ASSECOPOL, BOGDANKA, BORYSZEW, BZWBK, CCC, CYFRPLSAT, ENEA, ENERGA, EUROCASH, GRUPAAZOTY, GTC, HANDLOWY, INGBSK, JSW, KERNEL, KGHM, LOTOS, LPP, MBANK, ORANGEPL, PEKAO, PGE, PGNIG, PKNORLEN, PKOBP, PZU, SYNTHOS, TAURONPE, TVN. Next, the data was trimmed to the earliest possible point where the WIG30 index provides first observation (2012-12-28). Additionally the company quotes were shifted by 52 trading days to prepare signals for trading since the first day of WIG30 (2012-10-12).

Each of the benchmarks produced a visual output of Ichimoku components with price. Additionally, cumulative returns of the tested trend-following strategy were plotted along the classical "Buy & Hold" benchmark.

Visual examination of two different results of strategies helps understand the mechanics of gains and losses. The first example shows a situation where the cumulative return of trend following-strategy outperformed the Buy & Hold benchmark. Figure 4 presents the company with symbol GRUPAAZOTY and its Ichimoku components. Figure 5 presents cumulative returns.

Opposite to the previous example where Buy & Hold outperformed, the trend following Ichimoku strategy is presented on figures 6 (flags, lines) and 7 (cumulative returns).

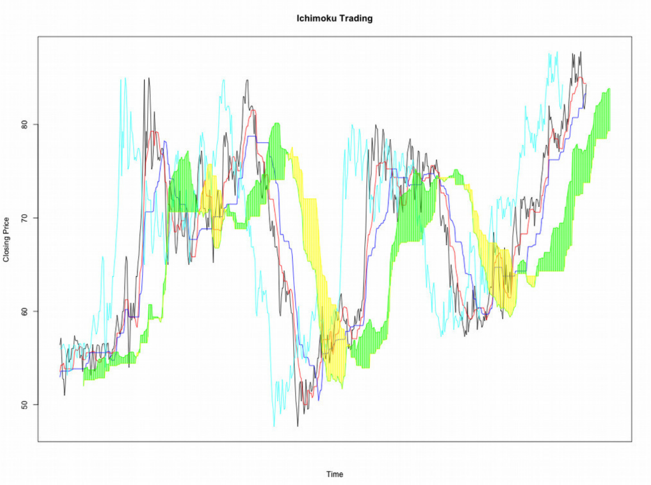


Fig. 4 Ichimoku flags and lines for GRUPAAZOTY.

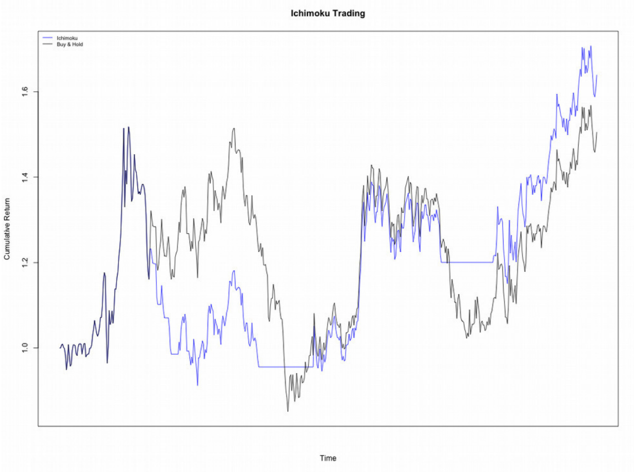


Fig. 5 Cumulative returns - comparison of Ichimoku and Buy & Hold strategies (GRUPAAZOTY).

Final summary of the results leave no doubts. The trend following strategy performed worse than Buy & Hold in 26 out of 30 cases. In most cases, the winning situations were merely protecting losing positions. The table 2 presents summary of the results for all the companies in the backtests. The numbers in column I and II show cumulative returns. Values below "1" are losses of initial capital.

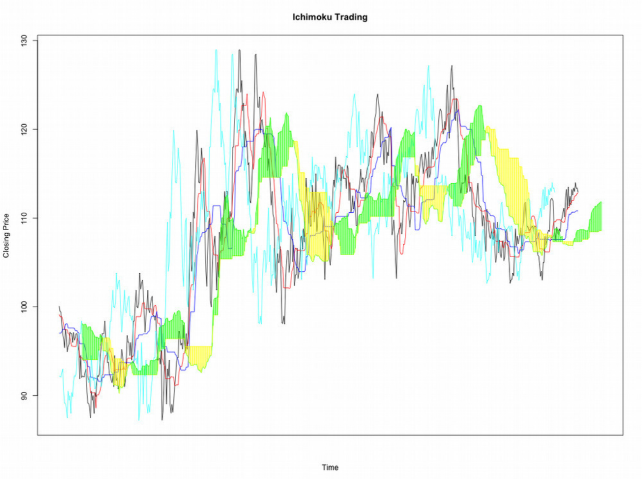


Fig. 6 Ichimoku flags and lines for HANDLOWY.

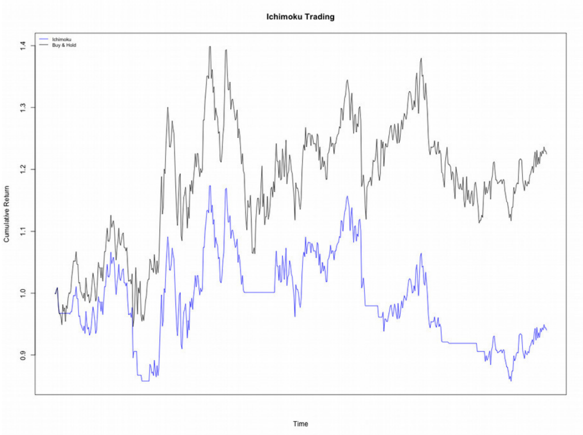


Fig. 7 Cumulative returns - comparison of Ichimoku and Buy & Hold strategies (HANDLOWY).

The statistics of cumulative returns across backtests is presented in table III.

## IV. Discussion

As shown in backtests buying winners on Warsaw Stock Exchange with Ichimoku short trend following strategy did not perform well. There might be a few reasons for such results.

First is past-winners do not have enough strength to beat the market in the future. Buy and hold benchmark seems to confirm this thesis. This is not a typical situation on well-developed capital markets where buying winners usually outperform the market [18]. The assets for the backtests were taken from the WIG30 index component list. Perhaps those assets were considered overpriced, due to the short term overreaction to the news, and therefore there was no steady potential to outperform the market. Investigation into other segments of the market might be helpful to verify this statement.

Second reason of the losses is the potential inefficiency of Ichimoku technique as the trend following strategy. There is a possibility that Ichimoku with used parameters does not follow trend as intended. In such cases, further backtests with different strategies might deny or confirm this thesis. However looking at the Buy & Hold strategy, it seems that the buying "winners" approach is the problem, not the chosen strategy itself. Nevertheless other strategies might shed some additional light on this matter.

In general, trend following scored worse in almost every case, except one: investment protection. The minimum Buy & Hold cumulative return is 0.1562, while, for trend following, the protection shut the trading down at 0.4646. Although this additional protection limits losses, it also trims gains.

Table II.
Statistics of WIG30 returns - various time resolutions

| Trend Following | Buy & Hold | Symbol | Result |
|---|---|---|---|
| 0,95 | 1,32 | ALIOR | LOSS |
| 1,38 | 1,39 | ASSECOPOL | LOSS |
| 0,60 | 0,65 | BOGDANKA | LOSS |
| 1,16 | 1,21 | BORYSZEW | LOSS |
| 1,05 | 1,39 | BZWBK | LOSS |
| 2,15 | 2,72 | CCC | LOSS |
| 1,34 | 1,56 | CYFRPLSAT | LOSS |
| 0,69 | 1,09 | ENEA | LOSS |
| 1,18 | 1,31 | ENERGA | LOSS |
| 0,68 | 0,78 | EUROCASH | LOSS |
| 1,64 | 1,51 | GRUPAAZOTY | WIN |
| 0,89 | 0,74 | GTC | WIN |
| 0,94 | 1,22 | HANDLOWY | LOSS |
| 1,52 | 1,62 | INGBSK | LOSS |
| 0,61 | 0,16 | JSW | WIN |
| 0,46 | 0,54 | KERNEL | LOSS |
| 0,59 | 0,69 | KGHM | LOSS |
| 0,73 | 0,84 | LOTOS | LOSS |
| 1,32 | 1,42 | LPP | LOSS |
| 1,19 | 1,48 | MBANK | LOSS |
| 1,16 | 0,87 | ORANGEPL | WIN |
| 0,78 | 1,21 | PEKAO | LOSS |
| 1,00 | 1,22 | PGE | LOSS |
| 1,13 | 1,20 | PGNIG | LOSS |
| 1,00 | 1,38 | PKNORLEN | LOSS |
| 0,84 | 1,06 | PKOBP | LOSS |
| 0,80 | 1,18 | PZU | LOSS |
| 0,75 | 0,94 | SYNTHOS | LOSS |
| 0,85 | 1,11 | TAURONPE | LOSS |
| 1,86 | 1,89 | TVN | LOSS |

Final conclusions are that joining recently established uptrends, in the case of WIG30 components in period 2012-12-28 to 2015-05-06, did not lead to excessive returns and did not provide any other kind of improvements. Of course, the presented backtesting approach is very general: the algorithms can be adjusted, additional variables can be specified, and more sophisticated models can be used. However, the broad backtests performed in this research

TABLE III.
STATISTICS BETWEEN RETURNS ACROSS ASSETS

| | **Trend Following (Ichimoku)** | **Buy & Hold** |
|---|---:|---:|
| **Min.** | 0.4646 | 0.1562 |
| **1st qu** | 0.7579 | 0.8830 |
| **Median** | 0.9768 | 1.2070 |
| **Mean** | 1.0424 | 1.1884 |
| **3rd qu** | 1.1890 | 1.3911 |
| **Max.** | 2.1539 | 2.7209 |

negate momentum, and suggest contrarian nature of WIG30 components.

## REFERENCES

[1] Lempérière, Y., Deremble, C., Seager, P., Potters, M., & Bouchaud, J. P. (2014). Two centuries of trend following. arXiv preprint arXiv:1404.3274.

[2] Sasaki, Hidendobu, “Ichimoku Kinko Studies”, Toshi Raider Publishing, 1996

[3] Burghardt, G., & Walls, B. (2011). Two benchmarks for momentum trading. Managed Futures for Institutional Investors: Analysis and Portfolio Construction, 99-127. B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.

[4] Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. Journal of Financial Economics, 104(2), 228-250.

[5] Rouwenhorst, K.G., (1998). International momentum strategies, Journal of Finance 53, 267-284.

[6] Rouwenhorst, K.G., (1999) Local return factors and turnover in emerging stock markets, Journal of Finance 54, 1439-1464.

[7] Hameed, A. and K. Yuanto, 2002, Momentum strategies: evidence from the pacific basin stock markets, Journal of Financial Research, 25(3), 383-397.

[8] Schiereck, D., W. DeBondt and M. Weber, 1999, Contrarian and momentum strategies in Germany, Financial Analysts Journal 155,104-116.

[9] Jegadeesh, N., 1990, Evidence of predictable behavior of security returns, Journal of Finance 45 881-898.

[10] Lehmann, B.N., 1990, Fads, martingales and market efficiency, Quarterly Journal of Economics 105, 1-28.

[11] DeBondt, W.F.M. and R. Thaler, 1985, Does the stock market overreact? Journal of Finance 40, 793-805.

[12] DeBondt, W.F.M. and R. Thaler, 1987, Further evidence on investor overreaction and stock market seasonality, Journal of Finance, 42, 557-581.

[13] Chang, R.P., D.W. McLeavey and S.G. Rhee, 1995, Short-term abnormal returns of the contrarian strategy in the Japanese stock market, Journal of Business Finance and Accounting 22, 1035-1048.

[14] Hameed, A. and S. Ting, 2000, Trading volume and short-horizon contrarian profits, evidence from Malaysian stock market, Pacific-Basin Finance Journal 8, 67-84.

[15] Kang, J., Liu, M. H., and Ni, S. X. (2002) Contrarian and momentum strategies in China stock market: 1993–2000. Pacific-Basin Finance Journal, 10(3): 243–265.

[16] Korczak, Jerzy, et al. "A-Trader—Consulting agent platform for stock exchange gamblers." Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on. IEEE, 2012.

[17] Ichimoku, Sanjin “Ichimoku Kinko Charts”, Keizai Hendo Kenkyujo, 1981.

[18] Jegadeesh, at al, “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency”, The Journal of Finance, vol. XLVIII, 1993.

DOI: 10.15439/2015F155
ACSIS, Vol. 5

# Fuzzy logic in the multi-agent financial decision support system

Jerzy Korczak, Marcin Hernes, Maciej Bac
Wrocław University of Economics ul. Komandorska 118/120, 53-345 Wrocław, Poland

e-mail:{jerzy.korczak, marcin.hernes, maciej.bac}ue.wroc.pl

***Abstract*—The article presents the application of a fuzzy logic in building the trading agents of the a-Trader system. The system supports investment decisions on the FOREX market. The first part of the article contains a discussion related to the use of fuzzy logic as an agents' knowledge representation. Next, the algorithms of the selected fuzzy logic buy-sell decision agents are presented. In the last part of the article the agent performance is evaluated on real FOREX data.**

## I. Introduction

FINANCIAL decisions are made under conditions of risk and uncertainty that influence their level of performance. These decisions are usually supported by decision support systems and various computational models. Among them, there are multi-agent systems [2] which use various methods based on mathematics, statistics, finance or artificial intelligence [3, 4, 6, 10, 13, 14, 18, 20, 21,25, 28, 48]. A-Trader [22] supporting investment decisions on the FOREX market (Foreign Exchange Market) may serve as the example of such a system. FOREX is one of the biggest financial foreign exchange markets in the world. Currencies are traded against one another in pairs, for instance EUR/USD, USD/PLN. Trading on FOREX relies on opening/closing long/short positions. A long position is a situation in which one purchases a currency pair at a certain price and hopes to sell it later at a higher price. This is also referred to as the notion of "buy low, sell high" in other trading markets. On FOREX, when one currency in a pair is rising in value, the other currency is declining, and vice versa. If a trader thinks a currency pair will fall, he will sell it and hope to buy it back later at a lower price. This is considered a short position, which is the opposite of a long position. The A-Trader receives tick data which are grouped to minute aggregates (M1, M5, M15, M30), hourly aggregates (H1, H4), daily aggregates (D1), weekly aggregates (W1) and monthly aggregates (MN1). The A-Trader supports a High Frequency Trading (HFT) and puts strong emphasis on price formation processes, short-term positions, fast computing, and efficient and robust indicators.

High frequency traders are constantly taking advantage of very small quote changes with a high rate of recurrence to generate important profit rates. As many HFT experts underline, the traders seek profits from the market's liquidity imbalances and short-term pricing inefficiencies. Hence, the minimization of time from the access to quote information, through the entry of an order until its execution, is vital. Generally speaking, to support traders, the systems must provide as soon as possible advice as to which position should be taken: buy, sell or do nothing. Time series forecasting is more difficult while online trading has to be served.

The architecture of a-Trader and the description of the different groups of agents have already been detailed [23, 24]. In general, the agents possess their own knowledge, they can continuously learn and change their knowledge in order to improve their performance.

Different methods of agents' knowledge representation can be applied in a-Trader. In our previous work [23, 24] we were focused on three-valued knowledge representation of this group of agents. Value "1" denoted „buy" decision, value "-1" denoted „sell" decision, value "0" denoted „ leave unchanged". Agents are implemented using the C# environment and MQL4 language.

The key part of the system is the Supervisor agent. Its task is, among others, to coordinate the work of agents on trading strategy and it presents the final strategy (suggestions of open/close positions) to the trader. The Supervisor uses various strategies and evaluates their performance.

The a-Trader allows also for making arbitrarily independent decisions by traders (experts)on the basis of their knowledge and experience. The traders' decisions can be stored in a database, evaluated and compared with strategies provided automatically by agents.

The purpose of this paper is to present a manner of applying a fuzzy logic as the agents' knowledge representation and evaluating the performance of selected agents in the a-Trader system.

In the first part of the article, the fuzzy logic as agents' knowledge representation is briefly discussed. The algorithms of three selected agents are then described. The final part discusses the results of the performance evaluation of these agents.

978-83-60810-66-8/$25.00©2015, IEEE

## II. Fuzzy logic as agents' knowledge representation

The literature on the subject presents many different methods for agents' knowledge representation. The main ones include first-order predicate logic, production systems, artificial neural networks, frame representation, ontologies such as semantic web and semantic networks, multi-attributes and multi-values structures, and multi valued logic [12, 17, 32, 33, 39, 40, 42, 43, 44, 46]. Some of these methods are closely related to fuzzy logic.

The first-order predicate logic that is one common knowledge representation is founded on the following general assumptions [11]:

- the knowledge representation is independent of physical media,
- agents' internal states are related to the objects of external environment,
- the knowledge representation consists of symbols forming the structure,
- reasoning is based on the manipulation of these structures to derive other structures.

Often agents' knowledge is represented as multi-attribute and multi-value structures which allow representation the real world environment in wide scope of objects features.

Multi-valued logic and fuzzy logic are more suitable methods for HFT. Three-valued logic is a very simple language consisting of proposition symbols and logical connectives. It can handle propositions that are known true, known false, or completely unknown. The set of possible models, given a fixed propositional vocabulary, is finite, so entailment can be checked by enumerative models. Inference algorithms for three-valued logic include backtracking and local–search methods and can often solve large problems very quickly [35]. Three-valued logic is reasonably effective for certain tasks, but does not scale to environments of unbounded size because it lacks the expressive power to describe the real world objects.

To reduce this weakness, a fuzzy logic can be applied in HFT. Fuzzy logic is an approach founded on "degrees of truth" rather than the usual "true or false" values (1 or 0). The idea of fuzzy logic was first proposed by Zadeh in the 1960s when he was working on the problem of computer understanding of natural language [47]. Fuzzy logic is a form of multi-valued logic derived from fuzzy set theory [3] to deal with approximate reasoning. In contrast to "crisp logic", where binary sets are processed by binary logic, fuzzy logic variables may have a truth value that ranges between 0 and 1 that allows the user to express imprecision and flexibility in a decision-making system [41], [45], [30]. Fuzzy logic is used, for example, in multi-agent systems for information extraction [34], energy management [26] or robotics [18]. Fuzzy logic was also used for trading on FOREX, for example, in Expert Advisor [31] or technical analysis system [9] or fuzzy time series forecasting [1], [8], [36]. However, in these systems, the probability of decisions is ranged to [0..1]. In trading systems it is unfavorable, because the trader can buy, sell or leave a currency unchanged. Therefore, in the a-Trader system, the confidence of decisions range is [-1..1], where "-1" level denotes "strong sell" decision, "0" level denotes "strong leave unchanged" decision and "1" level denotes "strong buy" decision. The positions can be open/close with different levels of confidence of decision. For example, the long position can be open, when a level of confidence is 0.6 or short position can be open, when a level of probability is 0.7. Therefore the timeframe for the opening/closing position is wider than in the case of three valued-logic. An example of this difference is presented in Fig 1 and Fig 2 ($A_i$ – denotes the ith agent). In the case of three-valued logic (Fig 1), the green color points denote a "buy" decision, the red color ones denote a "sell" decision, and the black color points denote a "do nothing" decision. There are often agents that generate buy/sell decisions too fast or too late. In the case of fuzzy logic, the ranges of decisions probability often cover the best point for trading. In Fig. 2, the green triangle denotes transition from "do nothing" decision to "buy" decision, and the red inverted triangle denotes transition from "do nothing" decision to "sell" decision, and the black color denotes a "do nothing" decision). Therefore it is possible to place open/close positions closer to the optimal decision than in the case of a three-valued logic. Of course, the level of probability of decision for open/close position plays a vital role. This level can be determined on the basis of trader experience, or by the Supervisor on the basis of, for example, a genetic algorithm.

Using the fuzzy logic as agents' knowledge representation allows the trading decision to be closer to real experts' decisions (made under conditions of risk and uncertainty) that are also taken with a certain level of probability.

Fuzzy logic can be also used by trading advisors for the following tasks:

- forecasting, i.e. the possibility to calculate the output value for input data lies outside the scope initially predicted,
- expressing the agent's knowledge in a flexible, intuitive way.,
- computation of decisions' probability level,
- implementation of different automated learning algorithms,
- validation and consistency measuring that can speed up automated learning and improve user interpretability,
- taking into consideration ambiguity – the "natural" way for expressing uncertain knowledge.

The next part of the article describes selected fuzzy logic buy-sell decision agents implemented in a-Trader.

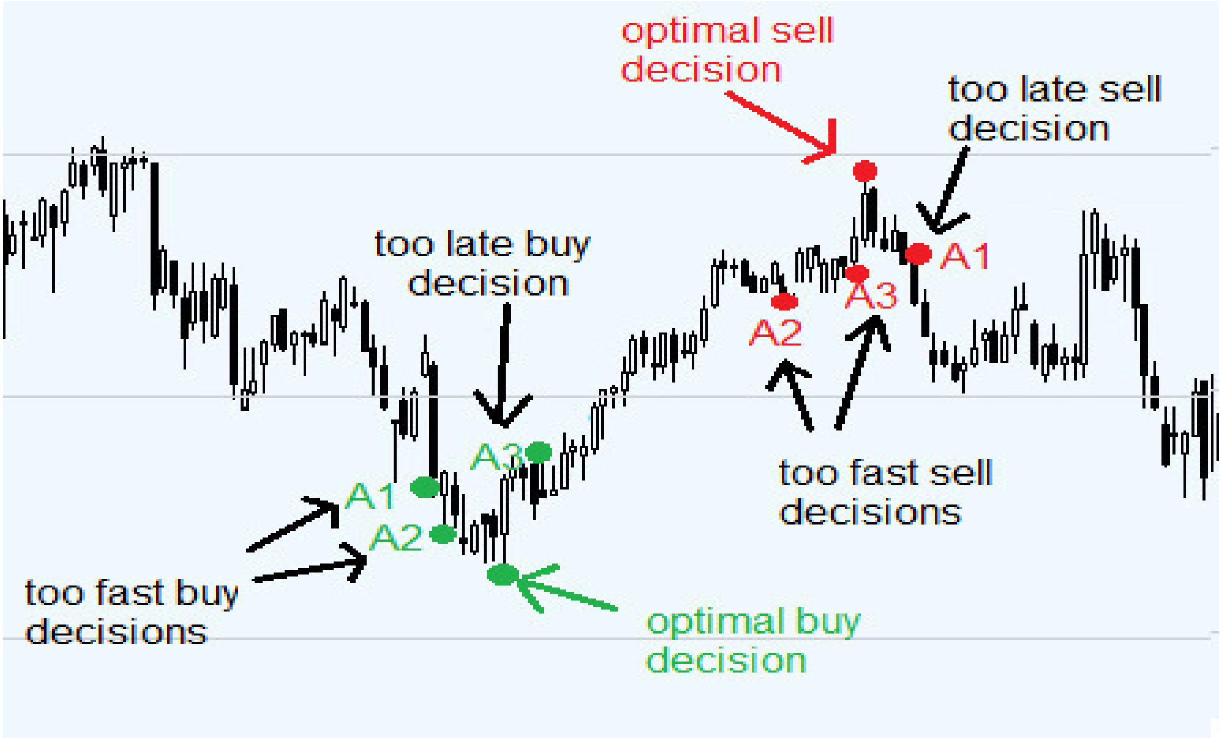


Fig. 1 Three-valued logic agents' decisions
Source: Own work.

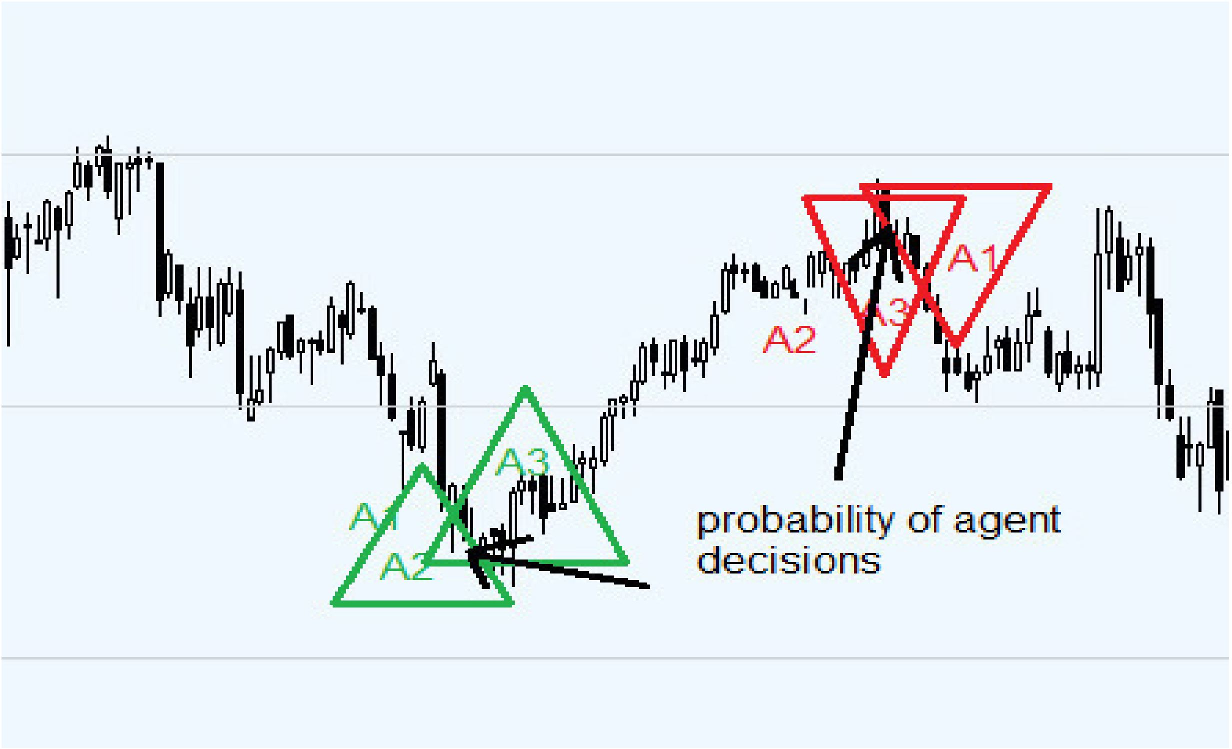


Fig. 2 Fuzzy logic agents' decisions
Source: Own work.

## III. Description of the fuzzy logic buy-sell decision agents

A-Trader contains approximately 1400 agents, including about 800 processing data agents (they calculate different indicators on the FOREX market, for instance trend indicators, oscillators) and 300 agents (running in all time periods) setting the buy-sell decision, including: 200 three valued logic agents and 100 fuzzy logic agents, also 200 agents providing the strategies. In order to illustrate the performance analysis, four agents were chosen: *BollingerFuzzy, WilliamsFuzzy, TrendLinear-RegFuzzy* and *ConsensusFuzzy*.

### *A. . The BollingerFuzzy agent*

The *BollingerFuzzy* agent is created on the basis of the Bollinger Bands indicator [5]. These bands are volatility constraints placed above and below a moving average. Volatility is expressed by the standard deviation, which changes as volatility increases and decreases.

The bands automatically widen when volatility increases and narrow when volatility decreases. The buy decision's probability level is calculated when the price is close to the upper Bollinger Band or breaks above it, and the sell decision is calculated when the price is close to the lower Bollinger Band or falls below it. The algorithm of this agent is the following:

**Algorithm 1**

```
Input:   q  //a value of quotation,
         bbandup // value determination by processing data agent
           named BBANDUP, which calculates the upper band,
         bbandlo //  value determination by processing data agent
           named BBANDLO, which calculates the lower band.
         sma // value determination by processing data agent named
           SMA, which calculates the simple moving average of
           quotation.
Output:  The fuzzy logic decision D (value range [-1..1]).
BEGIN
Let D:=0, calcBands:=0; //counter for fuzzification.
      maxcount:=0.  //maximum counter limit for fuzzification.
      Δ =Abs((sma-((bbandlo+bbandup)/2))/10).
If q<(bbandlo (+Δ)) then
      If (calcBands>0) then calcBands=0, calcBands:=calcBands-1.
      If (calcBands<-maxcount) then calcBands=-maxcount,
         D=calcBands/maxcount;
If q>(bbandup (-Δ)) then
      If (calcBands<0) then calcBands=0, calcBands:=calcBands+1.
      If (calcBands>maxcount) then calcBands=maxcount;
         D=calcBands/maxcount;
END
```

In a trading system the fuzzification is understood as a process of conversion of an input variable (i.e. signals determined by processing agents) to fuzzy set.

### *B. . The WilliamsFuzzy agent*

The *WilliamsFuzzy* agent is created on the basis of Williams %R indicator [19]. Williams %R is a momentum indicator that is the inverse of the Fast Stochastic Oscillator. Also referred to as %R, the indicator reflects the level of the close relative to the highest high for the look-back period. In contrast, the Stochastic Oscillator reflects the level of the close relative to the lowest low. %R corrects for the inversion by multiplying the raw value by -100. As a result, the Fast Stochastic Oscillator and Williams %R produce exactly the same lines, only the scaling is different. Williams %R oscillates from 0 to -100. The buy decision's probability level is calculated when Williams %R value falls below -80 and the sell decision is calculated when Williams %R value rises above -20. The algorithm of this agent is as follows:

**Algorithm 2**

```
Input:   q  //a value of quotation,
         williams  // value determining by processing data agent
           named WILLIAMS,   which calculates the Williams %R
           indicator.
         Δ – an external parameter denotes range of williams %R less
           than -80 or above -20 (it is assumed that the maximum
           value of williams %R does not have to be 0 and the
           minimum value of williams %R does not have to be -100.
           This parameter range is [1..20] and it is calculated by the
           genetic algorithm or determined by user.
Output:  The fuzzy logic decision D (value range [-1..1]).
BEGIN
Let D:=0.
If williams<=-80 then  D=(williams-(-80- Δ))/ Δ.
If williams>=-20 then D=-(williams-(-20)/Δ.
END
```

### *C. . The TrendLinearRegFuzzy agent*

The agent operates on the basis of the assumption that the trend of a certain number of *M* quotations is approximated with the straight line by the equation: $y = ax + b$. The straight line inclination depends on the value of the "*a*" parameter or the tangent value of the inclination angle with the use of linear regression [20], [27], [38]. The agent computes the probability level of a buy decision when the coefficient value changes from positive to negative and the probability level of a buy decision is calculated when the coefficient changes value from negative to positive. The change in the agent's decision is made with the use of hysteresis, the level of which is defined by means of the coefficient Δ, the value of which should be higher than the transaction costs.

The algorithm can be described as follows:

**Algorithm 3**

```
Input:   $w=<w_1, w_2, .... w_M>$ //The vector of quotation value of the
         quotations consisting of  M quotations (current quotation and
         M-1 previous quotations – the M is calculated by  the genetic
         algorithm or determined by the user),
         preva  // the previous value the a coefficient.
Output:  The fuzzy logic decision D (value range [-1..1]) with respect
           to w and preva value.
BEGIN
Let  sumy:=0; sumx:=0.0; sumxy:=0; sumx2=0.
      // where: sumy means the sum of the value of M quotations,
      sumx means the sum of the particular quotations' number in
      vector (suma numerów poszczególnych notowań np. jeśli M=5 to
      sumx = 1+2+3+4+5 czyli 15) in the vector, sumxy means the
      sum of the products of the quotation value and particular
```

```
        quotation number in the vector, and sumx2 means the sum of
        the squares of quotation numbers in the vector.
       D:=0, countTRL:=0; //counter for fuzzification,
       maxcount:=0.  //maximum counter limit for fuzzification.
For (i=1;i<=M;i++)
     sumy:= sumy+$w_i$; sumxy:= sumxy+$w_i$*i; sumx=: sumx +i;
     sumx2:= sumx2+i*i; i:=i+1;
c:= sumx2*M-sumx*sumx.
If c=0 then c:=0,1.
a:=(sumxy*M-sumx*sumy)/c.
If (a=preva=0) or (a<0 and preva<0) or (a>0 and preva>0) then D:=0.
If (a>0 and preva<0) then
   If (countTRL>0) then countTRL=0.
   countTRL:=countRL-1.
   If (countTRL<-maxcount) then countTRL=-maxcount.
   D=countTRL/maxcount;
If (a<0 and preva>0) then
   If (countTRL<0) then countTRL=0.
   countTRL:=countTRL+1.
   If (countRL>maxcount) then countTRL=maxcount.
    D=countTRL/maxcount;
preva:=a.
END
```

### *D. The ConsensusFuzzy agent*

The *ConsensusFuzzy* agent (detailed in [22, 23]) is founded on the consensus theory [15, 16, 29] and determines the decisions on the basis of the set of decisions generated by other fuzzy logic agents in the system.

The algorithm is as follows:

**Algorithm 4**

```
Input:  A= {$D^{(1)}$, $D^{(2)}$, .... $D^{(M)}$ } //The profile consists of M fuzzy logic
          agents' decisions, where M – number of fuzzy logic agents in
          the system, $D^{(1)}$, $D^{(2)}$, .... $D^{(M)}$ – decisions of particular agents
Output: The Fuzzy logic consensus CON  (value range [-1..1])
according to A.
BEGIN
Let  CON:=0.
Determine a sequence B by sorting elements of A profile in an
increasing order.
$k_1$=(M+1)/2 the element of B.
$k_2$=(M+2)/2 the element of B.
Set CON as any value from interval [$k_1$, $k_2$].
END.
```

It should be noted that currently in the system there are 100 agents using fuzzy logic representation. This set of trading agents may be easily extended if required. The evaluation of the performance of presented fuzzy logic agents will be shown further in the article.

## IV. Experiments

The agents performance analysis is performed for data within the M1 period of quotations from the FOREX market. For the purpose of this analysis, a test was performed in which the following assumptions were made:

1. EUR/USD quotes were selected from randomly chosen periods, notably:
   - 17-04-2015, 9:40 am to 17-04-2015, 9:50 pm, (710 quotations)
   - 20-04-2015, 0:00 am to 20-04-2015, 7:00 pm (1140 quotations),
   - 22-04-2015, 0:00 am to 22-04-2015, 7:00 pm (1140 quotations),
2. At the verification, the strategies (signals for open long/close short position-equals to 1, close long/open short position-equals to -1) of the Supervisor are based on different decisions' probability levels calculated by fuzzy logic agents described in section III (the example of strategy is presented in Figure 3, where the green line means the "long position" and the red one the "short position").
3. It was assumed that decisions' probability levels for open/close position are determined by the genetic algorithm (on the basis of earlier periods).
4. It was assumed that the unit of performance analysis ratios (absolute ratios) is pips (a change in price of one "point" in Forex trading is referred to as a pip, and it is equivalent to the final number in a currency pair's price).
5. The transaction costs are directly proportional to the number of transactions.
6. The capital management - it was assumed that in each transaction the investor engages 100% of the capital held at the leverage 1:1. It should be pointed out that the investor may define another capital management strategy.
7. The performance analysis was performed with the use of the following measures (ratios):
   - rate of return (ratio $x_1$),
   - the number of the transaction,
   - gross profit (ratio $x_2$),
   - gross loss (ratio $x_3$),
   - the number of profitable transactions (ratio $x_4$),
   - the number of profitable transactions in a row (ratio $x_5$),
   - the number of unprofitable transactions in a row (ratio $x_6$),
   - Sharpe ratio (ratio $x_7$)

$$S = \frac{E(r) - E(f)}{|O(r)|} \cdot 100\% \qquad (1)$$

where:

$E(r)$ – arithmetic average of the rate of return,

$E(f)$ – arithmetic average of the risk-free rate of return,

$O(r)$ – standard deviation of rates of return.

- the average coefficient of volatility (ratio $x_8$) is the ratio of the average deviation of the arithmetic average multiplied by 100% and is expressed:

$$V = \frac{s}{\left|E(r)\right|} \cdot 100\% \,. \qquad (2)$$

where:

$V$ – average coefficient of variation,

$s$ – average deviation of the rates of return,

$E$(r) – arithmetic average of the rates of return.

- Value at Risk (ratio $x_9$) – the measure known as value exposed to the risk - that is the maximum loss of the market value of the financial instrument possible to bear in a specific timeframe and at a given confidence level [7].

$$VaR{=}P{*}O{*}k \qquad (3)$$

where:

$P$ – the initial capital,

$O$ – volatility - standard deviation of rates of return during the period,

$k$ – the inverse of the standard normal cumulative distribution (assumed confidence level 95%, the value of k is 1,65),

- the average rate of return per transaction (ratio $x_{10}$), counted as the quotient of the rate of return and the number of transactions.

8. For the purpose of the comparison of the agents' performance, the following evaluation function was elaborated:

$$y = (a_1 x_1 + a_2 x_2 + a_3(1 - x_3) + a_4 x_4 + a_5 x_5 + \ldots$$

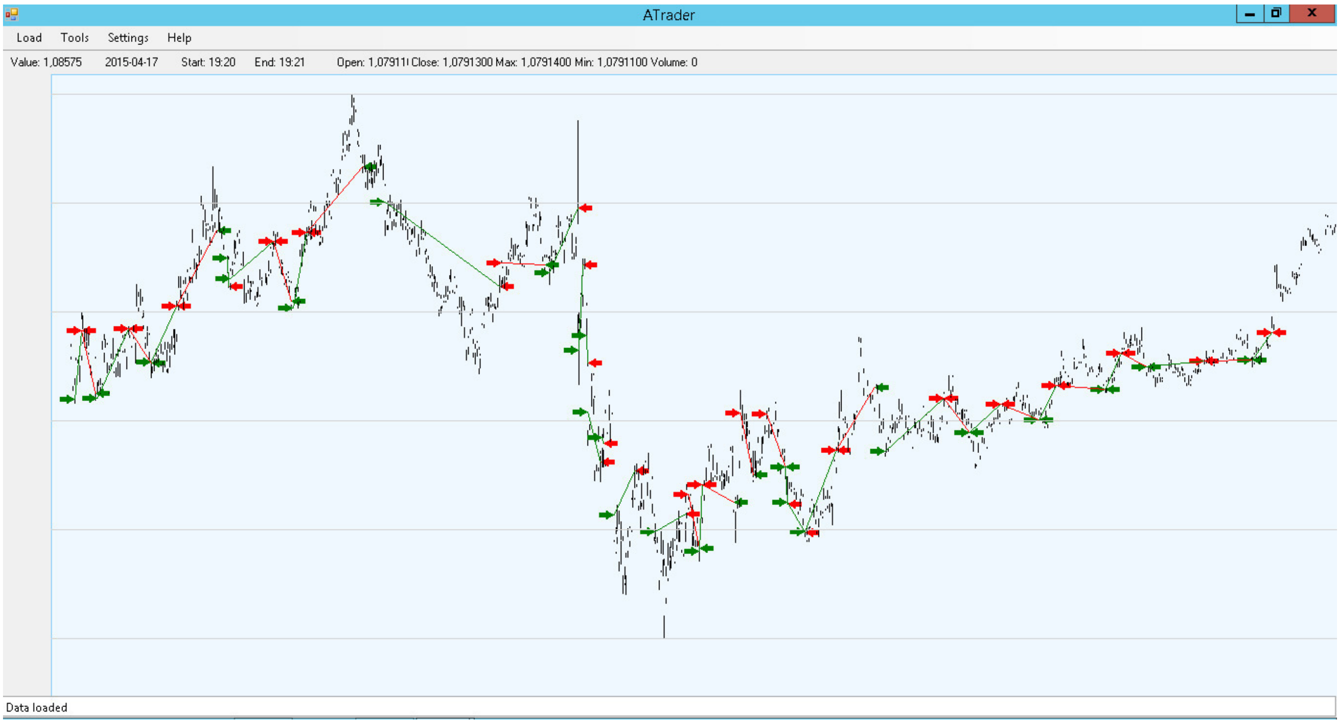
$$+a_6(1 - x_6) + a_7 x_7 + a_8(1 - x_8) + a_9(1 - x_{19}) + a_{10} x_{10}) \qquad (4)$$

where $x_i$ denotes the normalized values of ratios mentioned in item 6 from $x_1$ to $x_{10}$. It was adopted in the test that coefficients $a_1$ to $a_{10}$=1/10. It should be mentioned that these coefficients may be modified with the use of, for instance, an evolution/genetic method or determined by the user (investor) in accordance with his/her preference (for instance the user may determine whether he/she is interested in the higher rate of return with a simultaneous higher risk level or lower risk level, but simultaneously agrees to a lower rate of return).

The function is given the values from the range [0..1], and the agent's efficiency is directly proportional to the function value.

9. The results obtained by the tested agents were compared with the results of the Buy-and-Hold benchmark (a trader buys a currency at the beginning and sell a currency at the end of an investment period) and the EMA benchmark (Exponential Moving Average -- a type of moving average that is similar to a simple moving average, except that more weight is given to the latest data).

Table 1 presents the results obtained in the particular periods.



Value: 1,08575 2015-04-17 Start: 19:20 End: 19:21 Open: 1,07911! Close: 1,0791300 Max: 1,0791400 Min: 1,0791100 Volume: 0

Data loaded

Fig. 3 The example of strategy visualization.
Source: Own work.

TABLE I.
PERFORMANCE ANALYSIS RESULTS

| Ratio | BollingerFuzzy | | | WilliamsFuzzy | | | TrendLinearRegFuzzy | | | ConsensusFuzzy | | | EMA | | | B & H | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Period 1 | Period 2 | Period 3 | Period 1 | Period 2 | Period 3 | Period 1 | Period 2 | Period 3 | Period 1 | Period 2 | Period 3 | Period 1 | Period 2 | Period 3 | Period 1 | Period 2 | Period 3 |
| **Rate of return [Pips]** | 185 | 67 | 69 | -41 | 68 | 36 | 56 | 51 | -82 | 141 | 53 | -1 | 231 | -183 | -189 | 26 | -55 | -4 |
| **The number of transactions** | 34 | 38 | 32 | 9 | 11 | 11 | 41 | 126 | 63 | 42 | 51 | 43 | 134 | 201 | 164 | 1 | 1 | 1 |
| **Gross profit [Pips]** | 48 | 16 | 30 | 50 | 28 | 39 | 33 | 20 | 16 | 48 | 20 | 36 | 41 | 32 | 3 | 26 | 0 | 0 |
| **Gross loss [Pips]** | 54 | 43 | 53 | 27 | 27 | 28 | 37 | 21 | 17 | 36 | 32 | 40 | 35 | 28 | 23 | 0 | -55 | -4 |
| **The number of profitable transactions** | 25 | 24 | 22 | 3 | 7 | 6 | 18 | 58 | 19 | 30 | 29 | 26 | 63 | 79 | 67 | 1 | 0 | 0 |
| **The number of profitable consecutive transactions** | 13 | 8 | 6 | 1 | 5 | 4 | 4 | 6 | 3 | 9 | 6 | 7 | 7 | 5 | 6 | 1 | 0 | 0 |
| **The number of unprofitable consecutive transactions** | 2 | 2 | 2 | 2 | 2 | 2 | 4 | 9 | 12 | 3 | 3 | 3 | 5 | 9 | 4 | 0 | 1 | 1 |
| **Sharpe ratio** | 0.84 | 0.93 | 0.91 | 1.01 | 0.97 | 1.00 | 0.94 | 0.84 | 1.20 | 0.84 | 0.92 | 1.0 | 0.3 | 0.02 | 0.7 | 0 | 0 | 0 |
| **The average coefficient of volatility [%]** | 1.12 | 9.13 | 1.12 | 3.34 | 1.95 | 4.70 | 9.85 | 3.20 | 4.11 | 1.01 | 0.74 | 1.29 | 3.33 | 1.95 | 2.56 | 0 | 0 | 0 |
| **The average rate of return per transaction** | 5.44 | 1.76 | 2.15 | 4.56 | 6.18 | 3.27 | 1.36 | 0.40 | -1.30 | 3.36 | 1.04 | -0.02 | 1.72 | -0.91 | -1.15 | 26 | -55 | -4 |
| **Value of evaluation function (y)** | **0.26** | **0.42** | **0.43** | **0.02** | **0.12** | **0.61** | **0.07** | **0.31** | **0.01** | **0.54** | **0.38** | **0.43** | **0.20** | **0.16** | **0.21** | **0.03** | **0.01** | **0,02** |

Source: Experiment results

In general, it may be noticed that the fuzzy logic agents generated not only profitable decisions. In the performance analysis not only the rate of return was taken into consideration but also other ratios, including the level of risk involved in the investment. It may be noticed that the values of efficiency ratios of particular agents differ in each period: for instance the estimated values of such ratios as *Gross Profit* and the *Number of Profitable Consecutive Transactions*. The values of *Rate of Return, Sharpe Ratio* and *Average Rate of Return per Transaction* show significant dispersal among particular agents. It may also be noticed that in the case of the agents *WilliamsFuzzy, TrendLinearRegFuzzy* and *EMA*, the values of these ratios have shown variability in particular periods. The evaluation function provides the immediate choice of the best agent. It may be noticed that the values of the evaluation function oscillate in the range from 0.01 to 0.61.Thus, the use of this function reduces the deviation of the values of the ratios. The results of the experiment allow us to state that the ranking of agents' evaluation differs in particular periods. In the first period, the *ConsensusFuzzy* was the best agent, the *BollingerFuzzy* agent was ranked higher than *WilliamsFuzzy,* and *TrendLinearRegFuzzy* agents were ranked lower than the *EMA*. In the second period, the *BollingerFuzzy* was the best agent, and the *ConsensusFuzzy* and *TrendLinearReg* agents were ranked higher and *WilliamsFuzzy* was ranked lower than the *EMA*. Considering the third period, it may be noticed that the *WilliamsFuzzy* was the best agent and *BollingerFuzzy* and *ConsensusFuzzy* agents were ranked higher and the *TrendLinearRegFuzzy* was ranked lower than *EMA*. The *B&H* benchmark was ranked lowest in all the periods, and in second and third periods it generated the losses. It should be noticed that in the first period, the upward trend was observed, therefore *B&H*'s *Rate of Return* was positive. The second and the third periods shown a downward trend, and therefore the *B&H*'s *Rate of Return* is negative.

Taking into consideration all the periods, it may be stated that there is no agent ranked highest most often. Also, agents achieving the highest Rate of Return were not always ranked in the highest positions. The low level of risk was influenced by the ranks of the *ConsensusFuzzy* and *WilliamsFuzzy* agents. And, on the other hand, the *EMA* was often ranked low because of a high risk level (low value of *Sharpe Ratio*). Moreover, it generated a high number of transactions, so transactions costs are very high.

In the case of fuzzy logic agents, the value of buy-sell decision agents' evaluation is most often higher than the value of *EMA* and *B&H* benchmarks (see last row of Table 1). In the case of three-valued logic agents, instead, there are many cases where the value of buy-sell decision agents' evaluation is lower than the value of *EMA* and *B&H* (see [Korczak et al. 2013, 2014]). Also, the values of such ratios as *Rate of Return* and *Number of profitable transactions* were about several percent higher in the case of fuzzy logic. The risk measuring ratios (*Sharpe ratio, the average coefficient of volatility*) values were similar using fuzzy and three-valued logic.

The fuzzy logic has also demonstrated the better performance of the Supervisor strategies, because the opening/closing positions were generated closer to the optimal point determined by the genetic algorithm. In order to analyse the fuzzy logic agents' decisions efficiency it is also necessary to take into consideration the thresholds for open/close positions determined by the genetic algorithm (Table 2).

The optimal thresholds for opening/closing long/short positions differ in the case of particular agents. However, these levels often do not equal 1, 0 or -1(as in the case of three-valued logic). In addition, the levels for the open long position are different to levels for the close short position, and levels for the open short position are different to levels for the close long position.

TABLE II.
THRESHOLDS FOR OPEN/CLOSE POSITION

| Agent | Open Long | Close Long | Open Short | Close Short |
|---|---|---|---|---|
| BollingerFuzzy | 0.38 | -0.72 | -0.98 | 0.37 |
| WilliamsFuzzy | 1.00 | -1.00 | -0.73 | 1.00 |
| TrendLinearRegFuzzy | 1.00 | -0,88 | -0.89 | 0.88 |
| ConsensusFuzzy | 0.18 | -1.00 | -0.94 | 0.27 |

Source: Experiment results.

Therefore, the fuzzy logic agents may suggest the "out of market" status - in a period of uncertainty on the market - in a broader scope than three-valued logic agents.

## V. CONCLUSION

The fuzzy logic agents in the a-Trader system take independent buy-sell decisions with a certain level of probability. The analysis results presented in this article allow us to draw the conclusion that the application of fuzzy logic as an agents' knowledge representation allows for opening/closing long/short positions closer to the optimal level than the agents based on the three-valued logic.

In consequence the prediction performed by a-Trader were more precise, in periods with both upward and downward trends.

The implementation of fuzzy logic entailed the development of new agents and new trading strategies. The computational complexity of fuzzy logic algorithms is not higher than in the case of three-valued logic, so the computing time of trading positions was almost the same.

It can be also concluded that depending on the current situation on the FOREX market, the level of performance of a particular agent changes. There is no one agent which definitely dominates over the others. The automatic setting of the best agent in time close to real time is performed by the use of the performance evaluation function. It has, in turn, a positive influence on investment effectiveness.

Currently tests are being performed on the implementation of the fuzzy logic agents using fundamental analysis and the analysis of experts' sentiments. It is also planned to evaluate the a-Trader system on more periods and other quotations pairs.

## REFERENCES

[1] H. C. Aladag, U. Yolco and E. Egrioglu, "A new time invariant fuzzy time series forecasting model based on particle swarm optimization", Applied Soft Computing, 12 (10), 2012, pp. 3291-3299.

[2] M. Aloud, E.P.K. Tsang and R. Olsen, "Modelling the FX Market Traders' Behaviour: An Agent-based Approach", in B, *Simulation in Computational Finance and Economics: Tools and Emerging Applications*, . Alexandrova-Kabadjova, S. Martinez-Jaramillo, A. L. Garcia-Almanza and E. Tsang (eds.), IGI Global, 2012, pp. 202-228.

[3] B. C. Arabacioglu, "Using fuzzy inference system for architectural space analysis", *Applied Soft Computing*, 10 (3), 2010, pp. 926–937. doi:10.1016/j.asoc.2009.10.011.

[4] R.P. Barbosa and O. Belo, "Multi-Agent Forex Trading System", in: *Agent and Multi-agent Technology for Internet and Enterprise Systems, Studies in Computational Intelligence*, Volume 289, 2010, pp. 91-118.

[5] J. Bollinger, *Bollinger on Bollinger Bands*, McGraw Hill, 2001.

[6] O. Badawy, O. and Almotwaly, A., "Combining neural network knowledge in a mobile collaborating multi-agent system", *Electrical, Electronic and Computer Engineering*, ICEEC '04. 2004, pp. 325,328, 2004, doi: 10.1109/ICEEC.2004.1374457

[7] L. Chan, A. and Wk Wong, "Automated Trading with Genetic-Algorithm Neural-Network Risk Cybernetics: An Application on FX Markets", *Finamatrix*, 2011, pp.1-28.

[8] M. Y.Chen, "A high-order fuzzy time series forecasting model for internet stock trading". *Future Generation Computer Systems*, 37, 2014, pp. 461-467,

[9] W.M. Cheung and U. Kaymak, "A fuzzy logic based trading system", in: *Proceedings of the Third European Symposium on Nature inspired Smart Information Systems*, St. Julians, Malta. 2007.

[10] M. Dempster and C. Jones, "A Real Time Adaptive Trading System using Genetic Programming", *Quantitative Finance*, 1, 2001, pp. 397-413. doi: 10.1088/1469-7688/1/4/301.

[11] J. Ferber, *Multi-Agent Systems*, Addison-Wesley Longman 1999.

[12] R. Fikes and T. Kehler., "The role of frame-based representation in reasoning". *Commun. ACM* 28(9) 1985, pp. 904-920. doi:10.1145/4284.4285.

[13] S. Franklin and F.G. Patterson, "The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent". in: *Proc. of the Int. Conf. on Integrated Design and Process Technology*, San Diego, CA: Society for Design and Process Science, 2006.

[14] J. B. Glattfelder, A. Dupuis and R. Olsen, "Patterns in High-Frequency FX Data: Discovery of 12 Empirical Scaling Laws", *Quantitative Finance*, Volume 11 (4), 2011, pp. 599-614.

[15] M. Hernes M. and N.T. Nguyen, "Deriving Consensus for Hierarchical Incomplete Ordered Partitions and Coverings", *Journal of Universal Computer Science* 13(2) 2007, pp. 317-328.

[16] M. Hernes M. and J. Sobieska-Karpińska , "Application of the consensus method in a multiagent financial decision support system", *Information Systems and e-Business Management*, Springer Berlin Heidelberg 2015, doi: 10.1007/s10257-015-0280-9.

[17] M. A. Kadhim, A. Alam, M. and K. Harleen, "A Multi-intelligent Agent Architecture for Knowledge Extraction: Novel Approaches for Automatic Production Rules Extraction", *International Journal of Multimedia & Ubiquitous Engineering*, Vol. 9 Issue 2, 2014, p.95.

[18] O. Kazar, H. Ghodbane, M. Moussaoui and A. Belkacemi, "A Multi-Agent Approach Based on Fuzzy Logic For a Robot Manipulator" *JDCTA* 3(3), 2009, pp. 86-90.

[19] R. Karjalainen, "Using Genetic Algorithms to Find Technical Trading Rules", *Journ. of Financial Econ.*, 51, 1999, pp. 245-271.

[20] C. D. Kirkpatric and J. Dahlquist, *Technical Analysis: The Complete Resource for Financial Market Technicians*, Financial Times Press, 2006.

[21] J. Korczak and P. Lipinski, „Systemy agentowe we wspomaganiu decyzji na rynku papierów wartościowych", in *Rozwój informatycznych systemów wieloagentowych w środowiskach społeczno-gospodarczych*, ed. S. Stanek et al., Placet, 2008, pp. 289-301.

[22] J. Korczak, M. Bac, K. Drelczuk and A. Fafuła, "A-Trader - Consulting Agent Platform for Stock Exchange Gamblers", in *Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS)*, Wrocław, 2012, pp.963-968.

[23] J. Korczak, M. Hernes and M. Bac, "Risk avoiding strategy in multi-agent trading system", in *Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS)*, Kraków, 2013.

[24] J. Korczak, M. Hernes and M. Bac, "Performance evaluation of decision-making agents' in the multi-agent system", in *Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS)*, Warszawa, 2014, pp. 1171 – 1180. doi: 10.15439/2014F188.

[25] B. LeBaron, "Active and Passive Learning in Agent-based Financial Markets", *Eastern Economic Journal*, vol. 37, 2011, pp. 35-43.

[26] J. Lagorse, M.G. Simoes, and A. Miraoui, "A Multiagent Fuzzy-Logic-Based Energy Management of Hybrid Systems," *Industry Applications, IEEE Transactions on*, vol.45, no.6, pp.2123,2129, Nov.-dec. 2009, doi: 10.1109/TIA.2009.2031786.

[27] C. Lento, "A Combined Signal Approach to Technical Analysis on the S&P 500", *Journal of Business & Economics Research* 6 (8), 2008, pp. 41–51.

[28] S. Martinez-Jaramillo and E.P.K. Tsang, "An Heterogeneous, Endogenous and Co-evolutionary GP-based Financial Market", *IEEE Transactions on Evolutionary Computation*, Vol.13, No.1, 2009, pp.33-55.

[29] N. T. Nguyen, "Using Consensus Methodology in Processing Inconsistency of Knowledge", *in Advances in Web Intelligence and Data Mining, series Studies in Computational Intelligence*, M. Last et al. (Eds): Springer-Verlag, 2006, pp. 161-170.

[30] V. Novák, I. Perfilieva and J. Močkoř, "Mathematical principles of fuzzy logic", *Dordrecht,* Kluwer Academic, 1999.

[31] D. A. Oyemade, O. Godspower, O. Ekuobase and F. O. Chete, "Fuzzy Logic Expert Advisor Topology for Foreign Exchange Market", *Proceedings of the International Conference on Software Engineering and Intelligent Systems*, Ota, Nigeria, 2010.

[32] I. Palit, S. Phelps and W. L. Ng, "Can a Zero-Intelligence Plus Model Explain the Stylized Facts of Financial Time Series Data?", in *Proceedings of the Eleventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)* - Volume 2. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 653–660.

[33] M. Piunti and A. Ricci, "Cognitive Use of Artifacts: Exploiting Relevant Information Residing in MAS Environments", in *Knowledge Representation for Agents and Multi-Agent Systems*, J. Ch. Meyer, J. Broersen (eds.),Lecture Notes in Computer Science 5605, Springer Berlin Heidelberg, 2009 pp. 114-129, doi: 10.1007/978-3-642-05301-6_8.

[34] J. Ropero, A. Gómez, A. Carrasco and C. León, "A Fuzzy Logic intelligent agent for Information Extraction: Introducing a new Fuzzy Logic-based term weighting scheme", *Expert Systems with Applications*, Volume 39, Issue 4, 2012, pp. 4567-4581, http://dx.doi.org/10.1016/j.eswa.2011.10.009.

[35] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach* (2 ed.), Pearson Education, 2003.

[36] P. Singh and B. Borah, "Forecasting stock index price based on M-factors fuzzy time series and particle swarm optimization", *International Journal of Approximate Reasoning*, 55(3), 2014, pp. 812-833,.

[37] S.Srinivasan, D. Kumar and V. Jaglan, "Agents and their knowledge representations", *Ubiquitous Computing and Communication Journal*, 5(1), 2010, pp. 14-22, doi: 10.1.1.295.6890.

[38] TrendLinearReg,http://forexwikitrading.com/forex-indicator/trendlinearreg/ [access: 2014.02.02].

[39] X.L.Chen, L. M. Li, Y.Z. Wang, W. Ning and X. Ye, "ERPBAM: A Model for Structure and Reasoning of Agent Based on Entity-Relation-Problem Knowledge Representation System," *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09. IEEE/WIC/ACM International Joint Conferences*, vol.3, 2009, pp. 365,368.doi: 10.1109/WI-IAT.2009.302

[40] X.F. Zhang, G.J. Wang and G.W. Meng, "Theory of Truth Degree Based on the Interval Interpretation of First-order Fuzzy Predicate Logic Formulas and Its Application", *Fuzzy Systems and Mathematics*, 2006,20(2), pp. 8-12.

[41] X.Z. Wang and S.F. An, "Research on learning weights of fuzzy production rules based on maximum fuzzy entropy", *Journal of Computer Research and Development*, 43(4), 2006, pp. 673-678.

[42] G.J. Zhu and Y.M. Xia, "Research and practice of frame knowledge representation", *Journal of Yunnan University (Natural Sciences Edition)*, 28(S1), 2006, pp. 154-157.

[43] Z. Zeng, "Construction of knowledge service system based on semantic web", Journal of The China Society For Scientific and Technical Information, 24(3), 2005, pp. 336-340.

[44] J. Martin and J.J. Odell, *Object oriented methods: the foundations*, Englewood Cliffs, Prentice Hall, 1994.

[45] L. Valiant, *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*, New York: Basic Books, 2013.

[46] S.P. Li, Q.W. Yin, Y.J. Hu et al, "Overview of researches on ontology" *Journal of Computer Research and Development*,41(7), 2004, pp. 1041-1052.

[47] L.A. Zadeh, *Fuzzy Sets, Fuzzy Logic, Fuzzy Systems*, World Scientific Press, 1996.

[48] M. Żytniewski, R. Kowal, A. Sołtysik, "The Outcomes of the Research in Areas of Application and Impact of Software Agents Societies to Organizations so far. Examples of Implementation in Polish Companies", [in] Annals of Computer Science and Information Systems, Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS), Kraków, 2013 pp. 1165 – 1168.

DOI: 10.15439/2015F282
ACSIS, Vol. 5

# Secure service interaction for collaborative business processes in the inter-cloud

Björn Schwarzbach, Michael Glöckner,
Alexander Pirogov, Martin Max Röhling
Leipzig University
Grimmaische Straße 12, 04109 Leipzig, Germany
Email: {schwarzbach, gloeckner, pirogov,
roehling}@wifa.uni-leipzig.de

Bogdan Franczyk
Leipzig University
Grimmaische Straße 12, 04109 Leipzig, Germany
Uniwersytet Ekonomiczny we Wroclawiu
ul. Komandorska 118/120, 53-345 Wroclaw
Email: franczyk@wifa.uni-leipzig.de

***Abstract*—The emergence of a closer relationship between cloud service providers in the cloud computing market is the inevitable consequence of the computing as utility concept. The closer cooperation creates competitive advantages for providers and users of cloud services as well. Capacities and services can be used in a collaborative and flexible way. Despite the numerous potentials of composite cloud services, trust, policy and privacy are the major challenges resulting from the distributed and flexible data handling. The paper derives requirements and solutions in the field of inter-cloud service communication with a special focus on security. The proposed architecture is evaluated with a sample collaborative business process of inter-cloud service interaction.**

## I. Introduction

IN 2008 vice president of Gartner Research Thomas J. Bittman published his thoughts on future development of Cloud Computing (CC). In the early monolitic phase, cloud services were built on proprietary architectures of dominant Cloud Service Provider (CSP), such as Google, Salesforce or Microsoft. The second phase vertical supply chain distinguishes itself by the development of first ecosystems of smaller companies within the CC market. New CSPs use proprietary Cloud platforms of dominant providers, i.e. Google App Engine or Microsoft Azure, in order to provide their own services. In the last phase, smaller providers unite to a horizontal federation. That way, the union increases earnings by expanding their capacity and reducing costs through more efficient resource allocation. In parallel, open interoperability standards of service communication in intercloud-environment are developed [1].

The creation of a closer relationship between CSPs on the CC market is a inevitable extension of the computing-as-utility concept, which is about providing computing resources as a service over the Internet. Users of cloud based services may benefit in terms of cost reduction by renting their own distributed, virtualalized IT-infrastructure, improving service robustness and preventing provider dependence by means of interoperability standards [2], [3], [4].

A close cooperation creates certain advantages in competition for providers of cloud services. By using other CSPs' capacities, providers may deliver their products and services even faster and more effective to their clients. Further, making use of virtualization technology reduces costs for a flexible and customizable IT-infrastructure. Its dynamic and smooth scaling has a positive effect on service deployment time. Due to dynamic outsourcing of computational services, power consumption costs for computer centres can be significally reduced [3], [5].

Beside several advantages of collaborative cloud services, cloud specific issues concerning the security of service communication in an intercloud environment still exist. Trust is an essential precondition in order to create an intercloud federation. Without trust, security of cloud-interactions can't be guaranteed. Policy is another issue concerning intercloud interactions. It is essential to have effective control mechanisms so potential policy clashes, which would affect the safety of the whole system, are detected and removed. Identity and data privacy are other challenges to intercloud communication since users of cloud services transfer their personal data to the CSPs. Appropriate tools for access and identity management are essential for data protection [2], [6], [5], [7].

The paper focuses on the creation of an intercloud architecture, which enables secure service communication in collaborative business processes. The second chapter consists of an overview of relevant theoretical concepts while the next chapter introduces and explains a collobarative process of payment transaction. Communication models of services in the Intercloud environment are analyzed in chapter four. Further, chapter five introduces a draft architecture and suggestions for its implementation. Finally, the conclusion completes the paper.

## II. Theoretical background

The following chapter deals with the theoretical basis of service communication in the intercloud environment. Special attention is paid to security aspects of service interactions in collaborative business processes. Table I gives a resume of criteria for safe service communication that will be elaborated in this chapter.

The work presented in this paper was funded by the German Federal Ministry of Education and Research under the projects PREsTiGE (BMBF 16KIS0082K) and LSEM (BMBF 03IPT504X).

978-83-60810-66-8/$25.00©2015, IEEE

TABLE I
CRITERIA FOR SECURE COMMUNICATION IN THE INTERCLOUD

| **Criterion** | **Challenge** | **Solution** |
|---|---|---|
| Interoperability | Diverse, partly proprietary communication protocols | Broker; standardized interfaces; hybrid approach |
| Robustness | Single point of failure; workload balancing | Distributed implementation of the IT infrastructure |
| Optimal service provisioning, service selection, and service allocation | Orchestration of services and dependency resolution in real-time; automated selection of QoS-Criteria | Central platform that fulfills the orchestration in the intercloud; mechanisms for conflict resolution in case of conflicting QoS policies |
| Access & identity management | Management of multiple accounts of service users and CSPs; Establishment of a secure trust context for service interaction | Outsourcing of credential management to third parties; digital identities; Identity federation with the use of protocols such as Security Assertion Markup Language (SAML), eXtensible Access Control Markup Language (XACML), OpenID, OAuth, WebID, and SSO; secure authentication methods; multi factor authentification |
| Trust | Establishment of a secure trust context; dynamic determination of trust; de-perimeterization | PKI, XACML and SAML based communication protocols; reputation based, dynamic trust index |
| Policy | Inconsistency; inefficiency; semantic interoperability; static, predetermined SLAs, that hinder the ad-hoc service interaction | Mechanisms for the combination of policies on cloud federation level; monitoring and conflict resolution; dynamically, automatically, and instantaneously created federation-level agreements |
| Privacy | data privacy; identity privacy | Encryption; anonymization; pseudonymization |

### A. *Intercloud environment*

The term *intercloud* is not standardized in scientific literature. Though terms like *cross-cloud* [5], *multicloud* [7] or *cloud-federation* [2] can be found, but summarizing author-specific definitions do not differ basically.

Our work is based on the definition of the Global Inter-Cloud Technology Forum: "A cloud model that, for the purpose of guaranteeing service quality, such as the performance and availability of each service, allows on-demand reassignment of resources and transfer of workload through an interworking of cloud systems of different cloud providers based on coordination of each consumer's requirements for service quality with each providers SLA and use of standard interfaces." [8]

If not explicitly stated different, our work uses the term *intercloud*. In the authors' opinion intercloud is a better definition for a multicloud environment since it deals with a highly integrated environment where service communication is structured by the use of coordinating instances. Fig. 1 describes a typical intercloud environment with different types of Clouds. Closer cooperation of CSPs enables the usage of different strategies for resource consumption like outsourcing and cloud bursting. Service users are either the end consumers or other CSPs as well.

By combining different CSPs' services, the problem of being dependent on one provider, so called lock in effect, is solved in the intercloud environment. Moreover, flexibility as well as scalability and robustness of the whole system can be improved, because all intercloud CSPs are able to provide identical services. Further, energy can be used more effectively [3], [4].

The heterogeneity of this environment is a special challenge for technical implementation on all levels of intercloud architectures. The cloud-spanning integration of services emphasizes the importance of availability and access speed of services. Interoperability of interclouds is reduced by proprietary interfaces that are needed for service interaction.

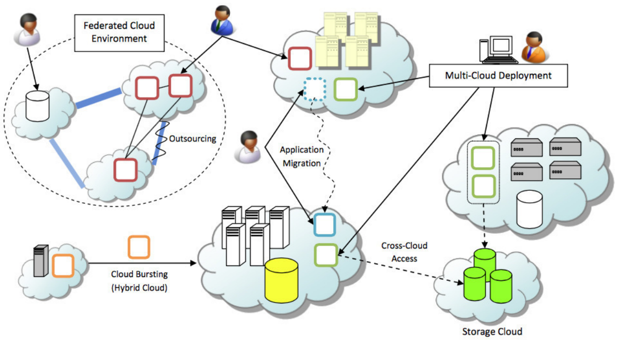


Fig. 1. Cloud interoperability [7]

In addition, the current service delivery model is incompatible with an open and dynamic collaboration, so using customer-specific tools leads the cloud interoperability to a vendor lock-in.

Due to limited access rights of unauthorized users and the bundling of provided services with other resources of the same provider, personalized service customization and cloud-spanning service composition are affected [6], [5], [4].

A secure collaboration of a multitude of CSPs in a heterogeneous environment is only enabled by complex intercloud architectures that meet several requirements. Fig. 2 provides an overview of intercloud architecture challenges. Several ideas of Toosi haven been applied [7]. This paper focuses mainly on the highlighted security relevant aspects.

From a technical point of view, interoperability can only be accomplished through a broker, which is in control of all communication between the CSPs, or via standardized interfaces. It's possible to use a hybrid of both ideas, depending on whether it is economically and technically useful. The

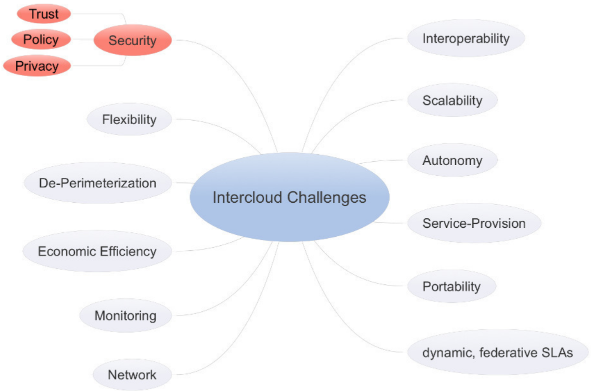


Fig. 2. Challenges of intercloud architectures

last option consists of an adaption of the architecture by all participants, such as the CSPs and the service users. A better option could be the introduction of a central platform, which would reduce the need for adaptation when joining the intercloud federation and which would provide a uniform interoperable vision of service communication.

### B. Cloud based collaborative business process

A collaborative business process is a dynamic process, run by several involved instances. All steps are conducted locally by interacting work-flow engines, following a common service definition. The engines of the interacting process partners are staying in control of all of their subprocesses [9].

### C. Trust

Due to de-perimeterization, trust is a very important factor in the intercloud environment. Since resources are outsourced to the CSPs, service users lose control over their personal data and that's why CSPs need to be trustworthy and service consumers need to be equipped with effective control mechanisms [6].

The transitive trust is essential when giving authorization to a third party. First, mechanisms of delegation have to allow a dynamic creation of service level agreements (SLA) in order to deny access to unauthorized user. Second, third parties must not modify service requests arbitrarily. Third, delegation has a time limit so service user representatives may only act in a certain time period legitimately. Another intercloud communication problem is the creation of a trust context between the interacting CSPs. Current intercloud trust models are based on a public-key-infrastructure-system (PKI), which judges the trustworthiness of entities in an absolute way and is therefore not suitable for a dynamic certification. The reputation-based trust complements the trust-context creation with dynamic aspects. When using parameters of reputation, utilizing relatively trustworthy computing resources in the intercloud is given, if they match with practical experience values of participating CSPs. Hence, it is possible to create a real trust federation along with third parties [10], [7].

Federated identity management is very important in the intercloud environment. It is possible to create an identity federation, which will support the Single-Sign-On (SSO) approaches and the management of digital identities by using formalized internet protocols like SAML and XACML specifications as well as open-source ones like OpenID, Oauth and WebID. This results in taking charge from interacting CSPs by outsourcing the credential management [5].

Literature addresses other dynamic and scalable approaches of confidentiality. The friend of a friend approach is interesting, since it provides machine-readable ontologies for object description. Despite of security concerns, the approach is flexible and executable without a centralized database [7].

### D. Policy

Policies are guidelines for the execution and monitoring of the interactions between the CSPs. Because of its inconsistency and incompatibility, policy-heterogenity causes the main security risk in an intercloud environment. Conflicts arise from a collision of local CSPs' policies and federated policies.

Policy-inconsistencies especially arise from a collision of access guidelines in distributed environments with a multitude of interacting entities. Different policies may be contradictory to each other if they have different effects on entities and their attributes. An exception occurs when two policies affect the same instance differently while being hierarchically linked. A policy-correlation leads to a partial conflict, where two overlapping policies treat one certain object differently, but only one of the two policies allows overlapping.

In an intercloud environment, policies can be collected in a list of guidelines, which could lead to significant reduction of the communication performance of access authorization. Firstly, policy-redundancies may occur in a way that one request is mapped to various policies with identical effects on one and the same object. Secondly, when combining policies, attention has to be paid to semantical and syntactical correctness of the new federated policy [6].

The SLAs complete the IT policies with legal aspects and are essential for policy-compatible intercloud interaction. At the moment, SLAs are limiting the dynamic intercloud communication because they reduce the flexibility of CC business models. More complex SLAs, which possess powerful management and monitoring tools, are essential for proper legal data processing [6], [7].

Another important policy aspect is quality of service (QoS). It helps clients to choose an appropriate service. The mechanisms of service selection have to be able to harmonize various rivaling and maybe excluding QoS objectives [2], [11]. The incorporation of QoS factors makes an intercloud system more flexible, customer-oriented and eventually more attractive to potential user.

### E. Privacy

Privacy is a strongly by legal restrictions influenced concept, which assures control over information and information flows and restricts access for illegitimate entities. Different laws (i.e. European and American) apply to the term Privacy in different ways, especially in terms of sensibility of personal data and legal precautions. Countries of the EU apply the data minimization concept: no access to personal data unless absolutely necessary. Service users may explicitly prohibit the usage of their personal data for advertising purposes. The USA do not have an equivalent to those legal restrictions [12], [13]. Therefore, it is extremely important to compare and adjust definitions, especially in terms of trans-regional interclouds.

The technical realization of these privacy requirements is a complex issue. It is necessary to distinguish between two basic privacy-strategies: data privacy and identity privacy. Data-privacy strategies consist of altering the content beyond recognition so data cannot be used by third parties easily. Data-perturbation completes original with *noise data* which subsequently becomes unreadable for non-legitimate instances. Unfortunately, the resulting redundancies may cause scalability problems in the intercloud. By using compression methods, cost for communication and request-handling can be cut. Perturbation approaches have to be flexible in order to find a compromise between the user and the privacy guarantee [13], [6].

Encryption by data transformation is one common privacy method. Besides several advantages, this method causes restrictions in the intercloud environment, from a service-communication point of view. First, data-utility aspect plays a more important role than the safety aspect. Second, *data at rest encryption* blocks data indexing and data search. Finally, no efficient methods for operation of data at transit are developed yet [13].

Identity privacy strategies aim at hiding the real identity of interacting instances from unauthorized user. By using anonymity, one CSP can authenticate a user without revealing his true identity. Unlinkability hinders CSPs from identifying users by using a transaction portfolio [14].

## III. Collaborative sample process

After providing basic information about intercloud environment in chapter II, an example is introduced in the following section. Building up on the example, chapter V will deal with the implementation of an architecture.

### A. Scenario

In the last few years providers like Google and Apple have contributed their solution of mobile payment to the market with Google Wallet and Apple Pay. Despite being two of the most successful companies, support by commerce for these providers is still missing. A better solution consists of retailer and customer are working with a neutral and flexible intermediary, i.e. payment provider.

This example deals with the scenario of electronic payment via smartphone. Mobile payment comprises customer authorization and realization of payment via smartphone. Instead of paying via credit card, the smartphone has to be put on a terminal to initiate the payment. Through near field communication (NFC) two electrical devices, in direct physical proximity, are able to exchange data. In this case, the payment amount is authorized via smartphone. After having put the smartphone on the terminal, the central CSP is contacted and requested for user identification of the smartphone. Afterwards, the user has to authorize the payment via finger print or PIN. When successfully matched, the transaction will be initiated by the central CSP. In order to explain which accounts are involved and which internal processes are initiated in the Central CSP, the process shown in fig. 3 will now be explained in more detail.

### B. Coupling of customer's account and payment provider

In this scenario, the CSP acts as an intermediary. The CSP does not have an account and cannot transfer any payments, it only delivers the payment order to the payment provider. Payment providers do not necessarily have to be one and the same nor even similar. Examples of payment providers may be PayPal, Visa or the German website sofortueberweisung.de. Payment provider have to be added to the central CSP in advance. This process is called linking and should ideally only be executed once. While linking, the user is authenticated at the provider and receives two tokens (access and refresh token). The client, in this case the central CSP, uses these tokens to authenticate future orders at the payment provider. After the linking, all payment providers are available via the CSP service and account. If a customer wants to use the payment service of the central CSP, only the authentication to the CSP is needed.

### C. Service discovery and execution of the transaction

Fig. 3 provides a detailed overview of the introduced process so all internal steps of the central CSP are visible. After placing the smartphone on the terminal and after the successful authorization, an automatic selection of a payment provider follows, according to rules and settings the customer has set before. Useful rules involve e.g. minimal transaction costs. Afterwards, the payment is initiated at the provider's side.

If the transaction failed, another provider will be chosen according to the rules set by the customer until the transaction is finally successful. This confirmation will be forwarded to the retailer in order to print the receipt. It is important to mention that the BPMN model is only constructed for a positive result. The exception of not finding an appropriate payment provider, due to strict selection criteria or only negative responses, has to be considered in future work.

### D. Evaluation

A big advantage in the introduced scenario is that the central CSP is provider neutral. If the CSP integrates new providers, customers and retailers will benefit. Further, the CSP is able to provide its orchestrated service as a SaaS in the intercloud federation.

## IV. Service Communication Models in heterogeneous Intercloud Ecosystems

The selected communication models of services provide a comprehensive conceptual view on the intercloud communica-

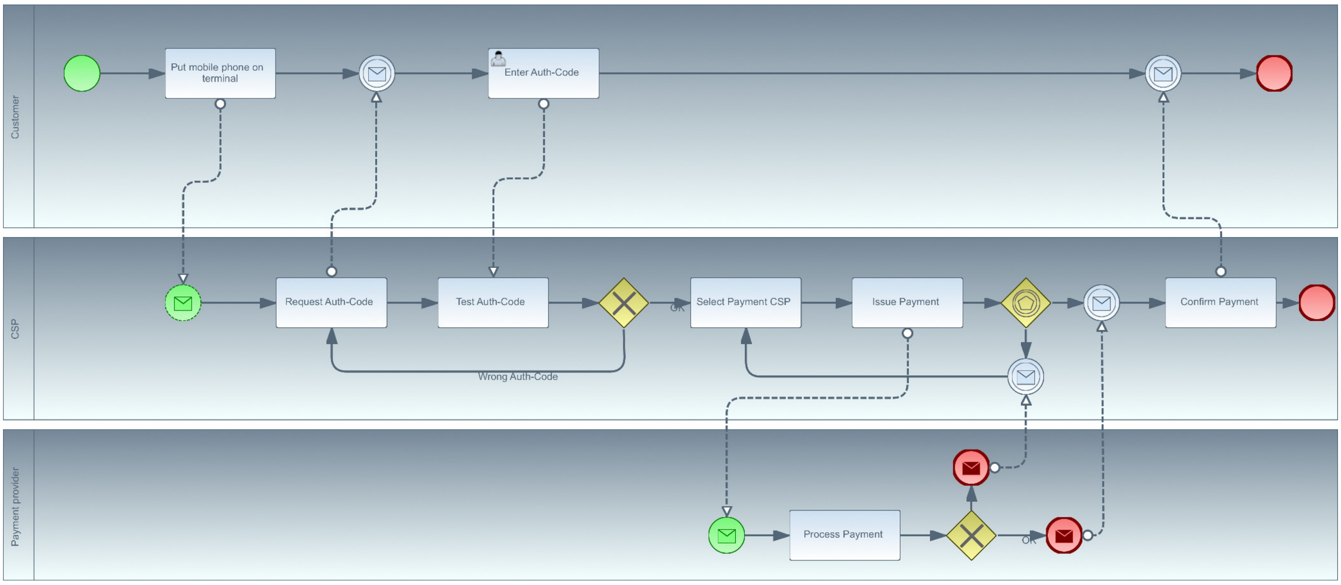


Fig. 3. Collaborative sample process

tion and provide a basis for our proposed architecture. Since some intercloud models, e.g. Demchenko and Makkes [15], [16], [17] and Lloret [4], are well known but focus on special areas of CC, e.g. Demchenko and Makkes focus on IaaS, we present only those models that are relevant for our work.

### A. Intercloud Domain-based Trust Model by Bernstein

The intercloud roots (IR), intercloud exchanges (IE), intercloud gateways (IG), and cloud computing resource Catalogs (CCRC) provide the basis of the domain-based trust model proposed by Bernstein [10]. Fig. 4 provides a comprehensive view of all components as well as the corresponding technologies and protocols.

The IR handle the broker functions to operate the service communication. They are structured hierarchical and self reproducing like nodes in peer-to-peer networks. IRs act as security trust service providers, handle the namespace, dynamic naming of the intercloud, and host the distributed CCRCs [19], [18].

The communication and collaboration of the heterogeneous intercloud environment is supported by the IEs, which utilize the information of the CCRC to provide an optimal computation resource matching. IEs act as the trust agents of one domain by collecting information of the confidentiality of other domains and providing the trust level of a domain while initiating the interaction between the domains [10].

The IGs provide authentication mechanisms and standards and protocols for interoperability. Another responsibility of the IGs is to check for availability of resources, the state of interactions and to transform the parameters of the communication from one cloud to another [10].

The CCRC provides a holistic and abstract view on computing resources in the intercloud federation. It enables the resource adjustment between individual CSPs on the basis of selected preferences and constraints. The catalog stores information on available resources, interoperability standards and guidelines, and SLAs [18], [10].

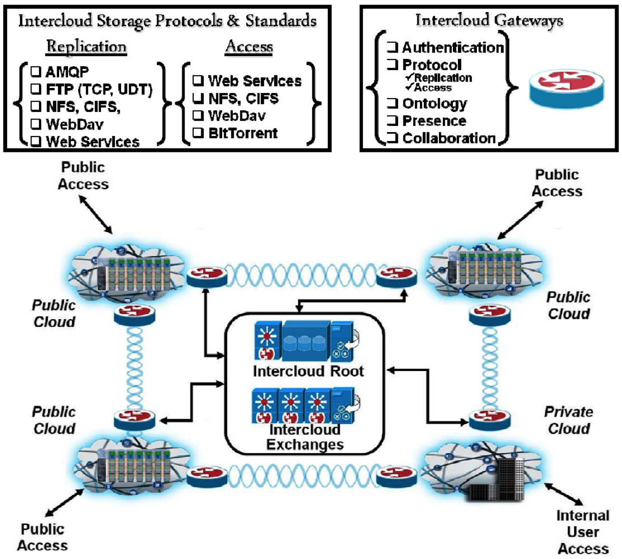


Fig. 4. Intercloud domain-based trust model by Bernstein [18]

The trust management of the intercloud consists of two components, the PKI provides services for the use of trust certificates and trust chains. The fixed-term certificates for short-term transaction are issued by IEs, the long-term transactions are certified by IRs. The PKI is of limited suitability for the certification of processes in the intercloud, since it classifies entities either as trustworthy or not [10].

CSPs are differentiated into confidentiality domains based on the dynamic, time-dependent trust index. The trust index

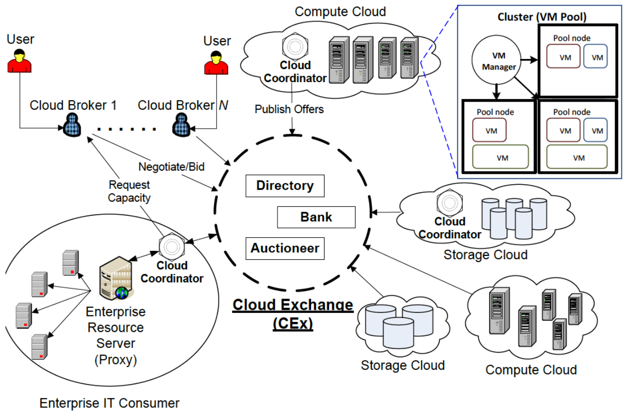


Fig. 5. Utility-oriented federation of clouds by Buyya [2]

comprises information on the CSP and its reputation. The application of the trust index enables the consumption of relatively trustworthy resources that may be located in another domain with a lower trust index [10].

### B. Utility-oriented Federation of Clouds by Buyya

Buyya proposed an intercloud architecture that focuses with economically efficient resource matching in the cloud market. The major components of the model are cloud coordinator (CCr), cloud broker (CB), and cloud exchange (CEx) as shown in fig. 5. The CCr, which is integrated into the CSPs' infrastructure, manages domain-specific clouds. It provides a programming, management, and installation environment for applications in the federation of clouds. The CCr acts as a proxy for communication between the participating CSPs and manages all corresponding processes for information exchange [2].

The CCr's service matching process is deeply influenced by economic aspects, e.g. energy consumption, heat output, and node utilization of virtualized cloud environments. It is also influenced by the QoS optimization problem that the service consumer may have selected multiple conflicting QoS objectives. The CCr resolves such conflicts by applying heterogeneous optimization algorithms [2].

The CEx manages all service requests and offers of the cloud federation and supports a SLA based matching process. Further it supports the exchange of information on the current state of allocated resources between the individual CSPs. Given that all participants provide standardized interfaces a directory service enables the CE to lookup desired service offers or requests. The trading process is managed by a dynamic bidding based service, which provides a trustful auctioneer and takes care of offer updates in the cloud federation. The payment management is implemented by an autonomous banking entity, which enforces the agreements on financial transactions of the global CExs. To ease the handling of financial transactions the integration of cloud based accounting systems with existing online payment systems, e.g. PayPal, is taken into consideration [2].

Finally, CB identifies appropriate cloud services on behalf of the users and agrees on the QoS based resource reservation with the CCrs. If a cloud is not able to process an incoming request locally, CB creates a new query comprising QoS information and forwards this query to the CE. Essential conditions for the successful completion of the matching process are: feasibility of QoS targets specified by the user and the equal resource distribution to the individual nodes [2].

### C. Cross-Cloud Federation by Celesti

Celesti proposes a process oriented ad-hoc cross-cloud federation, which comprises three phases: in the discovery phase a cloud looks for available resources of other clouds. Afterwards the most appropriate CSP is selected during the matching phase. Finally the authentication phase comprises establishment of a trust context for secure communication of the interacting clouds. The model distinguishes two types of clouds: the home cloud requests compute resources that are offered by the foreign cloud [5].

Fig. 7 provides an overview of the entities and components involved in creating the intercloud federation. In order to form a cross-cloud the internal architecture of the participating cloud must be converted to the following 3 layer structure, first: virtual machine manager (VMM), virtual infrastructure manager (VIM), and cross-cloud federation manager (CCFM) [5].

VIM is a dynamic orchestrator for virtual environments and enables the creation, installation, and management of virtual environments regardless of the underlaying technology. If the home cloud is not able to instantiate another virtual machine due to a lack of additional resources, it forwards the request to the cloud federation. Subsequently the VIM selects a suitable foreign cloud. To securely share and transfer resources with each other the VIMs of the individual clouds create a trust context with each other [5].

The CCFM orchestrates the three phases of the cross-cloud federation. The discovery agent supports the discovery process of the dynamic intercloud environment by publishing and updating the information on offered cloud services on behalf of the individual cloud [5].

The match-making agent makes authorization decisions based on policies and performs compatibility testing of communication policies of interacting clouds. In case of incompatible policy languages the match-making agent applies some transformation algorithms [5].

The authentication agent provides credential management to the cross-cloud federation and coordinates the creation of a security context for the cross-cloud communication using identity providers. By applying SSO after authenticating once with the system the services can be consumed without additional security checks. Using digital identities, which are provided by external identity providers, home clouds can connect to any foreign cloud [5].

The security context in cross-cloud federation is created in two layers: on authentication agent layer and on VIM layer.

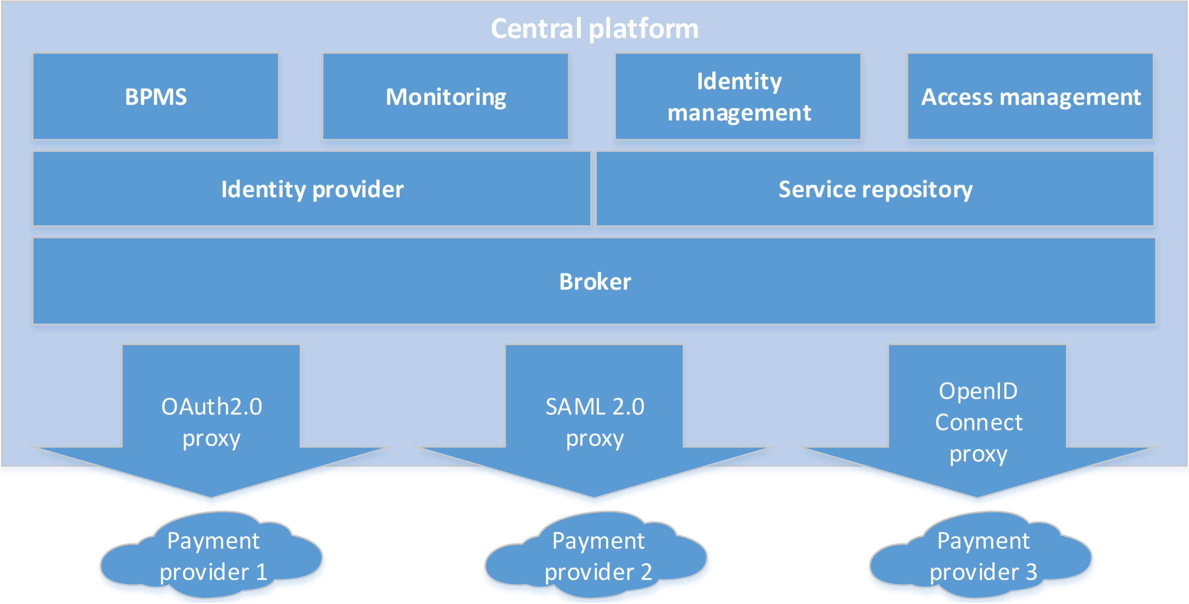


Fig. 6. Intercloud architecture for collaborative business processes

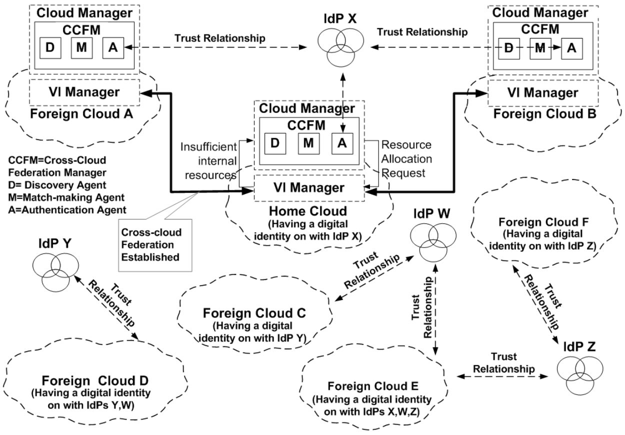


Fig. 7. Cross-Cloud federation by Celesti [5]

The latter act on a lower level compared to the CCFM and facilitate cross-cloud resource provisioning [5].

### D. Proxy-based Computing Cloud Framework by Singhal

Singhal proposes another way of tackling the secure communication problem in intercloud environments. In this model the proxy is the central component, which establishes secure communication without predetermined agreements or standardized interfaces and specification [6].

A proxy is an intermediary for traffic between interacting clouds and provides a confidential and trustworthy computing platform, which protects the user data and cloud applications from malicious attacks. The strategic proxy deployment is crucial for efficiency. Thus, in the proxy selection criteria such as latency and forecasted load are taken into account [6].

Since the proxies act on behalf of the users or CSPs, a mechanism for secure delegation is necessary. This mechanism includes dynamic SLA creation, prohibit the proxy access on the processed data, as well as protection against malicious proxy interaction, e.g. modification of transactions. Further a proxy is only allowed to act on behalf of an entity as long as its task is not completed. Such secure delegation can be realized by warrant-based-proxy signatures, PKI or OAuth protocol. Proxies monitor and eliminate conflicts based on policy heterogeneity while processing the service request to avoid potential security risks [6].

A cloud service can be accessed by multiple proxies, these proxies can be assigned different, specific roles. They operate automatically and need no further interaction with the commissioned instance. This interaction is realized by allowing proxies to communicate with other proxies and to initiate necessary service request by themselves. Hence, secure collaboration of multiple CSPs without predetermined agreements is possible [6].

The proxy based model features five types of intercloud architectures: cloud-hosted proxy, proxy as a service, peer-to-peer proxy, on-premise proxy, and hybrid proxy infrastructure. The individual scenarios differ in the implementation strategy of proxy instances.

In the cloud-hosted proxy scenario proxies are integrated into the infrastructure of the individual CSPs. This approach is advantageous from the CSPs point of view, because the CSPs maintain control over the proxies and can adjust the proxies to specific needs [6].

The proxy-as-a-service scenario proposes an autonomous proxy cloud, which provides proxies to the CSPs. Users can create an account either on the proxy cloud or the CSP's cloud. In both cases resource requests are handled by the proxy [6].

The peer-to-peer-proxy scenario organizes the proxies in a P2P-manner. The proxies are managed collectively by the CSPs or proxy providers or autonomously by the proxy nodes [6].

In the on-premise scenario proxies are part of the client's infrastructure. Users can submit their request either via a proxy or directly to the CSP. If the provider is not able to provide the requested resource this will be forwarded to the proxy that will continue the service discovery process with other CSPs [6].

If some or all of the previous variants are combined this is called hybrid-proxy infrastructure. The proxy selection may depend on the type of the requested service and infrastructural peculiarities of the CSP. This approach provides the greatest flexibility in term of intercloud creation [6].

## V. Architecture for secure intercloud communication

In this section we propose an architecture for secure intercloud communication in terms of trust, policy, and privacy and evaluate a prototype of this architecture with the sample process of chapter III.

### A. *Intercloud architecture for the sample process*

Basically the collaborative communication of the services is organized by a central platform, which applies concepts of the work shown in the previous section. To enable the communication with the payment providers of the sample process, we propose to use proxies to bind the payment services to the platform. Such a modular handling is necessary as there is no common standard for authentication in CC. Fig. 6 shows the main components of the architecture for secure collaborative business processes. OAuth 2.0, SAML 2.0, and OpenID Connect are the selected proxies that we have implemented as a proof of concept because these are the most widely used authentication standards in CC.

The components and a short description of their tasks is shown in table II. The connection of services through proxies enables flexible scaling of the platform. Another advantage is that the maintenance of the communication infrastructure remains in the CSPs' responsibility. In addition to the integration of the payment process into existing service ecosystems the payment process can be provided as a separate service that can be used by multiple CSPs. The evaluation environment assumes the integration of a payment terminal at the merchants, but there are far more scenarios possible. Since the service is part of payments it requires very high security standards. Hence, sophisticated security components need to be incorporated into the platform. The identity and access management system ensures it is clearly decidable and accountable who is authorized for performing certain actions with the resources, e.g. who is allowed to issue a payment. Such a system needs to authenticate with other system. Hence, protocols for authentication and authorization need to be supported. The most widely used protocols for authentication and authorization are shown in table III.

TABLE II
Core components of the architecture

| Kernkomponente | Funktion |
|---|---|
| BPMS | Business process management system to model, instanciate, and collaborative business process on the platform |
| Monitoring | Monitoring of the whole system |
| Identity Management | Secure and connect third party identities to the local identities; management of local identities |
| Access Management | Access policy enforcement and policy conflict handling |
| Service Repository | List of available services and their descriptions |
| Broker | Service selection and QoS |
| Proxy | Invocation of services |

The prototype we have implemented and evaluated for our research implements OpenID Connect. There are two reasons. On the one hand OpenID Connect is a protocol for authentication and authorization that is very new but already adopted or will be adopted in the very near future by plenty of the big CSPs, e.g. Google, Microsoft, and PayPal. On the other hand there are two kinds of protocols: centralized and federated ones. Since federated ones are more complex in terms of communication flow and the prototype should be recognized as a proof of concept we chose the federated OpenID Connect protocol.

When a business process is instantiated by the BPMS and the BPMS needs to perform a task of the business process, the broker selects an appropriate service of the service repository. The broker looks up the type of the service's invocation pattern and instantiates the right type of proxy. Then the proxy performs the actual service call and sends the data back to the broker. Prior to sending the user data to the proxy the broker applies privacy rules. This behavior is discussed in [20].

The proxies are composed of four internal components: communicator, privacy guard, data adapter, and service adapter. A service request first passes the communicator that ensures that all communication to the outside of the proxy is encrypted and secured in compliance with the policies. The privacy guard is applies the privacy policies to the payload of the service request by hiding, anonymization, and pseudonymization of data. The privacy compliant payload is then forwarded to the data adapter, which transforms the data from the internal data scheme to the payment provider's data scheme. This is necessary since every payment provider uses another set of parameters to issue the payment. Finally, the transformed payload is sent to the payment provider's service by the service adapter. After the procession of the service request the data goes the same way back threw all components of the proxy. While the data flows threw service adapter, data adapter, privacy guard, and communicator all changes applied to the payload are reverted. This approach ensures a maximal security and privacy level while consuming the service.

### B. Evaluation of the architecture

The proposed architecture ensures secure service interaction in collaborative intercloud scenarios with a maximal flexibility in terms of the participating cloud authorization protocols. Table IV shows the implementation of the requirements that have been put on the system.

## VI. Summary and outlook

This paper addressed the question how security challenges for service communication that arise due to the adoption of Intercloud environments can be addressed. To illustrate the identified challenges a sample business process has been introduced. This collaborative process that is operated by a central CSP is able to handle payment requests of merchants with a maximum flexibility in terms of the selected payment provider. To identify the optimal architecture multiple approaches of Intercloud service interaction have been reviewed. Based on the outcome of the reviews an architecture for secure service invocation in collaborative business processes has been developed. The architecture is based on proxies that are managed by a central broker engine.

During evaluation phase it turned out that this architecture is able to cope with the most challenges, e.g. interoperability, flexibility, scalability, and auditing. We also identified some limitations of the proposed architecture. First, due to the central management of the proxies it is necessary that the CSP provides enough resources to create and operate enough proxies, even when there are big peaks of requests. The centralized infrastructure also exposes the platform to the risk of outages if a component of the platform has a failure. To reduce this risk we implemented fault tolerance mechanisms into central parts of the platform on virtual machine level.

To ensure a maximal security of all components additional research is necessary. A first step has been made in [20], in this paper the communication flow and structure of the broker and proxy layer is described in detail. Another important field for additional research is the handling of access control to ensure data privacy throughout the whole business process.

## References

[1] T. Bittman. (2008) The evolution of the cloud computing market. [Online]. Available: http://blogs.gartner.com/thomas_bittman/2008/11/03/the-evolution-of-the-cloud-computing-market/

[2] R. Buyya, R. Ranjan, and R. N. Calheiros, "Intercloud: Utility-oriented federation of cloud computing environments for scaling of application services," in *Algorithms and architectures for parallel processing*. Springer, 2010, pp. 13–31.

[3] N. Grozev and R. Buyya, "Inter-cloud architectures and application brokering: taxonomy and survey," *Software: Practice and Experience*, vol. 44, no. 3, pp. 369–390, 2014.

[4] J. Lloret, M. Garcia, J. Tomas, and J. J. Rodrigues, "Architecture and protocol for intercloud communication," *Information Sciences*, vol. 258, pp. 434–451, 2014.

[5] A. Celesti, F. Tusa, M. Villari, and A. Puliafito, "How to enhance cloud architectures to enable cross-federation," in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, 2010, pp. 337–345.

[6] M. Singhal, S. Chandrasekhar, Tingjian Ge, R. Sandhu, R. Krishnan, Gail-Joon Ahn, and E. Bertino, "Collaboration in multicloud computing environments: Framework and security issues," *Computer*, vol. 46, no. 2, pp. 76–84, 2013.

[7] A. N. Toosi, R. N. Calheiros, and R. Buyya, "Interconnected cloud computing environments: Challenges, taxonomy, and survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, p. 7, 2014.

[8] GICTF, "Use cases and functional requirements for inter-cloud computing," 2010.

[9] Q. Chen and M. Hsu, "Inter-enterprise collaborative business process management," in *Data Engineering, 2001. Proceedings. 17th International Conference on*, 2001, pp. 253–260.

[10] D. Bernstein, D. Vij, and S. Diamond, "An intercloud cloud computing economy-technology, governance, and market blueprints," in *SRII Global Conference (SRII), 2011 Annual*, 2011, pp. 293–299.

[11] S. Ran, "A model for web services discovery with qos," *ACM Sigecom exchanges*, vol. 4, no. 1, pp. 1–10, 2003.

[12] B. Krumay and M. C. Oetzel, "Security and privacy in companies: State-of-the-art and qualitative analysis," in *Availability, Reliability and Security (ARES), 2011 Sixth International Conference on*, 2011, pp. 313–320.

[13] M. Mowbray, S. Pearson, and Y. Shen, "Enhancing privacy in cloud computing via policy-based obfuscation," *The Journal of Supercomputing*, vol. 61, no. 2, pp. 267–291, 2012.

[14] K. YIN and H. WANG, "A mutual authentication protocol providing security and privacy protection for intercloud environments," *Journal of Computational Information Systems*, vol. 10, no. 21, pp. 9087–9093, 2014.

[15] Y. Demchenko, C. Ngo, M. Makkes, R. Stgrijkers, and C. d. Laat, "Defining inter-cloud architecture for interoperability and integration," in *CLOUD COMPUTING 2012, The Third International Conference on Cloud Computing, GRIDs, and Virtualization*, 2012, pp. 174–180.

[16] Y. Demchenko, C. Ngo, C. d. Laat, M. X. Makkes, and R. Strijkers, "Intercloud architecture framework for heterogeneous multi-provider cloud based infrastructure services provisioning," *International Journal of Next-Generation Computing*, vol. 4, no. 2, 2013.

[17] M. X. Makkes, C. Ngo, Y. Demchenko, R. Stijkers, R. Meijer, and C. d. Laat, "Defining intercloud federation framework for multi-provider cloud services integration," in *CLOUD COMPUTING 2013, The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization*, 2013, pp. 185–190.

[18] D. Bernstein and D. Vij, "Simple storage replication protocol (ssrp) for intercloud," in *EMERGING 2010, The Second International Conference on Emerging Network Intelligence*, 2010, pp. 30–37.

[19] ——, "Intercloud security considerations," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, 2010, pp. 537–544.

[20] B. Schwarzbach, A. Pirogov, A. Schier, and B. Franczyk, "Inter-cloud architecture for privacy-preserving collaborative bpaas," Shanghai, 2015.

TABLE III
EVALUATION OF AUTHORIZATION AND AUTHENTICATION PROTOCOLS

| **Criterion** | **SAML 2.0** | **OAuth 2.0** | **OpenID Connect** | **CAS** |
|---|---|---|---|---|
| Exchange format | XML | URL parameter / JSON | URL-Parameter / JSON | URL-Parameter |
| Transfer protocol | HTTP, SOAP | HTTP | HTTP | HTTP |
| Token size | $5.7kB$ | $77B$ | $121B$ | $35B$ |
| IdP discovery | x | | x | |
| SSO | x | x | x | x |
| Strengths | Encryption | Wide spread | Authentication & authorization | |

TABLE IV
EVALUATION OF THE ARCHITECTURE

| **Requirement** | **Description** | **Implementation** |
|---|---|---|
| Interoperability | Communication has to be independent of the used authentication and authorization protocol | By the use of proxies the actual authentication and authorization protocol is been hidden from the other components. The broker instantiates the right proxy for the service provider. |
| Scalability | Flexible expandability | By adding new proxies the system can easily be expanded for new service providers. |
| Reliability | High availability of the payment service | In case on payment provider is not available the broker select the next suitable service provider. Hence, only if all suitable payment providers are unavailable the whole service is not working any more. |
| Optimal service provisioning | Service discovery, separation of combined services | The platform is divided into the components central platform, broker, service repository, and proxies |
| Access & identity management | Secure management of multiple identities; secure trust context | The identity management is done by the identity management system that is part of the central platform. This ensures no information is transferred to third parties unnecessarily. Additional firewall policies to protect the central platform are applied. Necessary communication on identities is realized by proxies, which implement secure authentication and authorization protocols. |
| Trust | Trust context | Externally by the use of proxies. Our sample process uses OpenID Connect and its PKI |
| Policy | Policy resolution | Monitoring |
| Privacy | Data & identity privacy | All communication, internal and external, is encrypted. The broker applies privacy control on the data. |

DOI: 10.15439/2015F188
ACSIS, Vol. 5

# The automatic summarization of text documents in the Cognitive Integrated Management Information System

Marcin Hernes
Wrocław University of Economics
ul. Komandorska 118/120, 53-345
Wrocław, Poland
Email: marcin.hernes@ue.wroc.pl

Marcin Maleszka
Wroclaw University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370
Wrocław, Poland
Email: marcin.maleszka@pwr.edu.pl

Ngoc Thanh Nguyen
Wroclaw University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370
Wrocław, Poland
Email:
ngoc-thanh.nguyen@pwr.edu.pl

Andrzej Bytniewski
Wrocław University of Economics
ul. Komandorska 118/120, 53-345
Wrocław, Poland
Email:
andrzej.bytniewski@ue.wroc.pl

***Abstract*—This paper presents issues related to a process of the automatic summarization of the text documents connected with economic knowledge performed by the cognitive agents in an integrated management information system. In contemporary companies, the unstructured knowledge is essential, mainly due to the possibility of obtaining better flexibility and competitiveness of the organization. Therefore more often the decision are taken in the enterprises on the basis of the summaries. The first part of the paper shortly presents the state-of-the-art in the considered field; next, the summarization process in the Cognitive Integrated Management Information System is characterized; the case study related with the summaries generating agent is presented in the last part of this paper.**

## I. Introduction

IN contemporary companies the unstructured knowledge is essential, mainly due to the possibility of obtaining better flexibility and competitiveness of the organization. The unstructured knowledge supports structuralized knowledge to a high degree. It is mainly stored in natural language, so it is processed with symbols (not numbers). One example of unstructured knowledge is experts' opinion about a predicted currency trading. Some experts may argue that the exchange rate of the currency will rise, others that it will decrease, and still others that it will remain unchanged. In addition, expert opinions include the reasons for these predictions. The number of such opinions in the Internet is usually very large (hundreds, thousands of the opinions). An investor who makes a decision, for example, on the financial markets, needs to analyze and summarize these opinions to formulate the correct decision. However, the manual realization of these processes is extremely difficult, and often impossible, due to time constraints [14]. Thus, often the processes of the analysis and summarization of the text documents are made automatically by computer systems, including the integrated management information systems. They may be constructed, for example, on the basis of the number of the cognitive agents [16]. Generally speaking, the cognitive agent is a smart program that not only concludes on the basis of the data received, takes specific actions to achieve the desired objective (this can be, for example, decision support), but also learns at the same time gaining experience. An example of a cognitive agent's architecture is The Learning Intelligent Distribution Agent (LIDA) [35]. This is a hybrid architecture that allows for symbolic and emergent knowledge processing and uses the semantic net with node and links activation level (the "slipnet") [19] to represent a knowledge.

Broadly understood, the analysis of the text documents is mainly based on document retrieval, information extraction, text mining and natural language processing. Summarization, instead, can include the contents of a document or set of documents. The basic idea of summarization is to get a summary that contains the most important information from the source document. One of the parameters of this process is the text volume. A good document summary frees the system user (investor, manager) from the need to read and analyze all of the text documents, and give the opportunity to focus his attention on aspects of the rapid and effective decision.

So far, the methods of summarization of the electronic text documents, containing economic knowledge and represented by the "slipnet" (semantic net with node and links activation level), has not been developed. It should be noted, however, that this type of representation allows the processing of both

978-83-60810-66-8/$25.00©2015, IEEE

knowledge represented in a symbolic way and knowledge represented in a numerical way.

The aim of this paper is to develop a method for automatic summarization of the electronic text documents, containing economic knowledge. This method will be implemented in the architecture of the cognitive agents running in the Cognitive Integrated Management Information System (CIMIS). The agent-based approach is used mainly because it enables taking automatic decisions and performing actions on the basis of summarization results.

This paper is organized as follows: the first part shortly presents the state-of-the-art in the considered field; next, the summarization process in CIMIS is characterized; the case study related with summaries generating agent is presented in the last part of paper.

## II. RELATED WORKS

The problems of automatic summarization have been widely considered in many papers and practical solutions. The simplest method for creating summaries is based on the assumption that the weight of the sentence depends on the weight of its words, calculated on the basis of their frequency in the text. In addition to the counted weight of words, other factors are also taken into account, such as the position of sentences in the text and the occurrence of words in the title or header [30]. An important work on the problem, was the project Parseval/GEIG, where the phrase structure grammars have been compared [4]. In this project using an anaphoric relations to create appropriate groups (consisting of paragraphs) has been proposed. The authors propose is that either all members of the group belong to the summary, or no element of the group appears in the summary. The solution proposed by [24] was based on several standard heuristics (e.g. sentence length, sentence position, keywords). However, the poor quality of the summaries generated using these tools has led to the search for alternatives. Another solution was tf-ifd system, used in ANES (Automatic News Extraction System), which determined the weight of words based on the number of their instances in the analyzed document [5]. R.Barzilay and M.Elhadad [3], instead, used an algorithm based on lexical chain for automatic creation of summaries. In the Tipster project a series of activities focusing on tasks such as summarization, translation and searching, were realized [18]. The paper [20] presents an overview of the different methods of summarization. Description of the elements of natural language processing and information retrieval methods have been widely presented in the works [2] and [7]. D.Weiss [43] presents algorithms of lemmatization and stemming, and a hybrid solution combining both approaches. P.Sołdacki [38] devoted his doctoral dissertation to an automatic processing of documents in natural language, in particular the use of text analysis methods for the shallow processing of documents in Polish. The problems with the automatic creation of summaries of texts document in Polish were well represented in the work of A.Dudczak [10]. Also E. Branny, M. Gajecki [46] present an algorithm for text summarization in Polish. Most of the mentioned works have been related to the creation of a single document summaries. Issues related to the creation the summaries of a set of documents can be found, for example, in the works [9], [30]. The papers [14], [27, 28] present two main streams of research related to the analysis of natural language texts:

- the formal description of the language and the real world's right,
- the statistical methods that bypass the problem of understanding of the text for the analysis of the prevalence of selected dependences.

J. Gramacki and A. Gramacki [15] used selected algebraic methods of analysis of textual data for automatic creation of summaries. Their paper presents the data structures and data modeling, which showed the essence of the reduced vectors' space of the mapped texts.

The paper [40] however, concerns the use of semantic roles in the process of summarization. The work [25] proposes a system of summarization of text documents by using a semantic net (semantic graph) for knowledge representation and mechanisms of the support vectors in the learning process. The methods rely on creating a new node for each sentence. (Fig.1).

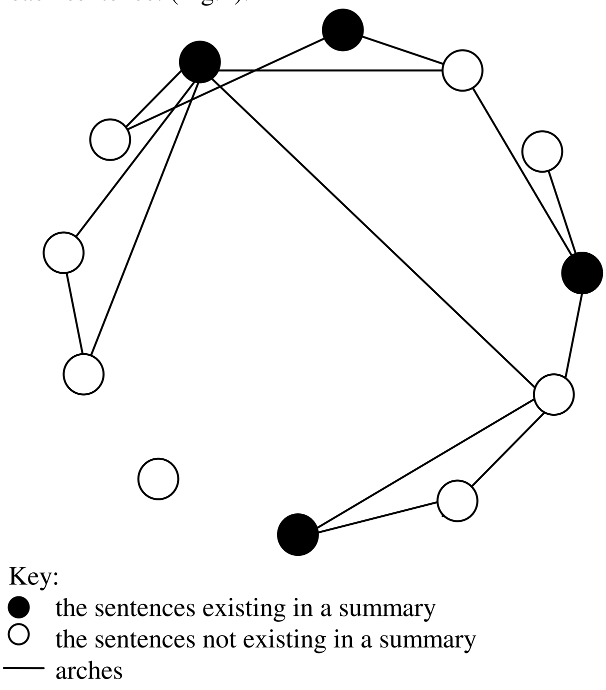

Key:
● the sentences existing in a summary
○ the sentences not existing in a summary
—— arches

Fig 1. The example of semantic graph.

If two sentences have common word then these sentences' nodes are connected by an arc (if the sentences have more than one common word - for each pair a separate arc is created). Sub-graphs whose nodes are connected by arches with other sub-graphs contain statements that define a good topic, while the nodes of sentences that have the highest cardinality (most arches) are the most significant sentences in the considered text, and should be included in the summary. However the applied semantic net does not include the activation levels of nodes and arcs.

Taking into consideration issues related to a text documents analysis process, nowadays the hybrid methods for processing unstructured knowledge are used; the methods involve structuralization of knowledge, followed by symbolic processing (e.g. with the use of expert systems or genetic algorithms) or converting knowledge into numerical representation followed by numerical processing (e.g. with the use of neural networks or fuzzy logic systems). In both cases, for knowledge processing, the following methods are used [32]:

- Information Retrieval,
- Information Extraction,
- Text Mining,
- Natural Language Processing.

There are two categories of analysis of text documents in the literature of the subject (for example [38]):

1. Deep Text Processing, that is a linguistic analysis of all possible interpretations and grammatical relationships occurring in natural text. The full analysis can be very complex. Moreover in many cases, the information obtained in this way may not be necessary. For this reason, more and more often there is a tendency to carry out only a partial analysis of the text, which may be much less time-consuming and is a compromise between precision and performance.
2. Shallow Text Processing, is defined as the analysis of the text to which the effect is incomplete in relation to the deep analysis of the text. The usual limitation lies in the identification of non-recursive structures or of limited recursion level, which can be diagnosed with a large degree of certainty. The structures requiring a complex analysis of the many possible solutions are overlooked or analyzed in part. The analysis is addressed mainly at recognizing proper names, noun phrases, verb groups without resolving their internal structure and function in a sentence. In addition, recognized are some main parts of sentences, for example the judgment or the judgment group.

There are several different methods within these categories.

For information retrieval the Boolean Logic Model (BML) or ranked-output systems are used [39]. A BML query consists of words or phrases concatenated with logical operators, such as AND, OR, and NOT. As a result, the set of documents is divided into two sub-sets: the first sub-set consists of documents matched with the query and the second sub-set consists of documents mismatched with the query. A ranking system, using vector algebra, assesses the probability of the content of documents matching the content of the query, and on this basis the ranking of the found documents is created. In this approach, for example, the Vector Space Model, Probabilistic Model or Interface Network Model are used [34].

One of the methods used in the shallow analysis of text documents is machine learning [33]. Under this method there are used, among other things, the naive Bayesian classifiers [12] and support vector machines [21]. In these approaches an analysis is made of the prevalence of individual words (terms) in the documents concerned. For example, in work [42] with the use of support vector machines it was determined, that the product attribute is considered part of the text, whereas in work [43] polarity of the opinion was expressed during the passage of the text.

Another method of both deep and shallow analysis of text documents is the use of rules, on the basis of which identification (annotation) of pieces of text for a specific topic is performed. Such rules are based on templates, taking into account the relationship between words and semantic classes of words [37]. The basis for the generation of rules can be automatic or manual analysis of the annotated corpus [31]. An analysis of documents by using rules rely on identifying the importance of text fragments in accordance with the principles enshrined in the rule. In certain cases it can be thought of as assigning the document to the category. The analysis of text documents with the use of the rules has been used, among others, to the extraction of spatial relationship [45], identification of the requirements for the IT projects~~,~~ expressed on Internet forums [41] or extraction of information from real estate ads [33]. In turn, the article [1 shows the characteristics of the project Semantic Monitoring of Cyberspace, that uses rules analysis in order to search cyberspace offers regarding illicit trafficking in drugs.

The text files are often represented in databases with key words contained in the document (symbolical representation of knowledge). However, such representation makes it difficult to compare documents, especially when it comes to measuring distance between documents - distance meaning similarity between the documents. Increasingly semantics nets are used to represent text documents, for example the topic maps. In work [11] it was found that the topic map allows to record information ontology and data taxonomy, structured semantically and at the same time it allows for knowledge mapping (both structured and unstructured) on a wide variety of hierarchical dependencies that exist between economic concepts and semantics (the concept of this type include, among others, ~~to~~ text documents in the field of management and economics).

However to match fully the needs of decision makers, a decision structure must consist of the level of certainty because the economic decision most are taken in terms of risk or uncertainty. Nowadays such structures [36] and consensus algorithms as regards these decisions~~,~~ are elaborated [23, 47, 50].

Thus it is possible to determine a certainty level of semantic relations between nodes (topics). In case of the economic knowledge it is very important issue, because the decisions making on the basis of this knowledge is performed usually takes place under risk and uncertainty conditions.

## III. Automatic summarization in the CIMIS

The CIMIS is dedicated mainly for the middle and large manufacturing enterprises operating on the Polish market (because the user language, at the moment, is Polish language). This does not preclude the implementation of other language text documents processing in the CIMIS. The process of analysis of text documents highly depends on the language. In addition, it is more difficult in case Polish language than, for example, in case English or German due to the greater complexity of Polish grammar [32].

The Learning Intelligent Distribution Agent (LIDA) architecture is used in CIMIS construction [8], [13]. In the construction of a LIDA agent mixed-used symbolically-connexionistic organization of memory is used in an attempt to ground the meaning of all symbols. It is necessary to properly process the unstructured knowledge, recorded mostly using natural language, such as customer's opinion about products. Grounding is meant as these cognitive processes, which are responsible for establishing and maintaining a connection between the language and the corresponding objects in the world [22]. The LIDA consists of the following modules [13]:

- Sensory Memory
- Perceptual-Associative Memory,
- Workspace,
- Transient Episodic Memory,
- Declarative Memory,
- Attentional Codelets,
- Global Workspace,
- Procedural Memory,
- Action Selection,
- Sensory-Motor Memory.

In the LIDA architecture it was adopted that the majority of basic operations are performed by the so-called codelets, namely specialized, mobile programs processing information in the model of global workspace.

In 2011, the CCRG group released the Framework LIDA, which allowed the use of this architecture by a wider circle of users. Framework LIDA is a software underlying the implementation of the cognitive agents. The Framework contains the object class (implemented in JAVA), performing operation in the field of agent architecture (definition and methods of handling all types of memory, communication protocols, methods for making reservation by the agent operations on real-world objects, such as searching for and identifying the objects, specify characteristics of objects, specifying associations between objects). The programmer's task is to fill in tools provided by the framework LIDA (write a program code) about aspects related to the specific domain of the problem - for example related to economics, management.

The CIMIS consist of following sub-systems: fixed assets, logistics, manufacturing management, human resources management, financial and accounting, controlling, CRM, business intelligence. These sub-systems are described in detail in [16]. In this paper we focus on summarization module, which is a part of CRM sub-system (Fig 2.). The module consist of following groups of agents:

- document retrieval,
- information extraction,
- text analysis,
- summaries generating.

Document retrieval agents search and retrieve~~,~~ from the internet sources the documents according to users' needs (consistent with the query retrieval ) and save full content of these documents in a database. Retrieving the documents is carried out by codelets operating in an agent environment. In order to achieve a higher level of retrieval effectiveness each agent running on the basis of different document searching engines, for example:

- Java Searching Engine,
- Microsoft BING API,
- Custom Search API,

The documents from the top of its list are stored in database. The number of stored documents is given in the form of a parameter in the agent configuration file (e.g. 10, 20, 50 documents). The format of retrieved documents is not limited (e.g. Pdf, PS, doc, html, xml), while in the database, each document is stored as text, except that html document is saved along with markers.

The role of information extraction agents is to identify essential information in text documents. For example, if opinions of mobile telephone users are to be analyzed, only those pieces of texts which contain opinions (advertisements are to be omitted) need to be extracted from text documents (saved in a database by documents searching agents). Each agent uses a different method of extracting information, for example:

- determining tags of beginnings and ends of essential fragments of texts in a document - the method is mainly used in case of files saved in html/xml format (for example, the series of characters ""/><p>" is treated as the beginning of an opinion whereas the series of characters "</p>" is treated as an end of an opinion (the characters are determined on the basis of a learning set of text documents);
- identifying essential fragments of texts on the basis of sets of key words or rules - the method is used in case of any document formats (for example an opinion about a phone may be identified on the basis of such

key words as "make", "model", "recommended/not recommended"),

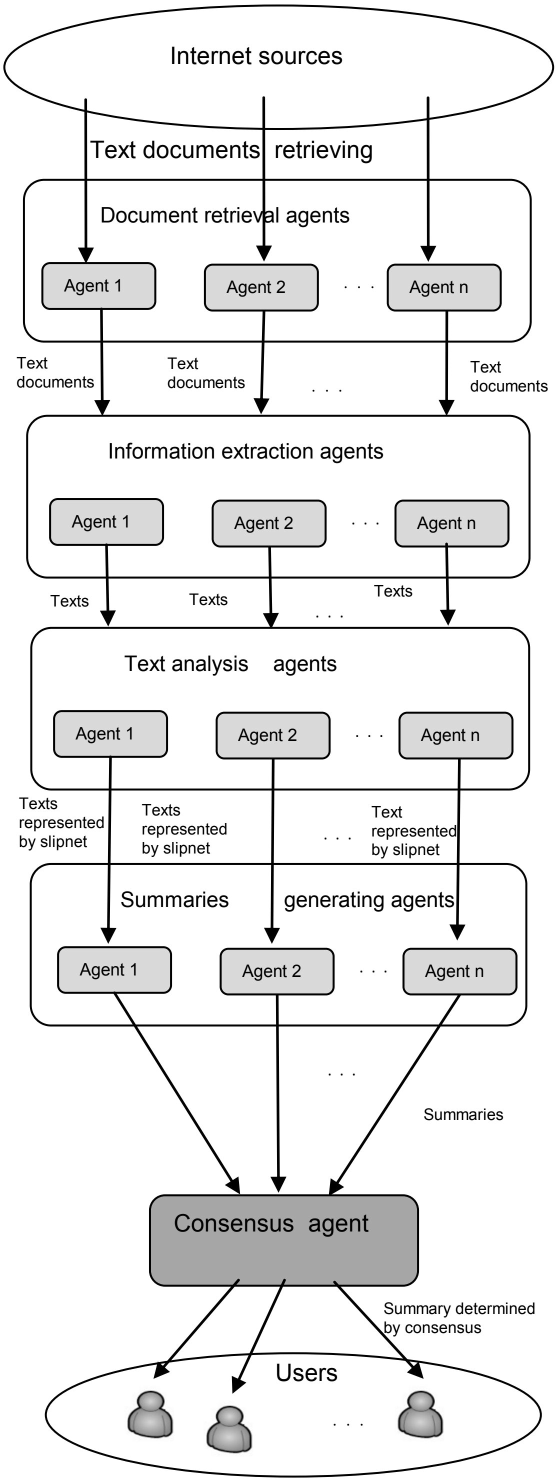


Fig 2. The functional architecture of automatic summarization module.

- identification of essential fragments of texts using documents representations in the form of a semantic network - semantic networks are created on the basis of a learning set, the networks serve as patterns representing particular classes of information (for example a semantic network representing a client's opinion on a given mobile phone), whereas in the process of IE a text document is saved also in the form of a semantic network, and then a cognitive agent searches for a given pattern in a semantic network representing a text document (the degree of similarity is defines as a parameter). This method allows for including a context of extracted information.

The results of the agents' running are stored in the text database.

The environment in which text analysis agents operate constitutes a set of text documents containing information which is the result of operation of agents extracting information (for example opinions of customers about mobile phones) located in a database of a system. Text analysis is performed in the following way:

1. A semantic network containing terms and connections between them is created in the perceptive memory on the basis of a learning set (for example a set of opinions about mobile phones). The perceptive memory stores also synonyms and different variations of words (thesaurus). In the perceptive memory of LIDA agents terms are represented by means of nodes, whereas connections are represented by means of links.

2. Individual text documents are added one by one into the sensory memory.

3. Opinions are analyzed by codelets, i.e. programs which search through texts according to certain criteria specified by means of configuration parameters. Values of the parameters may be indicated by users (parameters are saved in xml file structure and used in the program code of codelets). Codelets have been divided into two groups:

1) Codelets performing shallow analysis of texts. In the frame of this group the following codelets are running:

- tokenization,
- morphological analysis,
- removing the ambiguity,
- recognition of proper names, replacing pronouns,
- the distribution of complex sentences to the simple sentences,

2) Codelets performing in-depth analysis of texts. These codelets, by using thesaurus and results of shallow analysis codelets' operation, search for all possible interpretations and grammatical relations present in an analyzed document, represented in the form of a semantic network ("slipnet") in the perceptive memory of an agent. Results of documents' analysis, represented also in the form of a semantic network, are saved in a database.

4. The next step consists of passing the situation model to the global working memory, and from the procedural memory the following patterns of action are automatically selected: “saving results of opinions analysis into a database” (noSQL database – analysis of results – semantic network – are saved in XML format) and “entering another document into the sensory memory”.

The environment in which summary generating agents operate, on the other hand, consists of a set of text documents represented in the form of a semantic network constituting the result of operations of agents analyzing documents. Summary generating agents function in a way similar to documents analysis agents, however, codelets perform tasks connected with summarizing and they are divided into the following groups:

1. Codelets of text level units – responsible for analyzing relations between fragments of a semantic network, e.g.: probability, proximity in a text, common references, language ties (taken from a thesaurus for example), syntactic and semantic relations. Results of analyses are saved in the form of a semantic network with levels of nodes and links activation reflecting the degree of similarity and relations.

2. Discourse level codelets – responsible for performing analysis in the course of which the format of a whole document is taken into account, its rhetoric structure and issues touched upon in the document. The codelets use the method of creating semantic graphs and their subgraphs (graphs and subgraphs are also created in the form of a semantic network “slipnet”) using mechanisms of vectors facilitating the learning process. Levels of activating links in a graph depend on the level of the meaning of a given sentence (or phrase) in an analyzed text.

Next, the situation model is transferred to the global workspace, and from the procedural memory, the following patterns of action are automatically selected: “saving results of opinions analysis in a database”, “generating summaries in a natural language” (a summary in the form of a text in a natural language, which can be presented to a user, is created on the basis of a summary in the form of a semantic network) and “entering another document into the sensory memory”.

The integration of summaries is performing by consensus agent. This agent determines a summary presented to user (users). The use of consensus methods [17], [26], [29], instead, allows for the integration of summaries. The general meaning of the term consensus refers to an agreement. A consensus of a certain set (profile) of text documents may constitute a new document (hypothetical one) created on the basis of documents contained in the profile [48, 49].

Deriving consensus consists of three basic stages. In the first stage, one needs to determine a method of text documents representation. In this paper it has been assumed that the documents are represented in the semantic net. The next step requires defining the function of calculating distance between individual variants. The third stage involves developing consensus deriving algorithms, i.e. determining a representation of a set of documents (profile) where the distance between the representation (consensus) and individual documents of the profile is minimal (according to various criteria). If, for example, the different summaries of the documents describing the given phenomenon have been generated by cognitive agents, then using the consensus methods, on the basis of this summaries, the one variant of summary can be generated and presented to the user. This variant does not need to be one of the summaries generated by the cognitive agents. It can be a new variant determined on the basis of these summaries. Thus, all the summaries on a given phenomenon can be taken into account. Such a solution allows, among other, for shortening the time of determining the target summary (the user does not need to analyze the individual summaries and reflect on their choice - multi-agent system performs these steps automatically for user) and for decreasing the risk related to the choice of the worst summary (because all the summaries are taken into account in the consensus). As a result, the business processes within an organization can be implemented more quickly and efficiently.

## IV. Functionality of Summaries Generating Agent – the Case Study

In the frame of this case study the Polish language text documents will be taken into consideration.

The functioning of the module generating summaries of text documents will be presented on the example of experts’ opinions present in the Internet. A document-searching agent has received a task to find in the Internet documents whose content matches the following question: “In what stocks to invest during the uncertainty on the market?” For further investigation, one of the opinions obtained by an information-extracting agent was selected (the method of information extraction by determining beginning and end characters or marks of essential pieces of text in a document was used – see section III). The text of the opinion looks as follows: (sentences/phrases) of the opinion have been numbered by authors in order to simplify further analysis):

*“(1) The market is in the period of uncertainty (2) therefore investments are to be made into companies representing the defensive branches. (3) Among them, we can first of all mention telecommunication companies stocks, (4) as well as those which are related to public utility. (5) We are talking here about for example energy, (6) or all other companies which make investments using public money. (7) They are the least affected by changes in the economic situation (8). Many companies conduct different types of promotion actions aimed at making people invest in them. (9) However, one should not fall for that. (10) One should not invest in retail sector companies as their prices change greatly during a fluctuation on the financial market”.*

On the basis of a learning set, a semantic network containing terms and connections between them related to the investment topic, as well as a thesaurus for the terms were saved in the perceptive memory of a text analysis agent[1]. Next, a sample opinion was entered into the sensory memory of the agent, and a shallow analysis of the text was performed. As a result of the analysis, for example words “invest”, “investments” were changed into “investment”, and the phrase “changes in the economic situation” were changed into “fluctuation” (tokenization). Complex sentences were also broken into simple sentences (for example the sentence: “The market is in the period of uncertainty therefore investments are to be made into companies representing the defensive branches.” was broken into two sentences: “The market is in the period of uncertainty” and “therefore investments are to be made into companies representing the defensive branches). In the next step, codelets performing in-depth analysis of the text (see item 3) saved an opinion in the semantic form. Figure 4 on the other hand shows representation of the following sentence: “especially during of uncertainty on the market”.

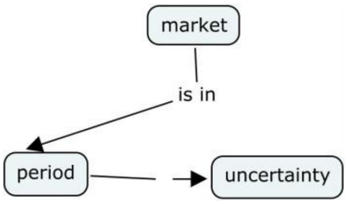


Fig. 3. A section of the network which relates to the following sentence: “especially during of uncertainty on the market”

Figure 4 presents a section of the network which relates to the following sentence: “Investments are to be made into companies representing defensive branches”.

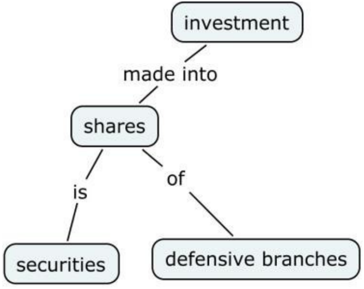


Fig 4. A section of the network which relates to the following sentence: “Investments are to be made into companies representing defensive branches”

It can be observed that the network has been enriched with the meaning of words “shares” and “companies”. It has been stated that they refer to “securities”.

The sentence: “Many companies conduct different types of promotion campaigns aimed at making people invest in them” (it is a dependent sentence so it has not been divided into two sentences) is represented in a way shown in figure 5. One may notice that the network has been enriched with the meaning of the phrase “promotion actions” – it has been specified that these are “actions”/”campaigns”. The remaining sentences or sentences of the opinions are represented in a similar way.

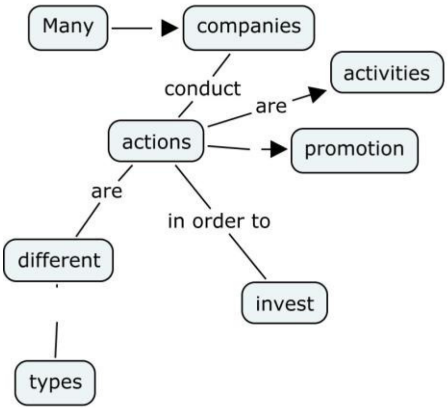


Fig. 5. A section of the network which relates to the following sentence: Many companies conduct different types of promotion actions aimed at making people invest in them”.

Next, the opinion, in the form of a semantic network, is sent to the sensory memory of a summary-generating agent. Discourse level codelets generate a semantic graph (figure 6) whose nodes refer to individual sentences of an opinion (the number of sentences is specified after breaking dependent sentences into simple sentences – in case of the particular opinion, there are 10 sentences). This graph corresponds to the opinion in Polish (see an Appendix). The nodes of the graph have been joined using links whose activation levels are directly proportional to the number of words common for the investigated sentences (activation level is defined in the following way: the longest sentence in a particular opinion consists of 10 words (excluding linkers) and the activation level is defined within the range [0..1], so in case of each common word the level of activation increases by 0,1.) [2].

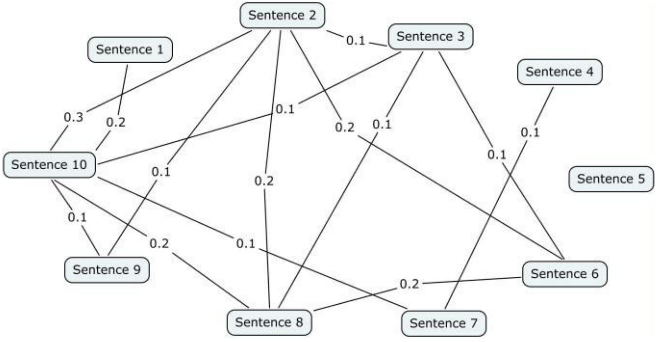


Fig 6. The semantic graph representing the considered opinion.

The most meaningful sentences are marked by nodes which have the highest total number of activation links which come out of them. The sentences should be included in a summary. Taking into account the presented semantic graph, these are the following sentences: “Sentence 2” (total

[1] The way for text analysis by cognitive agent is broader described in the paper [6].

[2] The longest sentence in the present opinion is 10 words (not including conjunctions) and the level of activation is defined in the interval [0..1], therefore, corresponds to one common expression on the activation level rise of 0.1.

of links activation levels: 0.9), “Sentence 6” (total of links activation levels: 0,5), and one of the sentence: “Sentence 10” (the sentences have the total of links activation levels: 1.0).

Apart from discourse level codelets (analyzing semantic graph) there are also text units level codelets. They enable finding sentences of different semantics despite comprising similar vocabulary. “Sentence 2” and “Sentence 8” serve as examples. The semantic graph suggests that it is enough to use just one of the sentences, however, after analysis performed by text units level codelets, it appears that “Sentence 2” (“Investments are to be made into companies representing the defensive branches”) refers to investing into shares, whereas “Sentence 8” (“Many companies conduct different types of promotion actions aimed at making people invest in them”) refers to taking action by companies (in Polish the words “shares” and “actions” are the same, i.e. “akcje” – see the Appendix) . Discourse level codelets mark sentences which shall be used in a summary by setting the level of activation of nodes for these sentences at the value 1 (value 0 for nodes representing sentences which shall not be placed in a summary).

Consequently, as a result of summary generating agent’s operation the following text of a summary of considered opinion has been defined:

*“During fluctuation on the financial market, investments are to be made into companies, representing the defensive branches, or all other companies which make investments using public money. One should not invest in retail sector companies”.*

It needs to be stressed that using representations of knowledge of a cognitive agent in the form of a semantic network “slipnet” enables taking into account a greater number of criteria which determine which sentences (or phrases) are to be included in a summary. For example, one can specify if more important are sentences whose nodes have a smaller number of links but whose levels of links activation are high (the sentence is “strongly” connected with other sentences, however the sentences refer to a small fragment of a summarized text), or whether more important are sentences whose nodes have a very high number of links of low activation level (the sentence is poorly connected with other sentences, however the sentences refer to a large fragment of a summarized text). Additionally, a cognitive agent, on the basis of an obtained summary, is capable of taking automatic decisions (on behalf of a user – investor) concerning abandoning investments in equities or starting investments in, for example gold. An agent will automatically sell equities (if an investor has been investing on a given market) and buy a gold.

On the basis of results of the preliminary research experiment performed by using 30 text documents, it has been state, that in about 70% cases the automatic generated summaries were very similar to summaries generating by users.

However, the main limitations of presented proposal are as follows:

- the results of text document analysis are not always properly, therefore a summaries generated on the basis of these results are also not always correct,
- in many cases, complex sentences weren’t broken into simple sentences,
- there are also summaries which consist of not correct generated text (linguistic error).

## V. Conclusion

The paper has demonstrated the development of a model of generating summaries of text documents containing mainly economic knowledge, represented by means of the semantic network “slipnet” (containing, apart from arch nodes, also levels of their activation). This sort of representation enables processing symbolically represented knowledge as well as numerically represented knowledge. It is possible then to define the level of probability or strength of semantic connections between terms, sentences, or phrases. Thanks to that, the cognitive agent takes into consideration user’s criteria which generated summaries should satisfy. Using cognitive agents in the devised model also enables taking automatic decisions and performing actions (on the basis of knowledge acquired in the process of summarizing documents).

Proper summarization of text documents is of great significance, particularly in integrated management information system design to support decision-making processes. Their functionality has a direct effect on user decisions and – ultimately – affects the organization as a whole.

Further research works will concern, inter alia, the developing, by using a fuzzy logic, a method for determining the activation level of nodes depending on the context of the sentence (or expressions).

## Appendix

The considered opinion in Polish is as follows:

„(1) Rynek znajduje się w okresie niepewności, (2) więc należy inwestować w spółki reprezentujące tzw. branże defensywne. (3) Wśród nich można wymienić przede wszystkim akcje spółek telekomunikacyjnych, (4) a także tych, które mają związek z użytecznością publiczną. (5) Chodzi tu np. o energetykę. (6) lub wszystkie inne spółki, które wykonują inwestycje z pieniędzy publicznych (7) W małym stopniu są one narażone na zmiany koniunktury (8) Wiele spółek przeprowadza różnego rodzaju akcje promocyjne aby w nie inwestować. (9) Nie należy im jednak ulegać (10) Nie należy inwestować w spółki działające w handlu detalicznym gdyż ich cena ulega dużym zmianom w okresie fluktuacji na rynku papierów wartościowych”.

## REFERENCES

[1] W.Abramowicz, E. Bukowska and A. Filipowska, “Zapewnienie bezpieczeństwa przez semantyczne monitorowanie cyberprzestrzeni”, *e-mentor*, 3 (50) , 2013.

[2] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., MA, USA, Boston 1999.

[3] R. Barzilay and M. Elhadad , “Using lexical chains for text summarization”, *Inteligent Scalable Text Summarization Workshop* (ISTS'97), 1997, pp. 10-17.

[4] E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Reukos, B. Santoni and T. Strzalkowski, “A procedure for quantitatively comparing the syntactic coverage of English grammars”, *DARPA Speech and Natural Language Workshop*, 1991.

[5] R. Brandow, K. Mitze and L. F. Rau, “Automatic condensation of electronic publications by sentence selection”, *Inf. Process. Manage.*, 31(5), 1995, pp. 675-685.

[6] A. Bytniewski, A. Chojnacka-Komorowska, M. Hernes and K. Matouk, “The Implementation of the Perceptual Memory of Cognitive Agents in Integrated Management Information System”, in: D. Barbucha, N. T. Nguyen, J. Batubara, *New Trends in Intelligent Information and Database Systems*, Studies in Computational Intelligence Volume 598, Springer International Publishing Switzerland, 2015, pp 281-290. doi: 10.1007/978-3-319-16211-9_29

[7] M. Ciura, D. Grund, S. Kulików and N. Suszczanska, “A System to Adapt Techniques of Text Summarizing to Polish*”, International Conference on Computational Intelligence*, 2004 p 117-120.

[8] *Cognitive Computing Research Group*, http://ccrg.cs.memphis.edu/, [29.03.2015]

[9] D. Das and A.F.T. Martins, “A Survey on Automatic Text Summarization”, *Literature Survey for the Language and Statistics II course at CMU*, 2007.

[10] A. Dudczak, *Zastosowanie wybranych metod eksploracji danych do tworzenia streszczeń tekstów prasowych dla języka polskiego*, Praca Magisterska Politechniki Poznańskiej, Poznań 2006-2007.

[11] H. Dudycz, *Mapa pojęć jako wizualna reprezentacja wiedzy ekonomicznej*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2013.

[12] E. Frank and R. Bouckaert, “Naive bayes for text classification with unbalanced classes”, *Knowledge Discovery in Databases: PKDD* 2006.

[13] S. Franklin, F. G. Patterson, *The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent*, in *Proc. of the Int. Conf. on Integrated Design and Process Technology.* San Diego, CA: Society for Design and Process Science, 2006.

[14] A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing,* Springer, Berlin 2012.

[15] J. Gramacki and A. Gramacki, *Automatycznie tworzenie podsumowań tekstów metodami algebraicznymi*, Wyd. PAK nr 07, s. 751-755, Gliwice 2011.

[16] M. Hernes, “A Cognitive Integrated Management Support System for enterprises”, w: D. Hwang, J. Jung, N.T. Nguyen N (eds.), *Computational Collective Intelligence Technologies and Applications*, Lecture Notes in Artificial Intelligence, vol. 8733, Springer-Verlag, 2014, pp. 252-261

[17] M. Hernes and N. T. Nguyen, “Deriving Consensus for Hierarchical Incomplete Ordered Partitions and Coverings”, *Journal of Universal Computer Science* 13(2)/2007, pp. 317-328.

[18] L. Hirshman, *Language understanding evaluation: lessons learned from MUC and ATIS*, LREC Granada,1998.

[19] D. R. Hofstadter and M. Mitchell, “The copycat project: A model of mental fluidity and analogy-making”, in D. Hofstadter, the Fluid Analogies Research group (eds), *Fluid Concepts and Creative Analogies*, Basic Books. Chapter 5, 1995.

[20] P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications - Text Retrieval, Extraction and Categorization*, John Benjamins Publishing Company, Amsterdam/ Philadelphia 2002.

[21] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features”, *Machine learning: ECML-98*, 1998.

[22] T.H. Duong, N.T. Nguyen, G.S. Jo, “A Method for Integration of WordNet-based Ontologies Using Distance Measures”, in: Proceedings of KES 2008. Lecture Notes in Artificial Intelligence 5177, 2008, pp. 210-219. doi: 10.1007/978-3-540-85563-7_31.

[23] J. Korczak, M. Hernes and M. Bac, “Risk avoiding strategy in multi-agent trading system”, [in:] Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS), Kraków, 2013, , pp. 1119 - 1126.

[24] J. Kupiec, J. Pedersen and F. Chen, “A Trainable Document Summarizer”, Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, 1995, pp. , 68 - 73.

[25] J. Leskovec, M. Grobelnik and N. Milic-Frayling, “Learning Sub-structures of Document Semantic Graphs for Document Summarization”, in 'Proceedings of the KDD 2004 Workshop on Link Analysis and Group Detection (LinkKDD)', 2004.

[26] M. Maleszka and N.T. Nguyen, “Approximate Algorithms for Solving O1 Consensus Problems Using Complex Tree Structure”, *Transactions on Computational Collective Intelligence 8*, 2012, pp. 214-227.

[27] A. Mykowiecka, *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*, Wyd. Polsko-Japońska WSTK, Warszawa 2007.

[28] A. Nenkova and K. McKeown , “Automatic Summarization”, *Foundations and Trends in Information Retrieval*, Vol 5, Issue 2-3, 2011, pp 103-233.

[29] N.T. Nguyen, *Advanced Methods for Inconsistent Knowledge Management*, Springer-Verlag London, 2008.

[30] D. Radev, H. Jing, M. Stys and D. Tam, “Centroid-based summarization of multiple documents”, *Information Processing and Management*, pp. 919-938, 2004.

[31] L.V. Pham and S.B. Pham, “Information extraction for Vietnamese real estate advertisements”, *Fourth International Conference on Knowledge and Systems Engineering (KSE)*, Danang 2012.

[32] P. Potiopa, „Metody i narzędzia automatycznego przetwarzania informacji tekstowej i ich wykorzystanie w procesie zarządzania wiedzą”, *Automatyka* 15(2) pp. 409-419, http://journals.bg.agh.edu.pl/AUTOMATYKA/2011-02/Auto40.pdf.

[33] F. Sebastiani, “Machine learning in automated text categorization”, *ACM Computing Surveys (CSUR).*, 34(1), New York 2002.

[34] A. Singhal: “Modern Information Retrieval: A Brief Overview”, *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4), 2001, pp. 35-43.

[35] J. Snaider, R. McCall and S. Franklin, “The LIDA Framework as a General Tool for AGI”, *The Fourth Conference on Artificial General Intelligence*, 2011.

[36] J. Sobieska-Karpińska and M. Hernes, “Consensus determining algorithm in multiagent decision support system with taking into consideration improving agent's knowledge”, Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS), IEEE Xplore Digital Library, Wrocław 2012, pp. 1035-1040.

[37] S. Soderland, “Learning information extraction rules from semi-structured and free text”, *Machine Learning*, 34(1-3), 1999.

[38] P. Sołdacki, „Zastosowania metod płytkiej analizy tekstu do przetwarzania dokumentów w języku polskim”, Praca Doktorska, Politechnika Warszawska, Warszawa 2006.

[39] S.L. Tomassen, *Semi-automatic generation of ontologies for knowledge-intensive CBR*, Norwegian University of Science and Technology, 2002.

[40] D. Trandabăţ, “Using semantic roles to improve summaries”, *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG '11).* Association for Computational Linguistics, Stroudsburg, PA, USA, 2011 pp. 164-169.

[41] R.E. Vlas and W.N. Robinson, “Two rule-based natural language strategies for requirements discovery and classification in open source software development projects”, *Journal of Management Information Systems*, 28(4), 2012.

[42] A. Wawer,” Mining opinion attributes from texts using multiple kernel learning”, *IEEE 11th International Conference on Data Mining Workshops*, 2011.

[43] D. Weiss, “A Hybrid Stemmer for the Polish Language”, *Technical Report RA-002/05,* Institute of Computing Science, Poznan University of Technology, Poland, 2005.

[44] T. Wilson, J. Wiebe and P. Hoffmann, “Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis”, *Computational linguistics,* 35(3), 2009.

[45] C. Zhang, X. Zhang, W. Jiang, Q. Shen and S. Zhang, “Rule-based extraction of spatial relations in natural language text”, *International Conference on Computational Intelligence and Software Engineering,* 2009.

[46] E. Branny, M. Gajecki: “Text Summarizing in Polish”, *Computer Science, Annual of AGH University Of Science and Technology,* 2005, pp. 31–46.

[47] J. Korczak, M. Bac, K. Drelczuk and A. Fafuła, “A-Trader - Consulting Agent Platform for Stock Exchange Gamblers”, in *Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS),* Wrocław, 2012, pp. 963-968.

[48] L. Sliwko, N. T. Nguyen, “Using Multi-agent Systems and Consensus Methods for Information Retrieval in Internet”, *International Journal of Intelligent Information and Database Systems* 1(2), 2007, pp. 181-198. doi:10.1504/IJIIDS.2007.014949

[49] N. T. Nguyen, “Using consensus methods for solving conflicts of data in distributed systems”, in: *Proceedings of SOFSEM 2000, Lecture Notes in Computer Science* 1963, 2000, pp. 411-419. doi: 10.1007/3-540-44411-4_30

[50] M. Hernes M. and J. Sobieska-Karpińska , “Application of the consensus method in a multiagent financial decision support system”, *Information Systems and e-Business Management,* Springer Berlin Heidelberg 2015, doi: 10.1007/s10257-015-0280-9.

DOI: 10.15439/2015F149
ACSIS, Vol. 5

# Transaction Based Business Process Modeling

Frantisek Hunka
University of Ostrava, Faculty of Science,
Dvorakova 7, 701 03 Ostrava,
Czech Republic
Email: frantisek.hunka@osu.cz

Roman Belunek
University of Ostrava, Faculty of Science,
Dvorakova 7, 701 03 Ostrava,
Czech Republic
Email: Roman.Belunek@osu.cz

***Abstract*—A term of transaction which has its origin in database processing, representing a set of operations that must be performed all or none of them. The notion of transaction is also used in some business process modeling approaches such as the DEMO (Design & Engineering Methodology for Organizations) and the REA (resource-event-agent) value modeling approach. The DEMO's transaction forms a basic building block from which a business process is composed. The REA value modeling approach utilizes transactions in an REA model representing a business process. In general, both methodologies utilize the notion of transaction, which has, however, a different meaning in these approaches. The aim of the paper is to describe the basic models of both approaches and to show, with the 'rent-a-car' example, the principal differences between the notions of transaction between them. The paper also reflects on possible mutual collaboration between both approaches.**

## I. Introduction

BUSINESS process modeling unquestionably belongs to software development process. It usually directly influences the database solution of the problem domain. Most of the business process modeling methodologies simply originated form 'best practice' without a vigorous theory from which the methodology is derived, see [9]. They mostly focus on production actions, which are usually described as an event that happens instantaneously or over a period of time. Generally, the most recommended notation of business processes is an UML activity diagram with swim lines, see [8]. Each swim line represents a human being, more precisely an actor role. Human beings are an inseparable part of business process modeling. However, the absence of a vigorous theory means that business process modeling approaches suffer from various incompleteness.

DEMO methodology stems from Enterprise Ontology [3] which represents a generic approach to business process modeling. The benefit of this methodology is that it perfectly identifies principal transactions that create the business process including human beings. It also provides necessary abstractions that enable us to obtain the essence of the modeling reality. On the other hand, this methodology is designed to be generic, which means that it registers production activities and is aware of them but without affecting them.

The other ontology which is at the core of our interest is REA ontology, see [1]. Its name is derived from three fundamental concepts, namely: *Resources*, *Events* and *Agents*. This modeling approach originated from accountancy systems but was developed into a fully-fledged tool for business process modeling. Economic resources are things of economic value that have utility for economic agents and for that reason they are planned, monitored, and controlled. Examples of economic resources are money, raw materials, labor, tools, products, and services. Economic events are activities within an enterprise that represent either an increment or a decrement in the value of economic resources. Some economic events occur instantaneously, some occur over time. Examples of economic events are sales of goods, rentals, and provision and use of services. Economic agents are individuals or organizations that participate in the control and execution of economic events. Examples of economic agents are customers, vendors, employees and enterprises. The structure of the paper is as follows: Section Two describes the main features of the DEMO methodology that are further utilized. The REA value modeling approach is clarified in Section Three. After a narrative description of the example, Section Four states the DEMO and REA solution to the example. Section Five discusses both approaches and Section Six summarizes the results achieved.

## II. DEMO Methodology

According to DEMO methodology [3], [4] an organization is composed of people (social individuals) that perform two kinds of acts, *production* acts and *coordination* acts. By performing production acts, people fulfill the aims of the organization. A production act can be either material or immaterial. By a material production act we mean a tangible act such as a manufacturing or transportation act. By an immaterial act we mean an intangible act such as the approval of an insurance claim or delivery of a judgment. By performing coordination acts human beings enter into and comply with commitments. They initiate and coordinate production acts. Abstracting from the particular subject that performs the action, the notion of the *actor role* is introduced. A subject in his/her fulfillment of an actor role is called an actor.

The result of successfully performing a production act is a *production fact*. An example of a production fact may be that the payment has been paid or an offered service was accepted. All realization issues are fully abstracted. Only the facts as such are relevant, not how they are achieved. The result of successfully performing a *coordination act* is a *coordination fact*. Examples of coordination acts are *requesting* and *promising* a production fact.

The diagram in Fig. 1 shows the standard transaction pattern (transaction). It contains two actor roles, the initiator and the executor and coordination and production acts between them. Each transaction starts with a *request* coordination act made by the initiator. In response to the *request*, the executor performs either a *promise* or *decline* coordination act. In short, a *decline* means the

978-83-60810-66-8/$25.00©2015, IEEE

end of a transaction. The *promise* goes on in a *production act* which results in a *production fact*. The production fact is *stated* to the initiator who can either *accept* it or *reject* it. The standard transaction pattern can be extended to the complete transaction pattern. In this case, the transaction pattern also contains four cancellation patterns that enable revoking of an act and completely model real conditions. For the purpose of the paper, only the transaction pattern will be used. The transaction itself can be expressed in a more condensed way, see Fig. 2.

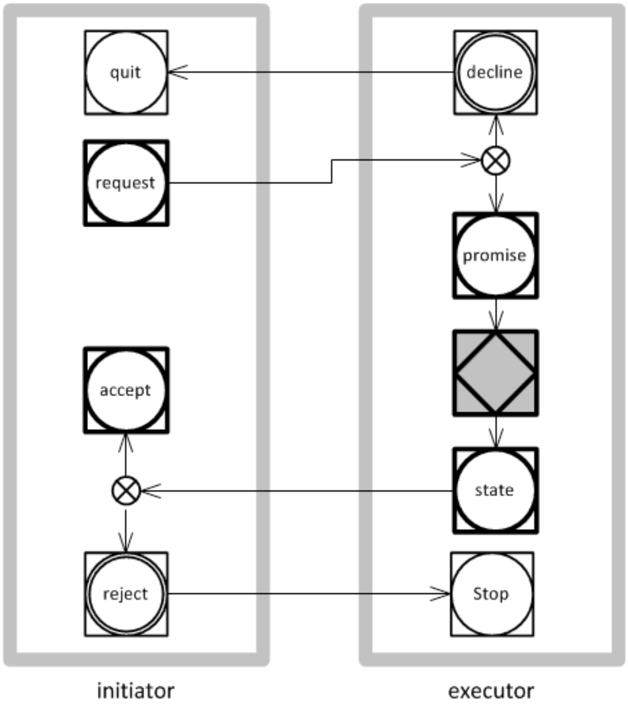


Fig. 1 DEMO transaction – standard pattern Source: [3]

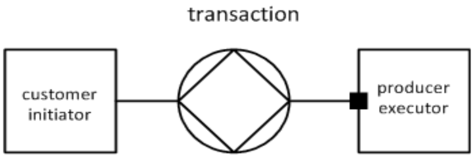


Fig. 2 Transaction in the Construction Model Source: [3]

The diagram in Fig. 2 shows the relation between the initiator (the relation is indicated by the plain line) and the executor (the relation is indicated by the dot at the executor). The DEMO methodology provides four mutually integrated aspect models. In this paper, only the Construction Model is used. The commission of the Construction Model is to identify actor roles and transactions. The task of the Process Model is to show how transactions are causally and conditionally related. The Construction Model has a crucial position in comparison with the REA model. The basic unit for declaring business processes is the notion of transaction. Transaction represents the basic building block between two social subjects. Apart from the basic states of *request*, *promise*, *production*, *state* and *accept* it also contains the other states that address erroneous states such as *decline* and *refuse* and the states which come from cancellation. Business process in DEMO methodology is defined as a set of enclosing transactions with a definite result (fact).

## III. REA Value Modeling Approach

The main benefits of the REA approach are being able to keep track of primary and raw data about economic resources. This explains why the REA approach offers a wider, more precise, and more up-to-date range of reports. All accounting artifacts such as debit, credit, journals, ledgers, receivables, and account balances are derived from data describing exchange and conversion REA processes. All reports based on accounting artifacts are always consistent, because they are derived from the same data by [6]. For example, data describing a sale event is used in warehouse management, payroll, distribution, finance and other application areas, without transformations or adjustments. REA ontology also benefits from the presence of a semantic and application independent data model, an object oriented perspective, and abstraction from technical and implementation details. These features enable the possibility of calculating the value of the enterprise's resources on demand, as opposed to calculation at pre-determined intervals.

Apart from keeping track of the past and current economic events, an REA model has the capability of modeling economic events that will occur in the future. However, the principal feature of REA ontology that originates from accountancy systems is that it explicitly distinguishes between past and current events and events performed in the future. And thus for events that occur in the future, REA introduces a new entity called a commitment entity. For this reason, the utilization of a commitment entity is not obligatory but depends only on the specific modeling context. That is, if the model does not address 'future events' there is no need for modeling a commitment entity. In addition, in many cases, only the data concerning real production, which means data about economic events and directly related entities (economic resources, economic agents), are utilized for further processing. The operational level is the part of the REA model which deals with the past and current events. The policy level is the upper part of the REA model which addresses the future events for which the commitment entity and contract entity are necessary. Although the operational level of the REA model can exist independently, the policy level of the REA model can exist only when mutually bound with the operational level. An REA model is illustrated in Fig. 3.

The commitment entity addresses the issue of modeling promises of the future economic events and the issue of the resources reservation, see [5]. Commitment entities and their relationships with other entities are shown in Fig. 3. This figure shows that the commitment entity copies the structure of the event entity to a considerable extent, by which we mean the existence of an increment and decrement commitment and exchange reciprocity relationship.The relationship of *committed provide* and *committed receive* means that some level of agreement about the future exchange must be achieved between economic agents. The *exchange reciprocity* relationship between the increment and decrement commitments identifies which resources are *promised* to be exchanged for which others. The reciprocity relationship is a relation of many-to-many (1..*, 1..*), see [6].

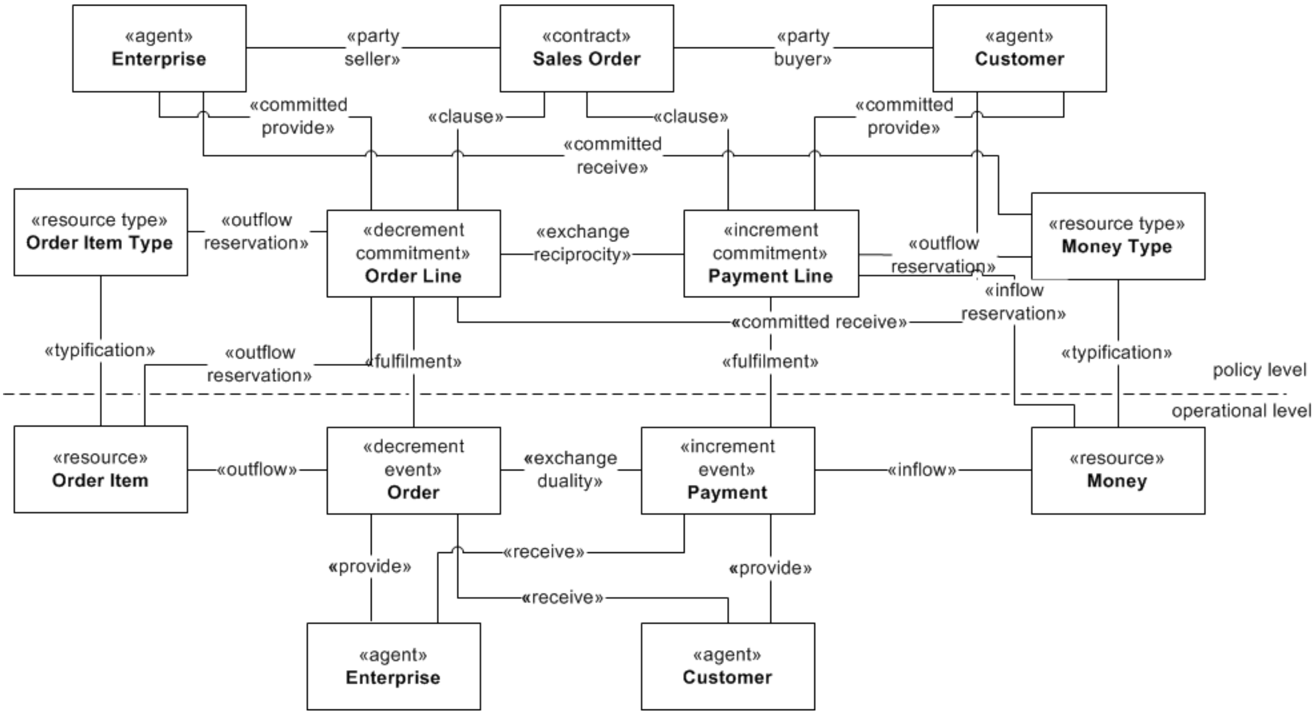


Fig. 3 REA model of an exchange process Source: adapted from [6]

Each commitment is related to an economic resource by a *reservation* relationship, which specifies which resources will be needed or expected by future economic events. The reservation relationship between the resource and commitment entities represents the obligation of economic agents to provide or receive rights to economic resources in exchange processes and represents scheduled usage, consumption or production of economic resources in conversion processes.

In its basic form, an REA transaction is represented by an economic event, an economic resource and a pair of economic agents. The economic event represents an event that happens or has happened in the past. REA transactions are related to each other by a duality relationship which is located at the operational level. In order to model future economic events, an REA model has to be extended with the policy level. This level contains commitment entities, contract entities and resource type entities. In short, a resource type entity represents a category item whereas a resource entity represents a physical item.

## IV. Rent-a-Car Example

This practical and probably familiar example can elucidate the differences and common issues of both methodologies. The Rent-a-car example covers both current and future events and it is not too complex to comprehend. To introduce the problem a short narrative description follows. Rent-a-car is a service which is provided either to walk-in customers or customers who make a rental reservation by telephone, fax or email. A car may be rented on the same day or may be reserved for a specific term in the future after a contract between an employee of the rental company and a customer has been signed. The company which rents out cars has many branches around the country. So the rented car may be picked up and dropped of at different branches. The rental payment depends directly on the number of days of rental and kind of car rented. The signed contrast states, among other things, the period of the rental and the name of the branch where the car will be dropped off. If the period of rental or/and the drop off branch do not coincide with the conditions in the signed contract, the customer is liable for a penalty payment. The contracted payment must be made by the starting day of the rental at the latest. Additional penalty payments must be made at the drop off point.

### *DEMO Solution*

This solution comprises the Construction model as it is fully in compliance with the necessity of both approaches comparison. The Construction model requires identification of the actor roles, transaction kinds and product kinds. In terms of actor roles there is an actor role CA0 who represents an employee of the rental company. The Construction model is illustrated in Fig. 4.

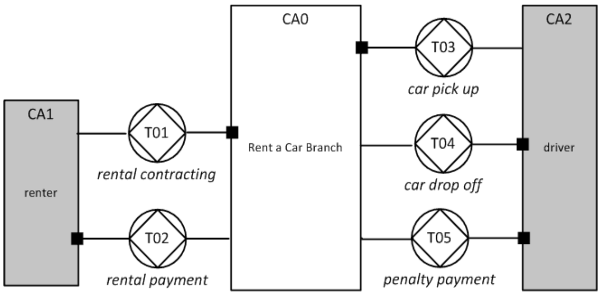


Fig. 4 Demo Construction model Source: [4]

The other actor roles represent a renter and a driver respectively. Utilizing the actor role principle provides a more accurate modeling perspective because a human being can have more than one actor role and one actor role can be represented by different human beings.As it is not a complex example, it is not difficult to identify essential transactions. The first transaction T01 - rental contracting covers signing a contract to rent a car. The next transaction T02 – rental payment must be promised before the contract is signed and this must be done by the first day of the rental at the latest. The car pick-up transaction (T03) includes the promise of the rental company employee regarding the starting conditions of the rental. The other part of this transaction covers the pick up of the car. The car drop-off transaction (T04) is composed of two parts. In the first part the customer promises to observe the conditions specified in the contract. The second part of the transaction represents the actual drop-off the car at a branch of the rental company. The penalty payment transaction (T05) is an optional transaction which is executed if the driver exceeds conditions agreed in the contract.

Going more carefully through the transactions it can observe that coordination steps (act & fact) enable mutual interconnecting (enclosing) of the transactions. The rental contracting transaction (T01) includes entering into and complying with a mutual commitment regarding the transaction T03, the car pick-up and transaction T04, the car drop-off. Both transactions involve detailed specifications about the beginning and end of the rental in the form of commitments. From the narrative description it is clear that the T02 rental payment transaction must be committed before the signing of the rental contract. This is due to the fact that in the text description the renter has to complete the payment by the first day of the committed car rental at the latest. The renter and the employee of the branch must also enter into and comply with any penalty payment to be paid if the period of rental is exceeded or the car is not returned to the branch where it should be dropped off. From this analysis it follows that at the rental contracting stage only the T01, rental contracting transaction, is accepted. The rental payment transaction (T02) must be paid by the starting day of the rental at the latest. The other transactions are usually promised but not yet accepted. The T03 car pick-up transaction is accepted only on the day that the rental starts.

As can be seen from the example, DEMO strictly follows actual conditions and provides great flexibility for a true description of a modelled reality. For example, in some cases the rental payment can be paid at the end of the rental or an advance payment must be made before the rental. This ability of DEMO is enabled by the distinction between the coordination and production acts and fact in a transaction. DEMO methodology is also beneficial when an immaterial product such as a judgement or a schedule is created within the transaction.

*REA Solution*

An REA solution can be modelled in two variations. The first variation represents the operational level of the REA model, whereas the second variation stands for both the operational and policy levels. The first variation enables the modeling of current and past economic events, whereas the second variation also enables the modeling of past, current and future events, described in [2]. Economic events, as a basic part of the REA transactions, must be identified first. Next, the corresponding economic resource and economic agents must be found. Once this has been done, a duality relationship between economic events must be established. The REA value modeling approach usually works with at least one pair of transactions which must be complementary to each other (give and take).

The first REA transaction is created by the rent-a-car economic event, the rent-a-car service economic resource and the renter and rental car company economic agents. The renter agent receives the rental service for a given period of time and the rental car company provides the car. The second transaction is formed by the rental payment economic event, the money economic resource and the pair of economic agents, the same as in the first transaction. What differs is the relationships between the agents and the events. In this case, the rental agent provides the money resource and the rental car company receives the money resource. The third transaction is similar to the second transaction apart from an additional event, which is the penalty payment economic event. These three transactions are related to each other by a duality relationship. This relationship allows us to keep track of which resources have been exchanged for which others.

The REA model illustrated in Fig. 6 represents the second variation of an REA model which includes future events. These events are in the REA approach represented by the commitment entities. The contract entity is composed of the commitment entities and terms. Under the conditions specified by the terms, a contract can create additional commitments. Thus, the contract can specify what should happen if the commitments are not fulfilled. Economic agents are related to the contract entity by the party relationship.

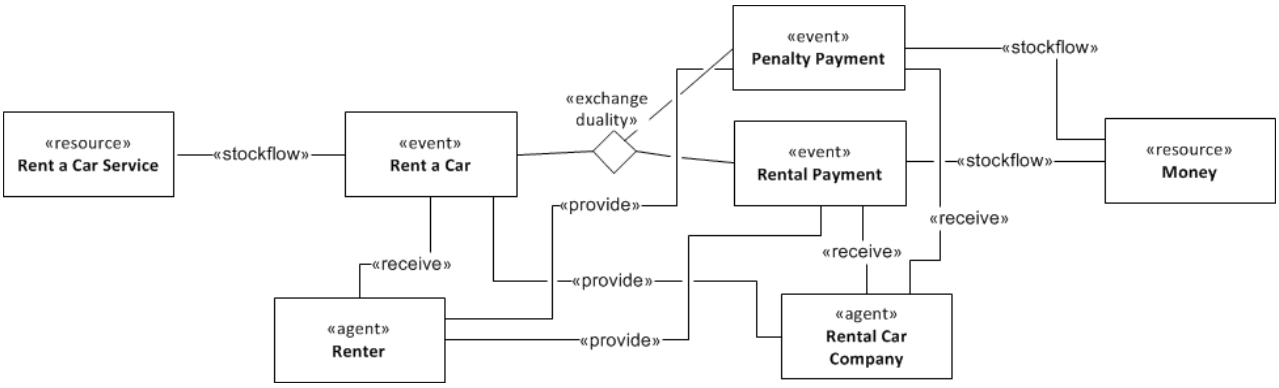


Fig. 5 REA model – operational level (first variation) Source: authors

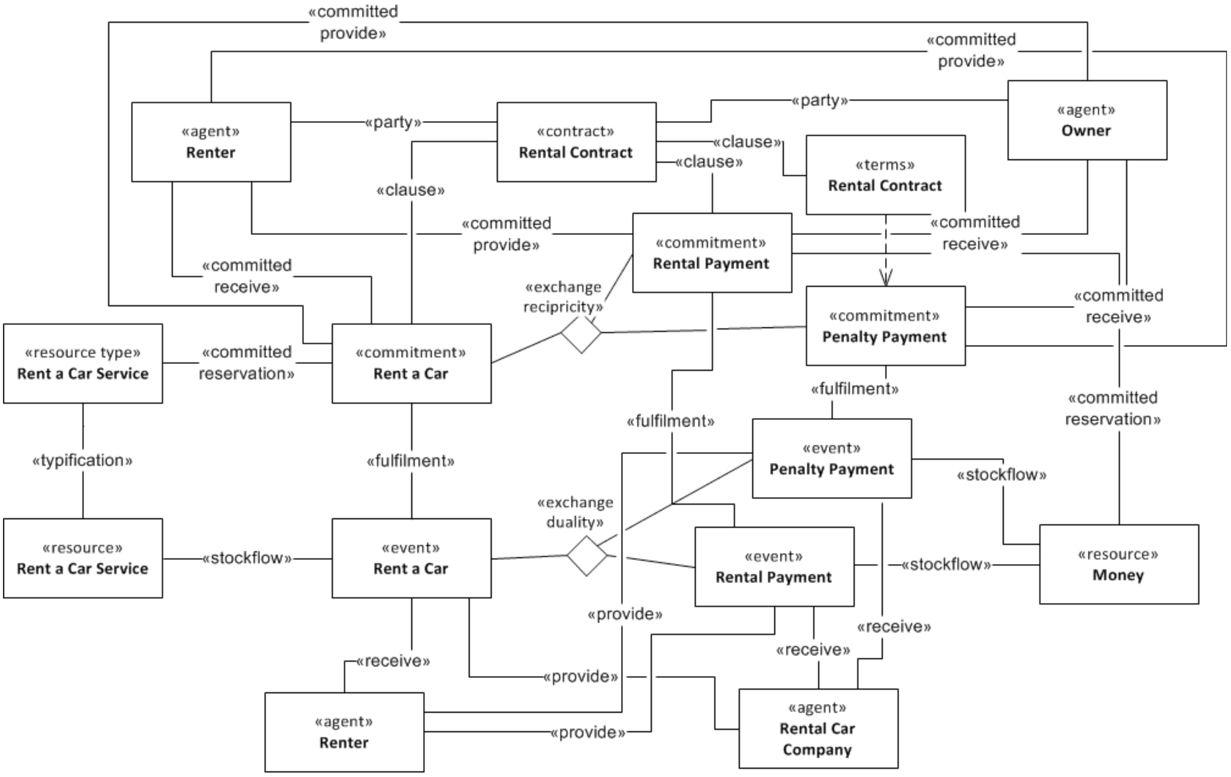


Fig. 6 REA model (second variation) Source: authors

The structure of the commitment entities corresponds with the structure of the event entities. The commitment entities are related to each other by a reciprocal relationship. The practical meaning of the reciprocal relationship is that the future events have been promised and that the resources have been reserved for the future events. The REA solution copies the domain rules.

## V. Discussion

Although, the DEMO transaction and the REA transaction have similar formal meaning their semantic representations are significantly different. This is caused by the fact that the DEMO methodology represents a generic ontology, whereas the REA modeling approach stands for a domain-specific ontology. DEMO methodology utilizes a 'single' transaction as a basic building block from which a business process is composed. The operation axiom divides actions into coordination and production actions and each action is concluded by the result (fact). The DEMO transaction model contains cancellation and revoking operations and is far the most suitable for modeling real world. The DEMO transaction integrates past, current, and future events into one consistent unit. Transactions in a business process are organized in a tree structure, which is far closer to a domain model than a flat sequential structure used by other business process methodologies, including the REA modeling approach.

Among others, it is the operation axiom that has brought DEMO far closer to the world of human beings. Coordination actions enable human beings to enter into and comply with commitments. A production fact comes into existence after it is accepted by the corresponding coordination fact. Being a generic ontology, DEMO knows about real world events with good empirical evidence but it has limited means to deal with production acts and facts.

Contrary to DEMO methodology, the REA value modeling approach does not distinguish between coordination and production activities. Above all, this modeling approach keeps track of the value of economic resources. These resources can either be exchanged for other economic resources or be converted to different economic resources. The REA model, which represents a business process, is composed of at least two mutually binding transactions representing 'give' and 'take' operations in an exchange process, or 'use', 'consume' or 'produce' operations in a conversion process. In an analogy with the DEMO building block, the REA building block is a structure that connects at least two semantically bound transactions. By semantically it is meant that the binding transactions are in compliance with economic laws for the exchange or conversion of economic resources. Despite the benefits of the REA approach which were summarized in the introductory section, there are several drawbacks which prevent this ontology from becoming more widespread. Among REA drawbacks is the explicit separation of past and current events and future events [7]. In the REA model, two separate very similar structures exist. One belongs to the commitment entities and the other to the event entities. It is said that there is mirroring between event and commitment structures [2]. Both structures relate to agent and resource entities. Despite the fact that the REA approach explicitly defines a specific contract entity there is still the open issue of how to create this type of entity which is immaterial in character. Another REA weak point is the absence of a clear state machine declaration. Currently, states are derived from the states of resources during exchange or conversion processes.

There is an idea of how both ontologies can be utilized in mutual accordance. DEMO is fully aware of everything that happens in the real world with high empirical evidence. Utilizing DEMO on the most generic level would mean having a modeling approach that covers broad modeling abilities and is closer to human beings. The REA modeling approach could be exploited to a higher degree and closely cooperate with the generic level. REA approach would be utilized due to domain specific knowledge and the ability to solve its specific production actions. However, this idea would require simulating the REA model (its coordination part) by the DEMO methodology. The problematic issue relating to this idea is the relatively large variety of the 'give' and 'take' transactions which are interrelated. In addition, these transactions may be delayed in time.

## VI. Conclusion

The aim of the paper was to show and with a simple example demonstrate the similarities and differences between the DEMO and REA concepts of transaction. The stated example illustrates the DEMO approach, which, with simple transactions, can model future commitments and in this way contracted actions. The REA value modeling approach that distinguishes between past, current and future events, and uses an explicitly defined contract entity for this reason. A contract entity contains related commitments and commitments that would be instantiated in exceptional circumstances. The paper also considered the idea of mutual collaboration of these different approaches of business process modeling.

## Acknowledgment

This paper was supported by the grant reference no.SGS14/PrF/2015 provided by Ministry of Education, Youth and Sports of the Czech Republic.

## References

[1] G. L. Geerts, and W. E. McCarthy, "The Ontological Foundation of REA Enterprise Information Systems". Paper presented at the Annual Meeting of the American Accounting Association, Philadelphia, PA., 2000.

[2] G. L. Geerts, W. E. McCarthy, "Polocy-Level Specification in REA Enterprise Information Systems". *Journal of Information Systems*. Vol 20, No. 2, 2006, pp. 37-63. DOI: 10.2308/jis.2006.20.2.37

[3] J. L. G. Dietz, "Enterprise Ontology – Theory and Methodology". Springer-Verlang, 2006.

[4] J. L. G. Dietz, "The Essence of Organization. An Introduction to Enterprise Engineering". Sapio bv, 2012.

[5] Ch. L. Dunn, O. J. Cherrington, and A. S. Hollander, "Enterprise Information Systems: A Pattern Based Approach". New York: McGraw-Hill/Irwin, 2004.

[6] P. Hruby, "Model-Driven Design Using Business Patterns". Springer-Verlang, 2006.

[7] F. Hunka, and J. Zacek, "Detailed Analysis of REA Ontology", *Lecture Notes in Busines Information Processing,* Vol. 174, 2014, pp. 61-75. DOI: 10.1007/978-3-319-06505-2

[8] R. Klimek and P. Szwed, "Verification of ArchiMate Process Specification Based on Deductive Temporal Reasoning". *Proceedings of the 2013 Federated Conference an Computer Science and Information Systems.* pp. 1103-1110.

[9] R. Wendler, "Delelopment of the Organizational Agility Maturity Model". *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems. Vol 2, 2014*, pp. 1197-1206. DOI: 10.15439/2014F79

DOI: 10.15439/2015F245
ACSIS, Vol. 5

# Critical Success Factors for Implementing Supply Chain Management Systems – The Perspective of Selected German Enterprises

*Full Paper*

Christian Leyh
Technische Universität Dresden,
Chair of Information Systems, esp. IS in
Manufacturing and Commerce,
Helmholtzstr. 10, 01069 Dresden, Germany
Email: Christian.Leyh@tu-dresden.de

Julia Thomschke
Technische Universität Dresden,
Chair of Information Systems, esp. IS in
Manufacturing and Commerce,
Helmholtzstr. 10, 01069 Dresden, Germany

***Abstract* — *The aim of our study was to provide a contribution to the research field of critical success factors (CSFs) with a focus on SCM system implementations. Therefore, we conducted a systematic literature review in order to identify CSFs for those projects. On the basis of that review, we conducted interviews within German large-scale enterprises and with consultants experienced with SCM system implementations. As a result, we showed that all the factors found in the literature also affected the success of SCM projects in the studied companies. Additionally, we were able to identify six further CSFs with the interview study. However, within those SCM projects, technological factors gained more importance compared to those factors which influence the success of ERP projects the most. For SCM projects, factors like Data migration, as well as SCM system tests, are even more important than Top management support or Project management, which are the most important factors for ERP projects.***

## I. Introduction

Today's enterprises are faced with the globalization of markets and rapid changes in the economy. In order to cope with these conditions, the use of technology, as well as information and communication systems, is almost mandatory. Specifically, the adoption of enterprise resource planning (ERP) systems as standardized systems that encompass the activities of an entire enterprise has become an important factor for today's businesses. The demand for ERP applications has increased for several reasons, including competitive pressure to become low-cost producers, expectations of revenue growth, and the desire to re-engineer businesses to respond to market challenges. A properly selected and implemented ERP system offers several benefits, such as considerable reductions in inventory costs, raw material costs, lead time for customers, production time, and production costs [1]-[4]. Therefore, the majority of enterprises around the world use ERP systems.

However, using ERP systems to optimize "the inside conditions" of an enterprise still may not be sufficient to be competitive in today's business environment. In the last decade and continuing through today, the logistics of an enterprise are strongly influenced by changing business conditions. These conditions have changed dramatically in recent years — e.g., customers expect that companies will respond quickly to their needs and wishes and that these companies will also be very flexible. In addition, the customers more frequently are requiring widely different product variants. In addition to this required variety, there is also a trend towards increasing collaboration with a variety of suppliers and the inclusion of a wide variety of distribution channels, mostly on an international level, caused by increasing globalization. To cope with these requirements, the optimization of enterprises and their processes has to extend beyond the companies' borders; therefore, information systems have to be able to cross these borders, too. Here, the focus lies on optimal planning, management and control of the material, and information flow across the entire value chain [5]. Confronted with these issues, ERP systems are often reaching their functional limits.

One approach to dealing with the needs for joint planning of production and logistics activities across enterprise borders is the supply chain management (SCM) concept. To provide the needed relevant information for adequate planning and calculations and to support the exchange of information between supply chain partners, information systems are essential components within the SCM concept [6]-[8]. As supply chain-wide information systems, SCM systems are becoming increasingly important for enterprises. The increasing importance of SCM systems is also emphasized by the fact that many companies have already implemented such systems, or at least plan to implement these systems in the near future. Therefore, the use of and the need for adequate SCM systems has increased in recent years [9].

However, the implementation of an information system (e.g., an ERP system or an SCM system) is a complex and time-consuming project during which companies face great opportunities, but at the same time also face enormous risks. To take advantage of the potential opportunities rather than get caught by the risks of these implementation projects, it is essential to focus on those factors that support the successful implementation of an information system [10], [11]. By

978-83-60810-66-8/$25.00©2015, IEEE

being aware of these factors, a company can positively influence the success of their implementation project and effectively minimize the project's risks [10]. Recalling these so-called critical success factors (CSFs) is of high importance whenever a new information system is to be adopted and implemented or a running system needs to be upgraded or replaced. Errors during the selection, implementation, or maintenance of information systems, wrong implementation approaches, or systems that do not fit the requirements of the enterprise can all cause financial disadvantages or disasters, perhaps even leading to insolvency. E.g., when considering ERP implementation projects, several examples of negative scenarios can be found in the literature (e.g., [12], [13]). Here, SCM implementation projects can result in an even more complex project structure since the companies face not only enormous internal challenges but many external challenges as well.

However, literature dealing with SCM projects and their critical success factors can only rarely be found, whereas CSFs of ERP projects have already been considered in numerous scientific publications. Several case studies, surveys, and literature reviews on CSFs of ERP projects have been conducted by different researchers (e.g., [4], [14]-[17]).

Hence, considering the increasing importance of the use of SCM systems, the aim of our study was to focus on the implementation of SCM systems, focusing in particular on the differences in CSFs of ERP projects and SCM system implementations. Therefore, we conducted a systematic literature review in order to identify CSFs for SCM projects and to update the existing reviews of CSFs. On the basis of the CSFs we identified, we conducted multiple interviews within German enterprises which have already implemented an SCM system, as well as with consultants from SCM manufacturers with specific experience in SCM projects, in order to obtain insights into the similarities and differences among CSFs for SCM system implementations. Overall, our study was driven by the following research questions:

Q1: What are the critical success factors of SCM system implementations?

Q2: What similarities and differences exist between critical success factors for ERP implementation projects and SCM implementation projects?

Therefore, the paper is structured as follows. The next section deals with the results of our literature review. We will point out which factors are most important and which factors seem to have little influence on the success of an SCM implementation project. Next, our data collection methodology is described before the results of the interviews are presented and discussed and the research questions are answered. Finally, the paper concludes with a summary of the results and discusses the limitations of our study.

## II. Literature Review of Critical Success Factors for SCM projects

A critical success factor is defined according to [15] as any condition or element that is seen as necessary in order for the system implementation to be successful. As mentioned in the introduction, in order to identify factors that affect the success or failure of ERP projects, several case studies, surveys, and literature reviews have already been conducted by a number of researchers (e.g., [15], [16], [18]). However, most of the literature reviews cannot be reproduced, because descriptions of the review methods and procedures are lacking. Some researchers have pointed out the limitations of the currently available literature review articles, specifically noting that they lack methodological rigor [19]. Therefore, in order to gain insight into the field of CSFs for SCM projects, we conducted a literature review by systematically reviewing articles in five different databases, as well as papers drawn from several international conference proceedings. The literature review to identify the CSFs was performed in several steps, similar to the approach suggested by [20]. Here, we adapted an approach which we have previously used to update the existing CSF frameworks for ERP projects (see [17], [21]).

The steps of our review procedure are presented below. An overview is given in Figure 1 with regard to the numbers of papers identified or remaining during/after each step. With each step, the number of papers was reduced according to the assembly of different criteria.

**Step 1 & Step 2:** The first two steps were to define the sources for the literature review and the search terms for the database-driven review. Therefore, several databases and conference proceedings were first identified. Keywords selected for this search were mostly derived and adapted from the keywords we used for our systematic review of the ERP CSF literature [17], [21]. To make our review reproducible, we have listed the databases and search terms in Table 1.

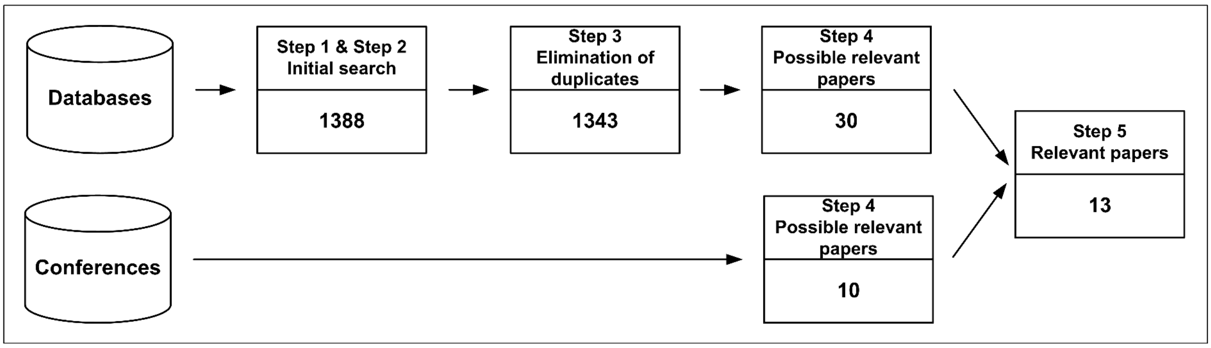


Figure 1. Progress of the literature review

TABLE 1. SEARCH FIELDS AND SEARCH TERMS FOR THE DATABASE-DRIVEN REVIEW

<table>
<tr><th>Database + Search fields</th><th colspan="2">Search terms / Keywords</th></tr>
<tr><td><b>Academics Search Complete:</b><br>“TI Title” OR “AB Abstract or Author supplied abstract”</td><td rowspan="5">SCM + system + success*<br>SCM + system + fail*<br>SCM + system + crit*<br>SCM+ system + fact*<br>SCM + system + csf<br>SCM + system + cff<br><br>APS + system + success*<br>APS + system + fail*<br>APS + system + crit*<br>APS + system + fact*<br>APS + system + csf<br>APS + system + cff</td><td rowspan="5">supply chain management + system + success*<br>supply chain management + system + fail*<br>supply chain management + system + crit*<br>supply chain management + system + fact*<br>supply chain management + system + csf<br>supply chain management + system + cff<br><br>advanced planning and scheduling + system + success*<br>advanced planning and scheduling + system + fail*<br>advanced planning and scheduling + system + crit*<br>advanced planning and scheduling + system + fact*<br>advanced planning and scheduling + system + csf<br>advanced planning and scheduling + system + cff</td></tr>
<tr><td><b>Business Source Complete:</b><br>“TI Title” OR “AB Abstract or Author supplied abstract”</td></tr>
<tr><td><b>Science Direct:</b><br>“Abstract, Title, Keywords”</td></tr>
<tr><td><b>SpringerLink:</b><br>“Title” OR “Abstract”</td></tr>
<tr><td><b>WISO:</b><br>“General search field”</td></tr>
</table>

Since the WISO database also provides German papers, we used the German translation of most of the search terms as well. For the conferences, only inappropriate search fields and search functionality were provided.

Hence, we decided to review the abstracts and titles of the conference papers in this step manually. We used the proceedings of four conferences:

- International Conference on Information Systems (ICIS)
- Americas Conference on Information Systems (AMCIS)
- European Conference on Information Systems (ECIS)
- Wirtschaftsinformatik (WI)

**Step 3:** During step 3 we performed the initial search according to step 1 and step 2, and afterwards eliminated duplicate results. The initial search provided 1,388 papers from the databases. After eliminating the duplicates, 1,343 articles remained. From the conference search, 10 papers remained. Altogether, 1,353 papers were identified during the initial search step.

**Step 4:** Step 4 included the identification of irrelevant papers. During the initial search, we did not apply any restrictions. The search was not limited to the research field of IS; therefore, papers from other research fields were also included in the results. Thus, these irrelevant papers had to be excluded. Additionally, the majority of SCM papers focused on the implementation of the supply chain management concepts itself without dealing with information systems. Therefore, these papers had to be excluded as well. The identification of these irrelevant papers was done by reviewing the abstracts of the papers and, if necessary, by looking into the paper’s content. Of the 1,353 papers, only 30 stemming from the database search remained, along with all 10 conference papers. Together, this yielded a total of 40 papers that were potentially relevant to the field of CSFs for SCM system implementations (see Figure 1).

**Step 5:** The fifth and final step consisted of a detailed analysis of the remaining 40 papers and the identification of the CSFs. Therefore, the content of all papers was reviewed in depth for the purpose of categorizing the identified success factors. Emphasis was placed not only on the wording of these factors, but also on their meaning. After this step, only 13 relevant papers that suggested, discussed or mentioned CSFs of SCM projects remained. In five of these 13 papers, CSFs of SCM projects were directly focused on within the conducted investigation, whereas in the other eight papers CSFs were discussed but these factors were not explicitly investigated with empirical studies. For each paper, the CSFs were captured along with the publication year, the type of data collection used, and the companies (e.g., the number and size) from which the CSFs were derived.

All 13 papers were published between 2000 and 2012. Table 2 shows the distribution of the papers by publication year. As is shown, there are not many papers published per year that deal with CSFs of SCM system implementation projects.

TABLE 2. PAPER DISTRIBUTION BY YEAR

| Year | Papers | Year | Papers |
|---|---|---|---|
| 2012 | 1 | 2005 | 1 |
| 2011 | | 2004 | 2 |
| 2010 | | 2003 | |
| 2009 | 3 | 2002 | |
| 2008 | 1 | 2001 | |
| 2007 | | 2000 | 2 |
| 2006 | 3 | | |

Overall, 22 factors influencing the success of SCM system implementations could be identified within the literature review. Table 3 shows the results of our review, i.e., the CSFs identified, ordered by each factor's total number of occurrences in the reviewed papers.

The factors *Top management support and involvement*, *Compatibility of the SCM system*, and *Data migration / data accuracy* are the three most-named factors, with each being mentioned in seven articles.

However, the differences in the CSF frequencies are only minimal and are related to the small number of identified papers. Therefore, to derive CSFs that are important for SCM implementation projects and to realize their different levels of importance is difficult due to the small number of studies focusing solely on these CSFs. Here, the minimal focus paid to SCM system implementations can be clearly seen as a research gap in the SCM system research field. To gain a deeper insight into this research field and to identify further factors not mentioned in the small number of articles, we set up an empirical study focusing on CSFs of SCM system implementations. We investigated these CSFs in depth by interviewing experienced SCM consultants as well as enterprises that have already implemented SCM systems. The results of this interview study will be part of the following sections.

Due to space constraints, detailed descriptions and definitions for each identified factor cannot be given within this article, but will be provided by the first author upon request.

TABLE 3. SCM PROJECT CSFS IN RANK ORDER BASED ON FREQUENCY OF APPEARANCE IN ANALYZED LITERATURE

| Critical Success Factor | Number of papers | Critical Success Factor | Number of papers |
|---|---|---|---|
| Top management support and involvement | 7 | Available resources | 3 |
| Compatibility of the SCM system (with other information systems and the IT infrastructure) | 7 | (Organizational) Fit of the SCM system | 3 |
| Data migration / data accuracy | 7 | Involvement of end-users and stakeholders | 3 |
| User training | 6 | External consultants | 3 |
| Balanced project team | 5 | SCM system tests | 3 |
| Project management | 4 | SCM system acceptance / resistance | 2 |
| Change management | 4 | Environment and organizational culture | 2 |
| Clear goals and objectives | 4 | Project team leadership | 2 |
| Company's strategy / strategy fit | 3 | Use of a steering committee | 1 |
| Cooperation with supply chain partners | 3 | Skills, knowledge, and expertise | 1 |
| Communication | 3 | Vendor relationship and support | 1 |

## III. Interview Study – Critical Success Factors for SCM implementation projects

### I. *Study Design – Data Collection Methodology*

To gain an empirical insight into SCM implementation projects and to gain an understanding of the CSFs for those projects, we used a qualitative exploratory approach within German large-scale enterprises and within German SCM system manufacturers. We chose to focus on German companies due to our cultural background.

The units of analysis in our study are the implementation projects carried out within the enterprises, as well as the SCM projects that the consultants have performed thus far in their careers. For the data collection, we conducted several interviews with members of the SCM implementation project teams and with consultants from German SCM system manufacturers in order to identify the factors that they determined to be relevant for the success of the projects.

During this process, we interviewed employees from nine large-scale enterprises located in Germany. The companies operate in different industry sectors and have implemented different SCM systems. Table 4 gives an overview of the companies and the interviewees. Within these enterprises, different SCM systems have been implemented (which cannot be named directly within this paper for data protection reasons).

Table 4. Overview of the Large-scale Enterprises and Interviewees (Number of Employees are Categorized due to Data Protection)

| Company | Industry sector | Number of employees | Interviewee |
|---|---|---|---|
| C 1 | Automotive industry | 1,000 – 5,000 | Head of the IT department |
| C 2 | Manufacturing of metal goods / Machine-building industry | 100 – 500 | Head of the *materials management* department |
| C 3 | Automotive industry | 1,000 – 5,000 | Head of the *company organization* department |
| C 4 | Consumer goods industry | > 15,000 | Project leader of the SCM project |
| C 5 | Automotive industry | 1,000 – 5,000 | Head of the *accounting and IT* department |
| C 6 | Building services engineering | > 15,000 | Business unit manager for *SCM* |
| C 7 | Construction industry | > 20,000 | Head of the *materials management* and *IT* departments |
| C 8 | Construction industry | 1,000 – 5,000 | Project manager for *logistics* |
| C 9 | Electronics industry | > 20,000 | Business unit manager for *supply chain innovations* |

Four companies have implemented the same SCM system; all the other companies have implemented different systems — some quite small and industry-specific systems and some more widespread systems on the SCM market. Most of the implementation projects took place in the mid-2000s. All of the interviewees were somehow directly involved in their respective companies' SCM system implementation projects.

In order to gather and include the SCM system manufacturers' perspectives in our study, we focused on information system manufacturers, with a specific emphasis on SCM implementations. Here, we interviewed three consultants from three different SCM manufacturers. Among the SCM consultants, we were able to interview consultants with longtime experience in several implementation projects.

To gain a deep and detailed view of the enterprises and their structures as well as of the consultants' experiences, we chose semi-structured interviews as our method of data collection. The interviews were conducted in retrospect regarding the SCM projects between July and October 2013. The interviews were designed as partially standardized interviews using open to semi-open questions as initial starting points for the conversation. Both personal (face-to-face) interviews and telephone interviews were conducted by the authors. An interview guideline was developed, based on the questions of [22], who conducted a similar study within the field of ERP implementation projects, and also on the basis of one of our previous CSF studies (in the field of ERP projects), which focused on smaller companies and their experiences in ERP implementations [21]. We changed the questions to align with our identified CSFs (see Table 3) to ensure that all of the factors were discussed in the interviews. The interview guideline consisted of more than 30 main questions with further sub-questions that referred to the identified CSFs. These questions were formulated in an open way, so that it would be possible to identify "new" CSFs from the interviews that were not identified in the literature review. This questionnaire was sent to interviewees prior to the interviews to allow them to prepare for their interviews. The complete listing of the formulated questions and their assignment to the success factors is not included in this paper, but will be provided by the first author upon request.

For a better analysis of the results, we recorded all of the interviews (the interviews typically took between 60 and 90 minutes) and transcribed them afterwards. As a first step, non-verbal and para-linguistic elements and other elements that were not relevant to the study were excluded. Next, in order to evaluate the CSFs, the interviews were analyzed with reference to each CSF question block. We matched the answers and statements of the interviewees to the respective factors. Therefore, we had to formulate respective coding rules. Afterwards, each CSF was ranked according to a three-tier scale (2–very important factor; 1–medium important factor; 0–less/non-important factor) and, for a finer classification, according to a five-tier scale (4–very important factor; 3–important factor; 2–factor was seen as relevant; 1–factor was mentioned but not seen as being very relevant; 0–factor was not seen as relevant or important/factor was not mentioned at all). This rating was done regarding the respective statements of the interviewees. We used these two scales to gain a preliminary understanding of whether differences would occur by using a finer/more detailed scale. Here, the five-tier-scale could be seen as more appropriate for determining the different levels of importance for the factors. After setting up this ranking of CSFs, we discussed the factor rating with other researchers in this field to reduce the subjectivity of the rating. Finally, this procedure resulted in a ranking of the CSFs according to the interviewees' statements and answers.

## *II. Results of the Interviews*

For each interview, a ranking of the critical success factors was set up by the authors. A final ranking was created including all interviews and all individual rankings (see Table 5). As shown, the top three most important factors for SCM system implementation projects according to our study are *Change management*, *Data migration / data accuracy*, *SCM system tests* and *Available resources* with around or above 35 out of possible 48 points. Each of the 22 factors stemming from the literature review was mentioned by at least one interviewee.

Additionally, six further factors (*Organizational structure*, *Business process reengineering*, *Troubleshooting*, *Knowledge management*, *Project champion* and *Vendor's tools and implementation methods*) could also be identified during the interviews. These factors are printed in bold in Table 5.

However, four of these additional factors seem to have less influence on the success of SCM implementation projects, since they are ranked as 20 or lower. Only *Organizational structure* and *Business process reengineering*—with nearly or above 30 out of 48 possible points—seem to have at least a medium influence on the SCM project success.

To categorize critical success factors, [18] suggest a matrix scheme. Here, they consider the tactical or strategic direction of the CSFs and divide them into organizational and technological factors [18]. Thus, tactical CSFs relate instead to short-term aspects and goals of the system implementation project itself, whereas strategic factors aim at the long-term impacts of activities with strong connections to the development of the organization in relation to the mission, vision and core competencies of the business activity. Considering the technological and organizational character of the CSFs, the specificity and significance of technological factors are strongly dependent on the SCM systems themselves, whereas organizational factors focus on corporate culture and its environment with its specific processes and structures [18], [23], [24]. Table 6 gives an overview of the categorization of the top twelve of the identified CSFs in our study with a focus on their ranking.

TABLE 5. CSFS ACCORDING TO THE FIVE-TIER-SCALE RATING

| **Rank** | **Factor** | **Factor rating (five-tier-scale)** | **Rank** | **Factor** | **Factor rating (five-tier-scale)** |
|---|---|---|---|---|---|
| **1** | Change management | 38 | **15** | Project team leadership | 27 |
| **2** | Data migration / data accuracy | 36 | | Vendor relationship and support | 27 |
| **3** | SCM system tests | 35 | | Communication | 27 |
| | Available resources | 35 | | Compatibility of the SCM system (with other information systems and the IT infrastructure) | 27 |
| **5** | Top management support and involvement | 32 | **19** | Company's strategy / strategy fit | 26 |
| | SCM system acceptance / resistance | 32 | **20** | **Troubleshooting** | 25 |
| **7** | User training | 31 | **21** | **Knowledge management** | 24 |
| | Project management | 31 | | Environment and organizational culture | 24 |
| | Skills, knowledge, and expertise | 31 | **23** | Involvement of end-users and stakeholders | 19 |
| | Clear goals and objectives | 31 | **24** | Cooperation with supply chain partners | 18 |
| | **Organizational structure** | **31** | **25** | External consultants | 17 |
| **12** | Balanced project team | 30 | **26** | Use of a steering committee | 14 |
| **13** | (Organizational) Fit of the SCM system | 28 | | **Project champion** | 14 |
| | **Business process reengineering** | 28 | **28** | **Vendor's tools and implementation methods** | 7 |
| 4–very important factor; 3–important factor; 2–factor was seen as relevant; 1–factor was mentioned but not seen as being very relevant; 0–factor was not seen as relevant or important/factor was not mentioned at all) / maximum possible rating on the basis of 12 interviews = 48 | | | | | |

We oriented around the classification and categorization of the factors according to [23], [24]. The factors of the top three are highlighted. It is shown that only a few CSFs (2 out of the top 12) are technological factors, whereas more than 50% of the factors (7 out of the top 12) are organizational factors with a strategic characteristic. However, the top 12 factors are spread out among all four categories, although most of them are part of the organizational category. Remarkably, two of the most important factors are part of the technological view.

TABLE 6. CATEGORIZATION OF CSFS (MODEL ADAPTED FROM [18], [23], [24])

| | Strategic | | Tactical | |
|---|---|---|---|---|
| | Critical Success Factors | Rank | Critical Success Factors | Rank |
| Organizational | **Change management** | **1** | User training | 7 |
| | **Available resources** | **3** | Skills, knowledge and expertise | 7 |
| | Top management support and involvement | 5 | Project management | 7 |
| | SCM system acceptance / resistance | 5 | | |
| | Clear goals and objectives | 7 | | |
| | Organizational structure | 7 | | |
| | Balanced project team | 12 | | |
| Techno-logical | | | **Data migration / data accuracy** | **2** |
| | | | **SCM system tests** | **3** |

## IV. DISCUSSION

Regarding these factors, the interviewees mostly acknowledge the descriptions of how the factors are summarized in the literature.

By looking at the rankings of the literature review, both at the ranking of the interview study and as a comparison at the ranking of an ERP CSF literature review from a former investigation (see [21]), the differences become obvious. Table 7 shows the respective top five factors.

As it is shown, the factor *Top management support and involvement* is always part of the top 5 factors. In both literature reviews this factor is the most important factor for the projects' success. Also, *Change management* is seen as a very important factor both for ERP projects and for SCM projects (as it is mentioned as the most important factor within the interview study). However, for SCM projects, technological factors (e.g., *Data migration / data accuracy* and *SCM system tests*) also seem to have a larger influence on the projects than they have on ERP projects. At least these factors are not ranked very high in the literature review for ERP implementations (see [21]).

An additional comparison of the results from the literature analysis and the interview study shows that the interviewees named and described twelve out of the 28 CSFs as they are described and defined in the literature, whether for SCM implementations or for ERP implementations. These factors are:

- Business process reengineering
- Knowledge management
- Change management
- Organizational structure
- Company's strategy / strategy fit
- Project management
- Compatibility of the SCM system
- Project team leadership
- Data migration / data accuracy
- SCM system tests
- Environment and organizational culture
- User training

Regarding the 16 other factors that are discussed by the interviewees in the way other than how they are described and defined in the literature, we cannot discuss the differences of all 16 factors at this point. However, we will point out these differences for some example factors:

- *Top management support and involvement:* The support of the top management was not seen as helpful by all interviewees. Some of them even mentioned that this had a negative influence on the project. For these interviewees, a lower level of top management support and less involvement from the managers were seen as beneficial since this would avoid long decision-making processes and other complications.

TABLE 7. COMPARISON OF THE TOP FIVE FACTORS

| Rank | Results of the SCM projects' literature review (13 papers) | Results of the ERP projects' literature review (320 papers) | Results of our interview study |
|---|---|---|---|
| 1 | Top management support and involvement | Top management support and involvement | Change management |
| | Compatibility of the SCM system | | |
| | Data migration / data accuracy | | |
| 2 | | Project management | Data migration / data accuracy |
| 3 | | User training | Available resources |
| | | | SCM system tests |
| 4 | User Training | Change management | |
| 5 | Balanced project team | Balanced project team | Top management support and involvement |
| | | | SCM system acceptance / resistance |

- *Balanced project team:* According to the literature for this CSF, project teams should consist of fixed and variable project members. However, such project team compositions were only used in two of the interviewed companies. The majority of the enterprises tended to use fixed project teams without variable members. Furthermore, in considering the literature, a high level of experience for the project team members in the field of software implementations is also an important aspect for the project team. However, this was not confirmed by interviewees. Mostly, the project members were not very experienced in this field.

- *Skills, knowledge, and expertise:* In the literature, experience, training and personal capabilities, skills and knowledge are all seen as positive impact factors on the project's success and can help to avoid errors and mistakes during the various project phases, or at least can lead to fast reactions. In the studied enterprises, the users' skills and knowledge (especially with regard to SCM systems or other software implementation projects) were described as really low. But this lack of experience and knowledge did not have any apparent negative impact on the projects' success.

- *Use of a steering committee / Troubleshooting:* Clear differences are also shown with respect to the use of a steering committee. According to the literature, it is almost essential to use a steering committee to manage and to supervise the implementation projects. However, only a few companies in our study had at least established some kind of a steering committee. Most of them did not see the necessity for a steering committee. In addition, the project plans of the companies contained no predefined troubleshooting plans or actions, which are seen as a mandatory aspect for software implementations according to the literature.

- *Communication:* As stated in the literature, in order to have adequate communication within the project teams, within the company itself and among SCM cooperation partners during the entire SCM system implementation, the development of a specific and detailed communication strategy is important. Yet, in the interviewed enterprises communication was done, for example, on a regular basis via meetings or telephone conferences. However, an explicit and specific communication strategy was developed only for one implementation project. Considering the communication, some interviewees also mentioned that it is sometimes easier just to inform the management rather than getting the managers too heavily involved. This again supports the discussion of the factor *Top management support and involvement*.

## V. Conclusion and Limitations

The aim of our study was to gain insight into the research field of CSFs for SCM implementation projects. Research in the field of software implementation projects and their CSFs provides valuable information that can enhance the degree to which an organization's implementation project succeeds [15]. As a first step, we carried out a systematic literature review to identify CSFs for SCM projects. Our review turned up only a small number of papers focusing on SCM system implementations—we could only identify 13 relevant papers in this research field. From these papers, we derived 22 different CSFs. However, compared to the CSF literature for ERP system projects (a separate literature review yielded 320 papers dealing with 31 different CSFs for ERP projects), this can be seen as a clear lack of research.

Here, to gain a deeper insight into this field, we set up an empirical study with a specific focus on SCM system implementations. We conducted several interviews within large-scale enterprises which have implemented SCM systems and with consultants who have several years of SCM project experience. Using a guideline consisting of more than 30 initial questions about CSFs, we conducted twelve interviews. We found that all 22 factors identified in the literature review were mentioned by at least one interviewee. Additionally, we were able to identify six additional factors during the interviews. Therefore, there are 28 factors which somehow affect the success of SCM system projects. However, contrary to the rankings resulting from an earlier literature review for ERP projects (see [17] and [21]), we identified factors with a more technological focus as being important for SCM projects. Here, the factors *Data migration / data accuracy* and *SCM system tests* are two out of the top three CSFs for SCM implementations. Hence, factors with an organizational characteristic could also be identified as part of the top 5 factors in our interview study (*Change management*, *Available resources*, *Top management support and involvement*, and *SCM system acceptance / resistance*).

Regarding research question 1 (Q1), we were able to clearly identify 28 factors that have an influence on SCM system implementations. However, we could also show that the importance of the factors for SCM projects differs from the CSFs' ranking for ERP projects (Q2). The implementing companies as well as the SCM system manufacturers have to be aware of these differences in the factors' characteristics, also focusing on technological aspects of the SCM system implementations rather than focusing mainly/only on the organizational factors.

A few limitations of our study must be mentioned as well. For our literature review, we are aware that we cannot be certain that we have identified all relevant papers published in journals and conferences since we made a specific selection of only five databases and four international conferences. Therefore, journals that are not included in our selected databases and the proceedings from other conferences might also provide relevant articles. Another limitation is the coding and ranking of the CSFs. We tried to reduce any subjectivity by formulating coding rules and ranking rules and by discussing the coding of the CSFs with several independent researchers. However, other researchers may code and assess the CSFs in different ways. For the interview study, the interviews conducted and data evaluated represent only an investigation of sample SCM projects in German enterprises. These results are limited to the specifics of these companies and the particular experiences of the consultants. In light of this, we will conduct further case studies and some larger surveys to broaden the results of this investigation.

## VI. References

[1] T. H. Davenport, *Mission critical: Realizing the promise of enterprise systems*. Boston, USA: Harvard Business School Press, 2000, DOI: 10.1225/9067.

[2] S. V. Grabski, and S.A. Leech, "Complementary controls and ERP implementation success," *International Journal of Accounting Information Systems*, vol. 8, no. 1, pp. 17–39, 2007, DOI: 10.1016/j.accinf.2006.12.002.

[3] S. C. L. Koh, and M. Simpson, "Change and uncertainty in SME manufacturing environments using ERP," *Journal of Manufacturing Technology Management*, vol. 16, no. 6, pp. 629–653, 2005, DOI: 10.1108/17410380510609483.

[4] T. M. Somers, and K. Nelson, "The impact of critical success factors across the stages of enterprise resource planning implementations," in *Proceedings of the 34th Hawaii International Conference on System Sciences (HICSS 2001)*, Hawaii, USA, 2001, DOI: 10.1109/HICSS.2001.927129.

[5] A. Busch and W. Dangelmaier, "Integriertes Supply Chain Management - ein koordinationsorientierter Überblick," in *Integriertes Supply Chain Management / Theorie und Praxis effektiver unternehmensübergreifender Geschäftsprozesse,* 2nd ed., A. Busch and W. Dangelmaier, Eds., Wiesbaden, Germany: Gabler publishing, 2004, pp. 1–24.

[6] H. Bartsch and P. Bickenbach, *Supply-Chain-Management mit SAP APO / Supply-Chain-Modelle mit dem Advanced Planner & Optimizer 3.1*, 2nd ed., Bonn, Germany: Galileo Press, 2002.

[7] H. Krcmar, *Informationsmanagement,* 5th ed., Heidelberg, Germany: Springer, 2001.

[8] G. Mangalaraj, A. Jeyaraj and E. Prater, "Technology Adoption in Supply Chain Management: A Meta-Analysis of Empirical Findings," in *Proceedings of the 12th Americas Conference on Information Systems (AMCIS 2006)*, Acapulco, México, 2006.

[9] M. J. Tarokh and J. Soroor, "Supply Chain Management Information Systems Critical Failure Factors," in *Proceedings of the IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI '06)*, Shanghai, China, 2006, pp. 425–431.

[10] A. Jones, J. Robinson, B. O'Toole, and D. Webb, "Implementing a bespoke supply chain management system to deliver tangible benefits," *International Journal of Advanced Manufacturing*

*Technology*, vol. 30, no. 9/10, pp. 927–937, 2006, DOI: 10.1007/s00170-005-0065-2.

[11] E. W. T. Ngai, T. C. E. Cheng, and S. S. M. Ho, "Critical success factors of web-based supply-chain management systems: An exploratory study," *Production Planning & Control*, vol. 15, no. 6, pp. 622–630, 2004, DOI: 10.1080/09537280412331283928.

[12] T. Barker, and M. N. Frolick, "ERP Implementation Failure: A Case Study," *Information Systems Management*, vol. 20 no. 4, pp. 43–49, 2003, DOI: 10.1201/1078/43647.20.4.20030901/77292.7.

[13] K. Hsu, J. Sylvestre, and E. N. Sayed, "Avoiding ERP Pitfalls," *The Journal of Corporate Accounting & Finance*, vol. 17, no. 4, pp. 67–74, 2006, DOI: 10.1002/jcaf.20217.

[14] P. Achanga, G. Nelde, R. Roy, and E. Shehab, "Critical Success Factors for Lean Implementation within SMEs," *Journal of Manufacturing Technology Management*, vol. 17, no. 4, pp. 460–471, 2006, DOI: 10.1108/17410380610662889.

[15] S. Finney, and M. Corbett, "ERP Implementation: A Compilation and Analysis of Critical Success Factors," *Business Process Management Journal,* vol. 13, no. 3, pp. 329–347, 2007, DOI: 10.1108/14637150710752272.

[16] F. F.-H. Nah, K. M. Zuckweiler, and J. L.-S. Lau, "ERP Implementation: Chief Information Officers' Perceptions of Critical Success Factors," *International Journal of Human-Computer Interaction*, vol. 16, no. 1, pp. 5–22, 2003, DOI: 10.1207/S15327590IJHC1601_2.

[17] C. Leyh, "Critical Success Factors for ERP System Implementation Projects: A Literature Review," in *Advances in Enterprise Information Systems II*, C. Møller, and S. Chaudhry, Eds. Leiden, The Netherlands: CRC Press/Balkema, 2012, pp. 45–56.

[18] J. Esteves-Sousa, and J. Pastor-Collado, "Towards the Unification of Critical Success Factors for ERP Implementations," in *Proceedings of the 10th Annual Business Information Technology (BIT) Conference*, Manchester, UK, 2000.

[19] J. vom Brocke, A. Simons, B. Niehaves, K. Riemer, R. Plattfaut, and A. Cleven, "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process," in *Proceedings of the 17th European Conference on Information Systems (ECIS 2009)*, Verona, Italy, 2009.

[20] J. Webster and R. T. Watson, "Analyzing the Past Preparing the Future: Writing a Literature Review," *MIS Quarterly*, vol. 20, no. 2, pp. xiii–xxiii, 2002.

[21] C. Leyh, "Which Factors Influence ERP Implementation Projects in Small and Medium-Sized Enterprises?," in *Proceedings of the 20th Americas Conference on Information Systems (AMCIS 2014)*, Savanah, Georgia, USA, 2014.

[22] F. F.-H. Nah, and S. Delgado, "Critical Success Factors for Enterprise Resource Planning Implementation and Upgrade," *Journal of Computer Information Systems*, vol. 46, no. 29, pp. 99–113, 2006

[23] J. Esteves-Sousa, *Definition and Analysis of Critical Success Factors for ERP Implementation Projects*. Barcelona, Spain, 2004.

[24] U. Remus, "Critical Success Factors for Implementing Enterprise Portals: A Comparison with ERP Implementations," *Business Process Management Journal*, vol. 13, no. 4, pp. 538–552, 2007, DOI: 10.1108/14637150710763568.

DOI: 10.15439/2015F251
ACSIS, Vol. 5

# Approach to Analysis and Assessment of ERP System. A Software Vendor's Perspective

Ilona Pawełoszek
Częstochowa Univeristy of
Technology, Management Faculty
Poland
Email: ipaweloszek@zim.pcz.pl

***Abstract*—The paper presents an approach to analysis and assessment of benefits brought by an implementation of an Enterprise Resource Planning system. The research has been conducted on a sample of 10 Polish companies using the Xpertis software, which is one of the popular applications for supporting the business activities of small and medium companies. The approach presented hereby aims at elaboration of the assessment method which can be easily applied by the software vendor, moreover it is acceptable by the customers because it does not disclose the confidential business information and still gives the results informative enough to be valuable as well for the software company as for its customers. Cluster analysis conducted through the Orange Data Mining tool was proposed as a technique of data analysis. The comparative study of the Xpertis ERP features and the customers' characteristics has been briefly presented and the directions of the future development of the considered software have been proposed.**

## I. Introduction

THE Assessment of benefits brought by an implementation of an ERP system is important from the point of view of two groups of stakeholders. The first group are companies which choose to use the system, incur costs of its implementation and maintenance. The other group's members are the system creators and vendors for which the system is a commodity for sale and a product covered by information and promotion campaigns. The aims of both groups of stakeholders – vendors and buyers – are somehow similar. In both cases there is a need for a forecast how the implementation of integrated information system will influence the functioning of the user-company in different areas. That information is crucial for a customer who buys the ERP system and wants to know how successful the investment will be and whether it will bring him operational effectiveness and competitive advantage. On the other hand the ERP vendors are strongly interested in effectiveness and usability of their system to use that knowledge in further development of their product and improvement of the level of their services (customer support, training, helpdesk applications etc.).

The domain of information systems analysis and assessment is a broad field related to many aspects of the company's functioning. The ERP assessment issues are well grounded in the literature, however most of the approaches represent the customer's perspective.

The critical success factors (CSFs) for an ERP implementation, with the number of over 80 [17]  have been well documented in the literature [13]. The factors such as: top management support, project team competence, interdepartmental cooperation, clear goals and objectives, project management, are considered to be monitored during the implementation of the system. [21]. But the question arises how to assess the system that has already been implemented and in the same time to check whether the successful implementation directly translates into the overall performance of the system during the consecutive years of its exploitation [11].

The main problem that confronts the current measurement frameworks is the fact that much of the benefits are strategic, therefore they are hard to quantify and may only appear several years after the implementation of the solution [4].

The aim of this paper is to present an approach that can be taken by software companies to create a method for evaluation of their products in terms of customer needs and identification of areas for the product's improvement.

The presented approach aims at bringing relevant and informative results that can also be presented to future customers to show them the usefulness of the system and reduce the information asymmetry which is cited as an important factor of ERP implementation failure [23]. The approach used in this study is based on the example of Xpertis ERP system offered by Macrologic Inc.

The structure of the paper is as follows. Section 2 introduces context for the study which is rooted in Polish market of ERP systems. The main focus has been on small and medium software companies and the issues they face in competing on highly demanding Polish market. Section 3 briefly describes the problems of ERP evaluation from the vendors perspective with special focus on software usability

978-83-60810-66-8/$25.00©2015, IEEE

and sustainability. Section 4 provides the description of research approach and design. Section 5 presents the data analysis approach based on clustering methods. Sections 6 and 7 present the research results, respectively the system features and the customers' characteristics. The conclusion section summarizes the outcome of the work, and suggests future research directions.

## II. The Polish market of ERP

ERP systems today grow in popularity and importance among enterprises of all sizes. For today's companies the support from well-tuned IT solution seems absolutely necessary although according to Central Statistical Office of Poland the share of Polish companies using ERP was only 18% in the year 2013 and grew to 22% in 2014, which is around 9% points less than average for 28 EU countries. [2]

Most of the small and medium Polish companies built their own systems (almost 40% of manufacturers with 100 to 1000 employees) or do not have the integrated IT system at all. The lack of integrated enterprise system hinders and delays information flow within an organization because the data is kept in many loosely coupled applications. Such a situation negatively impacts the company's competitiveness [14].

An analysis of the numerous reports from ERP market brings the reader to the conclusion that the market capacity is high. The potential target group are small and medium companies that do not have the ERP, but also large companies that need to replace or modernize their legacy system.

However the survey among managers of the Polish SME companies reveals little awareness of the need to implement the ERP and its impact on the company's profitability (only 33% of respondents declared to see the need to implement the ERP system) [24].

Many companies complain that after their huge investments in ERP systems, they find the systems do not bring the expected results in terms of new orders, new profits, or competitive advantage. [8]

For the majority of small and medium enterprises the decision of adoption as well as the selection of the appropriate ERP system is a difficult task. The main issues are shortage of financial resources, limited qualified IT personnel, lack of resources and time have been cited as the main factors that make this task difficult and risky [22]

The Polish ERP market is a very competitive one. There was 31 ERP vendors in Poland in 2014. The unquestionable market leader is SAP (39.6% share), which is followed by Comarch, Oracle, IFS and BPSC. The remaining market share of around 26% is divided among other small companies one of which is Macrologic Inc..

Large companies own enough financial and managerial resources to develop and bring to market new software products and to gain dominant share of the market. Contrarily, small software companies often meet difficulties in finances and staffing while running their businesses. Therefore they often choose to concentrate on a market niche, which is disregarded by large companies. Small companies also opt to build their competitive advantage mainly on the basis of their excellent responsibility and flexibility.

Actually, due to the limitation of resources faced by small and medium companies best practices which have proven in large firms might be too expensive or time consuming to perform. Accordingly, the recent researches start to find specific solutions to improve small companies' software processes in several aspects. [19]

## III. The Problems of ERP Evaluation From the Vendors' Perspective

The software systems today are more than just a commodity, they are strategic asset custom tailored to the issues each company is facing. They provide the ability to precisely adapt the software components to the needs of the organization. They are characterized by the requirement to provide maximum flexibility, understood primarily as a possible use by a variety of organizations [26].

Due to the complexity of enterprise systems, bringing a new product to the market is both expensive and risky for the vendor. Initial investment requires both highly skilled human effort, time and specialized IT tools. A software packages ERP life cycle has two linked segments: vendor's and customer's. The vendor's segment includes phases: analysis, design, developing and maintenance [27]. While planning and developing a new software product it is necessary to anticipate its complete lifecycle which is of even 10 to 20 years of exploitation. During its lifetime the ERP system needs to be securely operated, patched, upgraded, extended and integrated with other systems, and still preserve the functional requirements of the end-users.

From the marketing point of view, one of the most important factors is usability of the ERP system, which, according to ISO Standards can be viewed as a set of three factors: effectiveness, efficiency and satisfaction [10]. While users' satisfaction is quite easy to asses by asking the users for their opinion, there is more confusion when it comes to actually measure the system's efficiency or effectiveness especially in economic terms. It is hard to say to what degree the system impacts the overall performance of a company. Moreover the values to be measured are dynamic and change with the time that has elapsed since the completion of implementation.

Usability can be also described in narrower sense as a set of three following criteria [20]: navigation, presentation and learnability. Navigation is about accessing the information easy and finding system's functions by using the elements of the user interface. Presentation is the feature describing visual layout of the screen and the layout of printed documents. The layout of the interface elements such as menus, dialog boxes and controls should be logically structured and legible. Learnability is the most tricky to asses because it highly depends on other non-system issues such as users' skills, technical background, previous experiences with similar systems etc. Learnability can be measured as time needed to get basic knowledge to work with the system. Learnability is associated with accessible online or offline help files and additional training offered by the vendor. Learnability can be also measured as the degree of ease to learn how to use the system effectively, however this measure is very subjective. It often happens that the first implementation of the ERP system in a given company requires changes in business processes in such case the users can report difficulties in understanding the system as it does not correspond with the processes they are used to.

Security and reliability are also important issues that the vendor should be able to assess. This can be done by collecting data from system logs, especially those associated with error events. The events recorded by an ERP system should be semantically interpreted to find out the direct or indirect causes of failures.

In recent years, increasing environmental and social concerns posed a new challenge to the software vendors to improve sustainability of their products. Sustainability in broader perspective (in contrast with the narrow focus on reducing emissions of carbon dioxide) covers all aspects that potentially impact the use of any limited resource. Many resources in software development and operation can be considered as limited ones, for example: the number and expertise of administrators, the time of end users getting used to software or fulfilling a certain task, the available hardware etc., Sustainable software takes these constraints into account and balances them with the value added by the software [6].

The factors of sustainability are hard to predict and it is only during the exploitation phase that they can be determined. The evaluation of the system always requires cooperation of the stakeholders in determining where the highest probabilities of weak points exist. These observations are valuable as well for the vendor to gain knowledge about the products as for customers to help them improve their businesses.

## IV. Research Approach and Design

The aim of the study was to evaluate the impact of the ERP system on different areas of the company where the system is supposed to bring improvements in comparison to the situation before the system implementation. The studied product was the ERP system Xpertis developed by Macrologic Inc.. The system has been chosen as the object of the study due to the fact that it is exploited for educational purposes at Czestochowa University of Technology and so the author, working as an educator, had a chance to get familiar with its functions and business logic. The other important aspect is the pro-innovative attitude of the Macrologic's Executives who are substantially interested in continuous improvement of their product and competitiveness on the demanding Polish enterprise software market. The company is searching for new approaches to get valuable product and customer knowledge. These two kinds of knowledge are closely related to each other.

Product knowledge is an essential sales skill. The software manufacturer and vendor have their own subjective view of the product. To get the whole picture of the complex software product it is necessary to consider the customer's perspective. Moreover by acquiring the customers' knowledge the vendor gains deep understanding of the product features that allows the vendors to present the benefits accurately and persuasively. Customer knowledge is an essential asset since it represents a source of customer value improvement [5].

The market of ERP systems is overloaded with a lot of competitors offering very similar products. Therefore it can be hard to recognize the differences between them as well for the customer as for the vendor. In this situation the competitive advantage can be found in the way of communicating with the future and existing customers. The key element of communication seems to be the way of presenting them a value proposition. In marketing, a customer value proposition consists of the total sum of benefits which a vendor promises a customer will receive in return for the customer's associated payment [9].

Customers are a brilliant source of knowledge for the companies, because they gain expertise while using products or services [7]. However gaining customer knowledge is a difficult task. Survey practitioners often experience difficulties in collecting reliable data due to various privacy concerns. The exposed problem is mentioned in many publications. Some of the companies refuse to take part in the survey as their company policy do not allow sharing internal information with outsiders even for academic purposes [18]. Companies are afraid to conduct the research because it could reveal weaknesses in the management process, which are known but ignored by the executives of these companies [12]. Some of the companies' excuse for not taking part in the surveys is that the studies do not seem to bring any value for the respondents. To avoid such situation the goal of the study should be clearly defined and the utility value for the respondents should be the communicated at first.

The presented approach to analysis and assessment of ERP systems addresses the problem of commercially-sensitive information. This could be achieved through an adequate design of the survey questionnaire.

The approach, although it is tailored to fit the needs of the Macrologic Inc. and the Xpertis ERP, can be used by other software companies offering similar systems after some necessary modifications of the survey questionnaire according to the specificity of their product.

The survey has been conducted on the sample of 10 Polish companies, who have recently (during the past three years) completed the implementation of the Xpertis system. The participants for the survey were selected using a non-random decision rule intended to select the set of companies who satisfy the following conditions:

- The surveyed companies localized in different cities representing at least 50% of Polish voivodeships (administrative districts).
- The Company's management board's permission for the employees to undergo the survey.
- The Companies differentiated throughout the branches of industry (furniture, metallurgic, chemical, foundry, services and other).
- There were usually 7-10 persons who were surveyed in each of the companies. The persons are employees who are directly involved in operating the selected system's modules.

The survey questions were brainstormed by the team of three IT and management experts, all of them having experience in using the considered ERP.

The responders were employees operating the system in their daily work. They were asked to evaluate the impact of the system using the Likert-type scale from -2 to 2 (presented in Table I). The lack of evaluation were also allowed in cases when particular question was not relevant to the scope of the company's activities.

The reasons for the choice of the Likert scale instead of concrete values were twofold:

- the chosen scale is easy to interpret and compare because the values are normalized,
- the Likert-type scale was acceptable to the surveyed companies as a form of the survey that does not disclose their detailed financial and operating information.

TABLE I.
RATING SCALE AND ITS INTERPRETATION

| Very bad/ significant deterioration | bad/ deterioration | neutral/ no changes | good /improvement | Very good /significant improvement | Not applicable* |
|---|---|---|---|---|---|
| -2 | -1 | 0 | 1 | 2 | NA |

* The answer in case the given ERP module was not implemented or the company does not operate in the given area.

Source: Own elaboration

The survey questions covered 11 areas of the organization's activity:

1. Production - execution of orders
2. Sales
3. After-sales services
4. Recycling
5. Marketing
6. Procurement
7. Warehouse management
8. Economic analyses
9. Organization
10. Human resources and payroll,
11. Accounting and finance

The questions for all the aforementioned areas are organized into 3 categories and subsequent factors:

1. General factors:

- Compliance with the business requirements (Pre-implementation analysis)
- Compliance with business strategy
- functional adjustment to the business process specificity

2. User interface factors and usability factors:

- Overall satisfaction of the user
- Ease of use
- Functional adjustment to the user's tasks
- User friendliness of the interface
- User assistance in problem solving
- User's autonomy (necessity of the system's administrator assistance)
- Visual attractiveness of the printed documents

3. Specific factors – different for each area of company, associated with specific business operations.

The dataset yields a matrix containing 10 units of observation, each one is described by 205 factors..

The first aim of analysis was to extract the factors which are important and influence the functioning of the Xpertis ERP system, usability and effectiveness of the users' work and at the same time eliminating those factors which appear to be irrelevant (mainly assessed as 0 and ND)

The second aim was to find similarities between evaluations made by 10 surveyed companies. These similarities and differences can be useful to create the profiles of features and needs of different groups of customers. This research will allow to get knowledge about the needed future adjustments of the ERP system to the customers' needs.

## V. DATA ANALYSIS METHOD

The analysis could be performed on an intuitive way by a simple method of screening the data and selecting those factors which received the highest and the lowest scores in the customers' opinions. However establishing a more detailed approach to the classification of important features of ERP can be performed through the means of data-mining as it offers statistically based methods for identification of patterns in data sets.

For the considered dataset cluster analysis seems to be the appropriate choice because it allows to identify groups of similar records and moreover it is a well described and investigated method that is still being developed and ameliorated up till today [15].

Cluster Analysis or clustering is a common technique of statistical data analysis – specifically unsupervised machine learning. The objective of clustering is to assign observations to the groups (called clusters) in a way that the observations in the same cluster are similar to one another in some sense.

There are many clustering methods based on various algorithms that can be applied depending the nature of the dataset under consideration. Different clustering algorithms may render different results on the same data. Moreover the same clustering algorithm may bring different results on the same data, if it involves arbitrary initial parameters.

However interpreting the results of cluster analysis is not a trivial task, because it requires knowledge of semantic relations between investigated attributes of the ERP system and the domain it supports.

There are three important steps in the preparation of cluster analysis:

1. Selecting a distance measure – The similarity between various objects is defined by a distance measure. The distance measure plays an important role in obtaining semantically meaningful clusters. For simple datasets where the data is multidimensional, Euclidean distance measure can be employed. [25]. Moreover the Euclidean distance is appropriate for data measured on the same scale, as in this case.

TABLE II.
THE STATISTICS OF THE CLUSTERS

| Cluster | Number of features | Evaluation | | | | | |
|---|---|---|---|---|---|---|---|
| | | -2 | -1 | 0 | 1 | 2 | ND |
| cluster 0 | 87 | 0,0% | 0,0% | 11,1% | 31,7% | 55,1% | 2,1% |
| cluster 1 | 13 | 0,0% | 0,0% | 7,7% | 12,3% | 78,5% | 1,5% |
| cluster 2 | 36 | 0,3% | 0,3% | 32,1% | 37,6% | 18,7% | 2,0% |
| cluster 3 | 36 | 3,3% | 0,3% | 0,6% | 75,6% | 19,4% | 0,8% |
| cluster 4 | 11 | 2,7% | 4,5% | 50,9% | 34,5% | 0,9% | 6,4% |
| cluster 5 | 22 | 0,0% | 0,0% | 19,1% | 3,6% | 0,0% | 77,3% |
| sum | 205 | | | | | | |

* The answer in case the given ERP module was not implemented or the company does not operate in the given area.
Source: Own elaboration

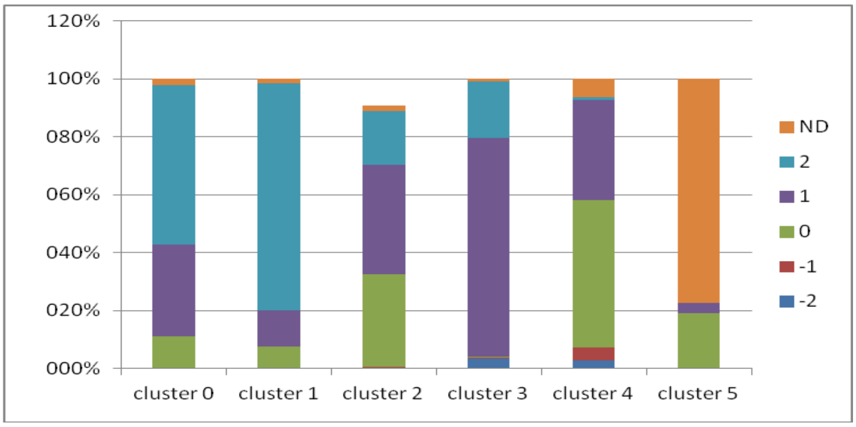


Fig. 1 Visualization of feature clusters statistics; Source: Own elaboration

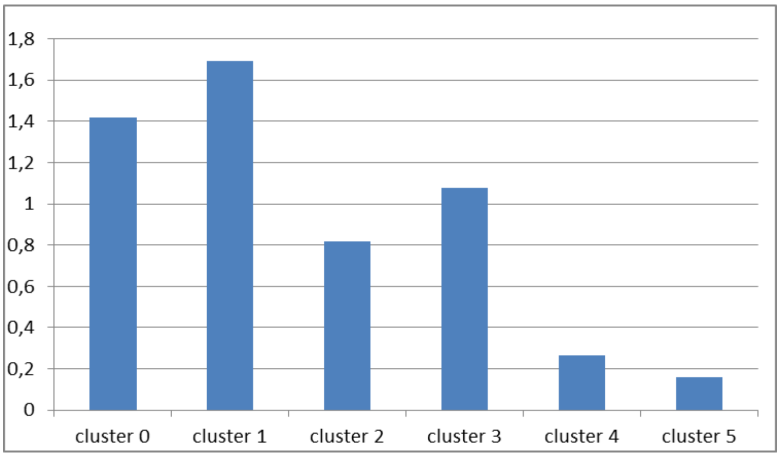


Fig. 2 Average evaluation for each cluster; Source: Own elaboration

2. Choosing clustering method and tool. A Hierarchical clustering method has been applied to classify the data set. The clustering algorithm were used and analyses were performed with Orange 2.7.1 software (developed by developed by Bioinformatics Lab at University of Ljubljana, Slovenia, in collaboration with open source community [3].

3. Determining the number of clusters in a data set. It is still an open problem in the machine-learning research community. There is no generally recognized state of the art statistical method to set the right number of clusters. The heuristics, rule of thumb can be recommended instead. The most important rule is to preserve an informative value of the clusters. Each of the clusters should have its own general characteristics differing from the others.

In the present case the cluster analysis can be applied to two purposes:

- to distinguish groups of the features which received similar ratings from the customers using the Xpertis ERP system. For this analysis data matrix should be prepared with the column names indicating the specific customers and the rows corresponding with the 205 studied features,
- to identify similarities between customers. In this case of analysis the matrix should be transposed (customers described by rows, features in columns).

## VI. Feature clustering results

In the case of feature clustering the dataset was divided on 6 clusters numbered from 0 to 5. Each of the clusters contains observation with similar values assigned by the users of the Xpertis ERP. Table II presents brief statistic characteristics of each cluster. As it can be easily seen from the Figure 1 (which is graphical representation of the data from the Table 2) the clusters differ markedly from each other. Also discrepancy in the average evaluation for each cluster can be seen.

Cluster 5 (Table III) contains attributes which are mostly irrelevant for the evaluation of the system or describe the areas that are not influenced by the system. The attributes in the cluster 5 describe the following areas:

- recycling (all the attributes from this group),
- human Resources and Payroll, – 3 attributes,
- accounting and finance (2 attributes).

The aforementioned attributes from the cluster 5 indicate that the companies do not involve in the recycling activities and do not see the need to support this area by ICT tools.

The other area in this cluster is human resources management. This one is probably seen as a diverse collection of “soft” skills that can be hardly supported by IT tools. The system could be extended by the functions of replacement workforce management in the area of registering the profiles of employees and job seekers. Additionally the decision support tools could be implemented to the recruitment process to make it more effective at getting the best person for a job or a project.

TABLE III.
THE ATTRIBUTES IN CLUSTER 5 – MOSTLY IRRELEVANT

| Area | Attributes |
|---|---|
| Recycling | Compliance with business requirements<br>Compliance with strategy<br>Matching to business processes and functions<br>Overall user satisfaction<br>Ease of use<br>Matching functional requirements (does the system support all the user’s tasks)<br>Interface - user-friendliness<br>User support in problem solving<br>User’s autonomy (the need for intervention of the system administrator)<br>Visual appeal of the printed documents<br>Evidence of post-consumer waste<br>Evidence of revenue resulting from the post-consumer waste.<br>Cooperation with recycling companies<br>Recycling rates<br>Recycling costs<br>Cost and time planning<br>Compliance with legal regulations |
| Human resources and payroll | Human resources management |
| | Creation of replacement workforce |
| | Employment of replacement workers in projects |
| Accounting and finance | Budgeting |
| | Daily expenditure plans |

Source: Own elaboration

Cluster 4 represents the areas on which the ERP system does not have the clear impact or the impact is slightly positive. The system attributes in this cluster are presented in the table IV.

TABLE IV.
THE ATTRIBUTES IN CLUSTER 4 – NO CLEAR IMPACT OR SLIGHTLY POSITIVE EVALUATION

| Area | Attributes |
|---|---|
| Production | User’s autonomy (the need for intervention of the system administrator) |
| | Integration of technical and administration areas (toolroom) |
| Sales | Easiness and speed of planning (ie. visits to customers) |
| | Speed of handling the users' complaints |
| After-sales services: | User’s autonomy |
| Marketing | Increased number of non-electronic customer orders |
| Procurement | Reduction of prices of supplies (by increased competition between suppliers)<br>Identification of delays and their causes |
| Warehouse management: | Loss of value of commodities under storage<br>use of storage space |
| Organization | Creation of the e-communities of employees |

Source: Own elaboration

Analyzing the results of the evaluation we can draft some recommendations and suggestions for changes and further development of the system. In the areas of production and after-sales services there is a need for more user autonomy. However the need for the administrator's intervention was mostly indicated by two of the surveyed companies.

Creation of e-communities seems to be interesting future development direction of the ERP system. E-communities may focus on different aspects of the company's functioning and also support the knowledge flow and retention. There is a movement taking place in the IT industry that is really driven by major consumer technology vendors like Apple, Google, and Facebook. In fact, employees and managers of most any business are already communicating with each other through various Web 2.0 technologies. However all of this communication is taking place outside the bounds of formal and secure IT systems [1]. Adding collaborative Web 2.0 technologies to ERP applications could be a way to address these challenges both for business users and IT.

TABLE V.

THE ATTRIBUTES IN CLUSTER 3 – MODERATELY GOOD AND VERY GOOD EVALUATION

| **Area** | **Attributes** |
|---|---|
| All the areas apart from recycling | User-friendliness of the interface |
| | Visual appeal of the printed documents |
| Production - execution of orders | Work planning |
| Sales | The User's autonomy |
| | Document automation |
| Marketing | User's autonomy |
| | Increasing number of contractors |
| | Increasing the number of customers |
| | Quick response for the customers' needs (new products offerings) |
| Procurement | The User's autonomy |
| | Increasing number of contractors (wider choice of raw materials and manufacturing tools) |
| | Differentiation of suppliers |
| Warehouse management | The User's autonomy |
| Economic analyses | The User's autonomy |
| Organization | The User's autonomy |
| | Corporate dictionaries |
| Human Resources and payroll | The User's autonomy |
| Accounting and finance | The User's autonomy |

Source: Own elaboration

The cluster 3 is characterized by moderately good (75,6%) and very good (19,4%).evaluation of the Xpertis ERP features. There are 36 elements in the cluster 3 - mainly the features associated with the user's interface and visual aspects (Table V).

Although the overall assessment is positive there are also signals pointing to the problems regarding the users' autonomy in different modules of the system. This may indicate the poor practical skills and theoretical background of some of the users. The lowest ratings (-2) were given by

TABLE VI .

THE ATTRIBUTES IN CLUSTER 2 – GOOD AND VERY GOOD EVALUATION

| **Area** | **Attributes** |
|---|---|
| All the areas apart from recycling | Compliance with the business strategy |
| Production - execution of orders | Lower production costs |
| | Decreased number of production downtimes |
| Sales | Acceleration of the sales process |
| | Automation of notifications |
| | Number of orders |
| | Number of completed orders |
| | Acceleration of orders completion |
| | The time of answering for the customer's questions |
| After-sales services | Compliance with business requirements (pre implementation analysis) |
| | Functional adjustment to the processes |
| | Ease of use |
| | Information on the customer's complaints |
| | Identification of problems |
| | Decreased cost of after-sales services |
| | Loyalty systems and analysis of their effectiveness |
| | Revenues from after-sales service |
| Warehouse management | Virtual warehouses |
| | Decrease in stocks |
| | Costs savings on warehousing |
| | Decrease in the storage period |
| | Reduced time of warehousing operations |
| | Reduced costs of non-moving articles |
| | Using data collectors |
| Organization | Project management support |
| Accounting and finance | Information on effectiveness of resources |
| | Increase in the financial revenue and decrease of loan service costs |

Source: Own elaboration

the customers who have been using the system for the shortest time (3 and 6 months).

This may indicate the need for additional training during the implementation and post-implementation phase.

The Cluster 2 contains attributes associated with all the areas of the companies' activities (apart from recycling). The relevant areas and attributes are presented in the Table VI.

As it can be seen from the survey results the users evaluation is good or very good for most of the attributes in this cluster. However there are also 5 attributes that received mainly 0, these are the following: "decreased cost of after sales services", "loyalty systems and analysis of their effectiveness", "revenues from after-sales service", "project management support", "increase in the financial revenue and decrease of loan service costs". The 0 marks mean that the system does not influence the aforementioned areas. However interpretation of the users' assessment indicates interesting possible ways of enhancing the functionality of the system modules. The CRM module could be extended with the possibility to create targeted loyalty programs with relevant and personalized rewards for the customers. Also project management is not directly supported by any module of the system. The project management support should be offered as an optional module integrated with Human Resources and financial modules. Although for now not all the customers practice project management, this discipline is becoming more widely recognized and used throughout the world [16]. Therefore it can be expected that the enterprise project management support could be one of the important features distinguishing the ERP product.

Cluster 1 (Table VII) contains features that are considered very positively influencing the business (mainly marks 2 and 1). This cluster mainly describes the areas associated with enterprise-wide communication and information management. The Users reported significant improvement in marketing effectiveness and efficiency relative to the performance before the implementation of Xpertis system.

The considered ERP system allows a company to manage its business with potential benefits of improved process flow, reduced inventories, more comprehensive data analysis and better customer service.

The Xpertis system has significant impact on acceleration of information flow, document circulation, reporting and analyses. The four companies of the 10 surveyed, claimed that the system does not influence the number of offers sent to their customers and the number of received responses. This fact can be caused by the business specificity and long-term contracts with their customers.

Cluster 0 is the largest one with 87 factors of the system evaluation. The data in this cluster reflect the very positive or moderately positive impact of the system on the surveyed areas of the business. The cluster contains data associated with 10 modules (apart from the recycling module). The most often appearing features are:

- ease of use,
- overall satisfaction or the users,
- support in problem solving,
- automation and acceleration of the specific operations.

Regarding the large number of factors in this cluster and the limited size of the hereby publication the factors are not listed.

TABLE VII.

THE ATTRIBUTES IN CLUSTER 1 – VERY POSITIVE EVALUATION

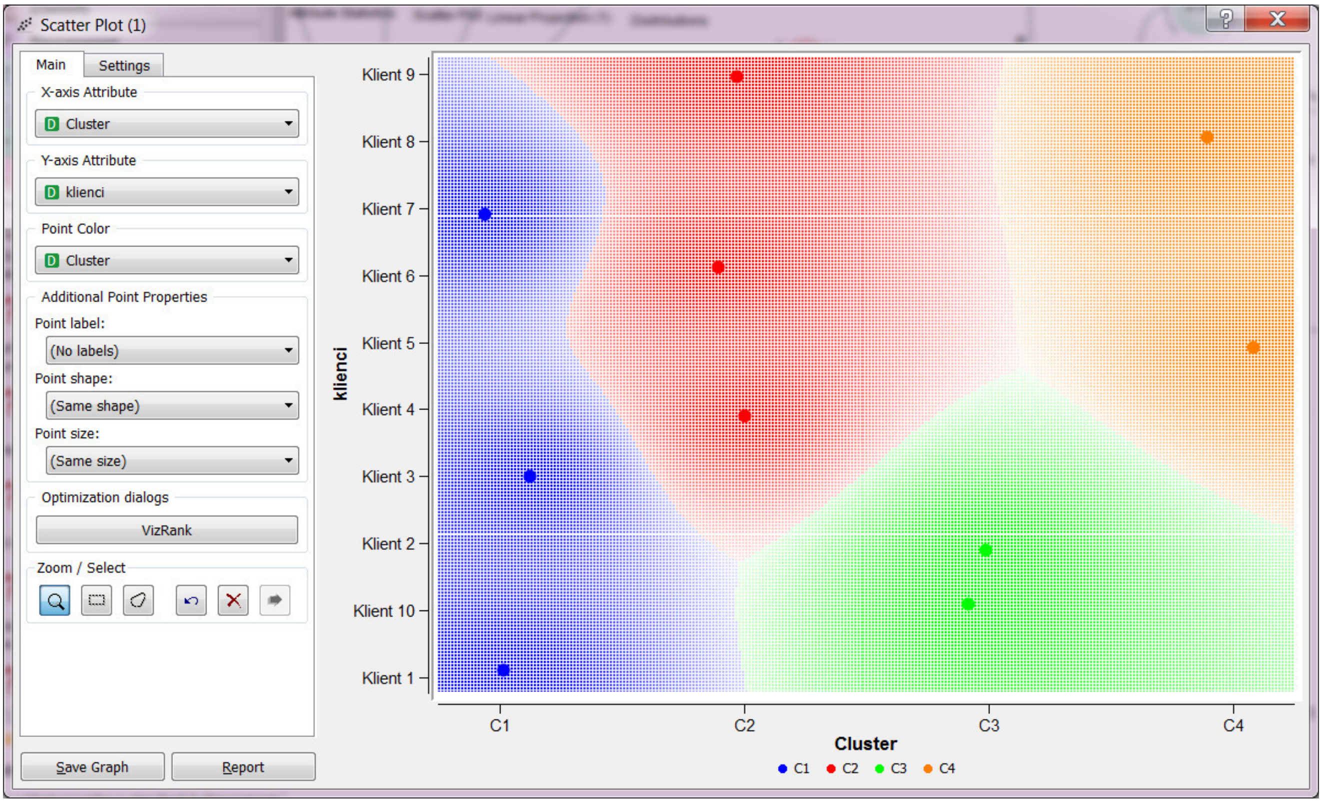
| Area | Attributes |
|---|---|
| Marketing | New medium of communication |
| | Number of offerings sent |
| | Number of responses received |
| | Number of electronic orders |
| Procurement | Acceleration of order completion |
| Warehouse management | Acceleration of stocktaking |
| | Raw materials assigned to specific orders/processes |
| | Materials management according to orders |
| Economic analyses | Greater scope of reporting |
| | Analysis of effectiveness of accessible fixed assets |
| Organization | Additional communication medium |
| | Communication regardless of location of the employees |
| Accounting and finance | Monthly balance sheet and reporting online |

Source: Own elaboration

The above analysis of the effects of the Xpertis ERP system on the business in 10 surveyed companies, shows that the considered software in most of the cases contributes to an improvement of the users' work in various aspects (communication, reporting, acceleration of operational tasks).

So far the potential areas of weakness seem to be factors from the cluster 5, and 4. Particularly the need for administrator's intervention reported by two of the 10 surveyed companies. To facilitate such analyses in the future some of the factors should be reconsidered in the questionnaire:

- human resources management (probably the problem was framed too broadly),
- creation of replacement workforce and employing the replacement workers in the projects (probably the surveyed companies do not practice project management)



Scatter Plot (1)

Main | Settings

X-axis Attribute: Cluster

Y-axis Attribute: klienci

Point Color: Cluster

Additional Point Properties

Point label: (No labels)

Point shape: (Same shape)

Point size: (Same size)

Fig. 3 Customer clusters visualization in Orange Data Mining Software

The question about the system's compliance with business strategy is recognized by the surveyed employees as correlated with the support for operational tasks.

## VII. The customers' characteristics

In practice of marketing management segmenting is the popular technique of putting customers into groups based on similarities. Segmenting technique assumes the groups are pre-defined. Here we propose clustering method to find similarities in customers so that they can be grouped, and therefore segmented. Then the groups generated by the clustering algorithm undergo semantic interpretation. Semantic interpretation requires some additional insight to find out the reasons of the similarities between the cluster member.

This part of the research approach is aimed at identification and analyses of regularities and similarities between companies exploiting the ERP system.

Although the sample was only 10 companies, the large number of factors taken into consideration makes the problem appropriate for the use of statistical methods. This time the clustering based on k-means algorithm was the choice. The dataset is partitioned into 4 clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.

The aforementioned Orange Data Mining tool was used to perform clustering. The Figure 3 presents the visualization of the clusters. The index on the vertical axis refers to the particular customer and the horizontal axis represents the cluster number.

By comparing the data on the time that has elapsed since the completion of the implementation with the clusters created by k-means algorithm it is easy to see some regularities in the evaluations made by customers assigned to the same cluster. It should be noted that the clustering was performed only on the basis of the customers evaluations of the system. Such an approach was taken intentionally to check how far the evaluation depends on the customers' characteristics. The table VIII presents the features of the surveyed Macrologic's customers along with the division on clusters.

A it was mentioned before the customer 5 and 8 reported problems with the users' autonomy, that were caused by little experience because the time after the completion of the system's implementation was relatively short (3 and 6 months). It can be reasonably expected that the evaluation will get better with time. The both observations were located in the cluster 4.

Regarding the data in Table VIII it can be assumed that the period of 1 year is the threshold time after which the users get enough experience to easily operate the system. Also the one year of exploitation gives the comprehensible view of achievements of targets assumed in pre-implementation phase.

The survey results also show the low evaluation cases in spite of the long post-implementation period (customer 2). In case of this customer the interview and observation showed the classical example of information asymmetry between the buyer and the vendor of ERP system.

In each of the considered cases the process of implementation had been preceded by pre-implementation analyses and the Xperits system had been customized to fit the customer's needs and business specificity. The information asymmetry could be eliminated or at least mitigated by devoting more time to discussions, consulting and collaboration between the vendor and the buyer. On the other hand it would significantly extend the time and increase the costs of the ERP implementation. The time-cost tradeoff problem is an important issue in the scheduling the large information system implementations.

By considering available data regarding the specificity of the surveyed companies along with the evaluations of the ERP system it can be found that there is no clear dependence between branch of industry and the customer perception of the system. For example the customer 1 and 8 both represent the metallurgic industry, however their production is different. Their evaluations of the system are very divergent so they were assigned to the different clusters and represented by the two points on figure 2 that are far apart from each other.

The closer look at the company 1 shows that the customer is very demanding in terms of the IT system and at the same time has very carefully elaborated procedures and is highly aware of the IT capabilities and own expectations. The implementation of the Xpertis system for this customer was difficult and many issues occurred that made it necessary to modify the system or change some process as well as the employees' habits. However the cumbersome and costly implementation resulted in achieving of the expected results and now after the 36 months of exploitation the system is considered as very valuable for the company.

Another example that proves the hypothesis of interdependence of time factor and the user's evaluation of the system is cluster 2 which contains two customers (9 and 4). Both customers evaluations were similar and the time after the completion of implementation is 18 months in both cases. The clear correlation between the number of IT staff, the form of employment and the customers' evaluations could not be identified.

TABLE VIII.

THE ATTRIBUTES OF THE CUSTOMERS

| Customer | Time since ERP implementation (months) | Cluster | Number of IT staff | Branch |
|---|---|---|---|---|
| **Customer 1** | 36 | 1 | 1 | Metallurgy |
| **Customer 2** | 36 | 3 | 1 | Furniture |
| **Customer 3** | 12 | 1 | Casual employee | Metallurgy |
| **Customer 4** | 18 | 2 | Additionally acting | Chemical |
| **Customer 5** | 3 | 4 | 1 | Services, projects |
| **Customer 6** | 30 | 2 | 9 | Services |
| **Customer 7** | 20 | 1 | 1 | Plastics and paper processing |
| **Customer 8** | 6 | 4 | 1 | Metallurgy |
| **Customer 9** | 18 | 2 | 2 | Metallurgy, molding |
| **Customer 10** | 12 | 3 | 1 | Production, plastics |

Source: Own elaboration

## VIII. CONCLUSION

The problem of post-implementation analysis of the ERP systems deserves further exploration. The study presented in the hereby paper is an attempt to elaborate the easy to use methodology for the software vendors to get more insight about their customers' needs and the possibilities of improvement of their products. The main issues in the ERP implementations seem to be precise description of the customer expectations and confronting it with the system capabilities. The approach presented hereby let the vendor create an insight on the overall system features in a way that does not reveal private information of the investigated companies. Therefore the research results can be presented to potential customers to better illustrate the areas of the impact of the ERP system on business processes.

The cluster analysis is an easy to use and flexible method to discover structures in a data set however it does not provide any explanation itself. The semantic interpretation of clustering effects can be performed by combining the data of each cluster with the knowledge of experienced members of the implementation team. The exploited tool – Orange Data Mining is free and easy to use because it is based on visual data modeling.

The analysis presented in this paper could be extended further to identify more similarities between the customers and the specific modules of the ERP system. For example clustering can be performed to analyze the features of each separate module. Such analysis only requires to prepare the data set limited to the selected features. The new dataset can be then easily loaded to the model built in Orange Data Mining tool.

The approach presented in this paper could be an inspiration for small and medium software companies to elaborate their own methods for products evaluation and comparison of their customers. The sophisticated research approaches based on data mining are often considered as reserved for large businesses with expansive budgets and creative departments. However, with open source data visualization and analysis tools available it is easier than ever for small and medium companies to gain knowledge on their products and customers to let them more effectively compete with larger companies on software market.

## ACKNOWLEDGMENT

The survey prepared is a collaborative effort. Therefore the author thanks Cezary Stępniak for the idea of the research and cooperation on preparation of the survey questionnaire. The author would like to thank Janusz Jakóbczak consultant to the Management Board of Macrologic Inc. for consulting the survey questionnaire, gathering data and his comments to the draft of the paper.

## REFERENCES

[1] Anderson D., Enterprise 2.0 and social media coming to ERP, InfoWorld Dec 22, 2010, retrieved may 02, .2015 from: http://www.infoworld.com/article/2624804/erp/enterprise-2-0-and-social-media-coming-to-erp.html

[2] Berezowska J., Huet M., Kamińska M., Kwiatkowska M., Orczykowska M., Rozkrut D, Wegner M., Społeczeństwo informacyjne w Polsce Wyniki badań statystycznych. Główny Urząd Statystyczny Urząd Statystyczny w Szczecinie 2014

[3] Demšar, T. Curk J.T., and Erjavec, A. Orange: Data Mining Toolbox in Python; Journal of Machine Learning Research 14(Aug):2349−2353, 2013

[4] Dyczkowski M., Korczak J., Dudycz H., Multi-criteria evaluation of BI systems. The case study of InKoM dashboard, [in] Prace Naukowe Informatyka Ekonomiczna, Business Informatics, 2014 (in print).

[5] Frauendorf, J. Customer processes in business-to-business service transactions. Wiesbaden: Deutscher Universi̋tts-Verlag 2006.

[6] Gudbrod, R.. Wiele, C.: The Software Dilemma: Balancing Creativity and Control on the Path to Sustainable Software. Springer, Berlin 2012

[7] Hong Tang, S., Homayouni, M. and Alaei, H., The role of intelligent agents in customer knowledge management. African Journal of Business Management Vol. 5(16), 18 August, 2011, 7042-7049

[8] Hsu, P., Commodity or Competitive Advantage? Analysis of the ERP Value Paradox," Electronic Commerce and Research and Applications, Vol. 12, No. 6, 2013, pp. 412-424

[9] Hutt M.D., e-Study Guide for: Business Marketing Management: B2B 10th Edition Study Guide, Cram101 Textbooks Reviews 2012

[10] ISO. 1997. ISO9241-11: Ergonomic requirements for office work with visual display terminals (VDT's). Part 11-guidelines for specifying and measuring usability. International Standards Organization (ISO). Available at:http retrieved may 02, .2015 from: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=16883.

[11] Jenatabadi, H.S. and A. Noudoostbeni,, End-User Satisfaction in ERP System: Application of Logit Modeling. Applied Mathematical Sciences, 2014. 8 (24): p. 1187-1192.

[12] Kot, E. M. "How to Conduct the Audit of Intellectual Capital in Polish Tourism Business?" Electronic Journal of Knowledge Management Volume 7 Issue 4,(pp459 - 468), Retrieved May 2015 at: www.ejkm.com/issue/download.html?idArticle=197

[13] Leyh, C., Critical Success Factors for ERP Projects in Small and Medium-sized Enterprises – The Perspective of Selected German SMEs. In: Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, FedCSIS 2014, September 6 - 10, Warsaw, Poland, pp. 1181-1190. https://fedcsis.org/proceedings/2014/pliks/243.pdf

[14] Mejssner B., Mniejsze przedsiębiorstwa w kolejce po ERP, Computerworld 16 April 2014 http://www.computerworld.pl/artykuly/395608/Mniejsze.przedsiebiorstwa.w.kolejce.po.ERP.html

[15] Nekvasil M., Evaluation of Semantic Applications for Enterprises, Prague, 2010 Retrieved from: http://nekvasil.eu/files/papers/1012%20-%20dissertation%20thesis%20-%20Evaluation%20of%20Semantic%20Applications%20for%20Enterprises.pdf

[16] Newell, M. W., and Grashina, M. N. (2004). The project management question and answer book. New York, NY: AMACOM

[17] Ngai EWT, Law CCH, Wat FKT. Examining the critical success factors in the adoption of enterprise resource planning. Computers in Industry; 2008, 59:548–564.

[18] Senathiraja R., Fernando M.D., An Emperical Study on the Impact of Multiple Intelligences on Team Development in The IT Industry in Sri Lanka, South East Asia Journal of Contemporary Business, Economics and Law, Vol. 2, Issue 1 (June) 2013 p.47-58

[19] Shen Y., Software Engineering Challenges in Small Companies Helsinki 2008 http://www.cs.helsinki.fi/u/paakki/Shen.pdf

[20] Scholtz B, Calitz A., Cilliers C., "Usability Evaluation of a Mediumsized ERP System in Higher Education" The Electronic Journal Information Systems Evaluation Volume 16 Issue 2 2013, (148-161) , Retrieved 01.07.2015 at www.ejise.com

[21] Somers, T.M., and Nelson, K., The Impact of Critical Success Factors across the Stages of Enterprise Resource Planning Implementations'. Proceedings of the 34,h Hawaii International Conference on System Sciences (HICSS-3) Maui, Hawaii. (CD-ROM), January 3-6, 2001.

[22] Stefanou, C. J. 2014, Adoption of Free/Open Source ERP Software by SMEs, in: Information Systems for Small and Medium-sized Enterprises, Devos, J., van Landeghem, H., Deschoolmeester,D. (Eds.) (pp. 157-166): Springer Publishing Company, Incorporated, Berlin, Germany 2014.

[23] Tazyeen F., Modeling government ERP acquisition methods using system dynamics, Massachusetts Institute of Technology 2012, Retrieved may 2015 at: http://web.mit.edu/smadnick/www/wp/2012-04.pdf

[24] Tchorek-Helm T., Dlaczego firmy nie kupują ERP, 16 June 2011, Retrieved, May 2015 at: http://www.komputerwfirmie.pl/informacje/raporty/pelny/4636/dlaczego-firmy-nie-kupuja-erp

[25] Vimal, A., Valluri. S.R., Karlapalem. K., An experiment with distance measures clustering. Technical report. Center of Data Engineering. II1T. Hyderabad 2008, retrieved. May 2015 at: http://www.cse.iitb.ac.in/~comad/2008/PDFs/61-ankita.pdf

[26] Wieczorkowski J., Pawełoszek I., Polak P., Software Standardization in the Context of the Innovativeness of Enterprise Operations, in: P. Kommers, P. Isaias (red.), Proceedings of the International Conference e-Society 2015, s. 215-222, IADIS Press, Lisbon, 2015.

[27] Wieczorkowski J., Polak P., "Analysis and Implementation Phases in the Two-Segmental Model of Information Systems Lifecycle", in 2012 Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS, PTI, IEEE, Wrocław 2012, pp. 1041-1046

DOI: 10.15439/2015F44
ACSIS, Vol. 5

# Efficiency of formal verification of ArchiMate business processes with NuSMV model checker

Piotr Szwed
AGH University of Science and Technology
E-mail: pszwed@agh.edu.pl

***Abstract*—We investigate an application of model checking techniques to automated verification of business processes expressed in ArchiMate language. As a verification tool the state of the art symbolic model checker NuSMV is used. The proposed approach consists in fully automated translation of behavioral elements embedded in ArchiMate models into a corresponding representation in NuSMV language and then verifying its properties specified in CTL. Since our goal is to build an interactive verification tool, we focus on time efficiency of the verification process. We report results of tests performed on artificial process models of various complexity, as well as on a real business process example. The results show, that the described approach can be applied successfully, however, verification of complex business process specifications may face the problem of state space explosion. In such a case, to make the verification feasible, various reductions and simplifications can be applied.**

***Index Terms*—ArchiMate, business process verification, model checking, NuSMV**

## I. Introduction

THE GOAL of business process verification is to check if processes intended to be used or already implemented within an organization exhibit desired behavioral properties. The analyzed models may represent combinations of manual tasks performed by employees with operations supported by IT tools, as well as fully automated specifications run by process execution engines. In both cases process verification is appreciated by business organizations, as it can detect potential errors, design flaws and ambiguities.

Although graphical process modeling tools offer support for local syntax checking, e.g. correct use of links between elements of the diagram, some structural errors remain undetected, especially those resulting from incorrect use of synchronization mechanisms [1]. Partial analysis of model behavior can be performed by simulation techniques, however, only application of formal methods can give unequivocal answer that the verified system meets formally specified requirements.

Obviously, the landscape of business process modeling notations is dominated by such languages, as BPMN [2] or EPC [3], [4]. In this work, however, we decided to focus on verification of process models defined in ArchiMate, a contemporary, open and independent language intended for description of enterprise architectures [5]. Although ArchiMate comprises a variety of elements intended to model important aspects of enterprise architectures, constructs allowing to model behavior can be found only in the *Business* layer. They include events, processes (also understood as activities), interactions, collaborations and several types of junctions.

Formal system verification can by done either by deductive reasoning or model checking [6]. Deductive reasoning consists in formulating theorems specifying desired system properties and proving or falsifying them using manual or automated techniques. However, deductive reasoning methods gives very little information on causes, if the verified property does not hold.

Model checking allows to verify a concurrent system modeled as a finite state transition graph against a set of specifications expressed in a propositional temporal logic. It employs efficient internal representations and quick search procedures to determine automatically, whether the specifications are satisfied along the computational paths. Moreover, if a specification is not met, the verification procedure delivers a counterexample that can be used to analyze the source of the error. The main problem faced by model checking is state space explosion [7]. At the very beginning only small examples could have been processed. A significant progress in this technique was achieved with application of ordered binary decision diagrams (OBBD) [8] allowing to model systems consisting of millions of states and transitions.

Although formal tools reached the state of the art, they are not commonly used in the engineering practice. According to Huuck [9] three factors decide on successful application of formal tools: they should be simple to use, the time spent on model preparation and verification should be comparable with other user activities, and, finally, a tool should provide a real value, i.e. deliver information that was previously not available.

We were motivated by an idea of developing a software tool that fully automatically translates behavioral elements of a business model expressed in ArchiMate language to a corresponding finite-state graph required by the model checker. Then, after running the verification and detecting errors, valuable information about specifications not met and counterexamples can be returned to the process designer.

The concept of verification system is presented in Fig. 1. The business model is defined within Archi [10], a well known ArchiMate modeling tool.

As a verification platform the state of the art symbolic model checker NuSMV [11] is used. NuSMV allows to enter a model comprising a number of communicating finite state machines (FSM) and automatically checks its properties specified as Computational Tree Logic (CTL) or Linear Temporal Logic (LTL) formulas. We have developed an Archi plugin that

978-83-60810-66-8/$25.00©2015, IEEE

extracts a subgraph of ArchiMate behavioral elements and transforms it into NuSMV model descriptions. Specifications of desired properties are defined by a process designer, however, a part of them is generated automatically by an analysis of the process structure.

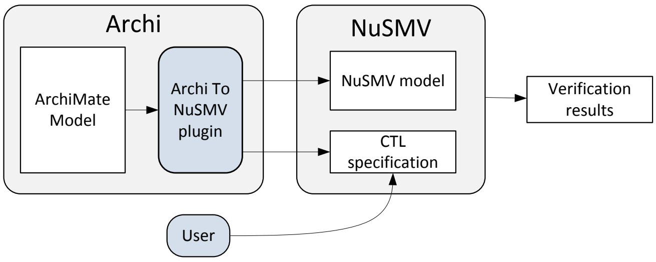


Fig. 1. The concept of the verification system

This paper is a continuation of our previous work [12], where an initial concept of a model checking verification system dedicated to ArchiMate models was described. Since our goal is to build an interactive verification tool, the main concern in this work is the *time efficiency* of the verification process. In order to assess it, we performed tests a for set of artificial process models of various complexity, as well as we verified a real business process example. In all cases we collected information related to the size of state space and processing time.

The paper is organized as follows: next Section II discusses various approaches to the verification of business models. Section III presents basic concepts of ArchiMate language. It is followed by Section IV, which describes details of the NuSMV model generation procedure. Section V reports results of time efficiency tests. Section VI provides concluding remarks.

## II. Related works

Application of formal methods to verification of business processes was surveyed by Morimoto [1]. Author distinguished three prevalent approaches: based on automata, Petri nets and process algebras. The first approach consists in translating the process description into a set of communicating automata (state machines) and performing model checking with such tools, as SPIN [13] or UPPAAL [14]. In analysis of Petri net models, basically simulation techniques are used, especially in the case of more expressive colored Petri nets.

Model checking has an established position in verification of business processes. It was applied in [15] to BPMN models extended with temporal and resource constraints. In [16] verification of of e-business processes was achieved by translation to CSP language and checking refinement between two specifications. In [17] authors implemented a system that translated BPEL specification into NuSMV language, what allowed them to check properties defined as CTL formulas. Three types of correctness properties were analyzed: invariants, properties of final states and temporal relations between activities. The first two can be classified as *safeness*, the last as the *liveness* property. Similarly, in work by Fu et al. [18] CTL was applied to the verification of e-services and workflows with both bounded and unbounded numbers of process instances. The work [19] discusses verification of data-centric business processes. The correctness problem was expressed in the LTL-FO, an extension to the Linear Temporal Logic, in which propositions were replaced by First Order statements about data objects.

Wynn et al. [20] discuss verification tools developed for models in YAWL language [21], with a special focus on OR-joins and cancellation constructs. The verified process properties are predefined. e.g. a soundness property is a combination of three conditions: a process should always complete, after the completion all its subprocesses should be inactive and every its part should be executable. Verification algorithms for YAWL are closely related to those of Petri nets, i.e. analyzing reachability graphs and in hard cases applying Petri nets reduction techniques.

In our previous works [22], [23] we proposed a method for verification of ArchiMate behavioral specifications based on deductive reasoning. The described approach consisted in transforming ArchiMate model into a set of LTL formulas, then extending it with formulas defining desired system properties and formally proving them using semantic tableaux method.

Although verification of business processes have been investigated for at least 15 years, surprisingly, there is very little information provided about efficiency of applied techniques. This in particular concerns the formal verification with the model checking approach. However, quantified results for a quite complex process specified in YAWL are given in the work [20] entitled remarkably “Business Process Verification - Finally a Reality!”. Without reductions the soundness verification of the process stages took about a few dozen seconds; applying reductions decreased the verification time of the whole process below ten seconds.

NuSMV [11] is a state of the art model checker that has been succesfully used for various verification tasks including formal protocol analysis [24], verification of requirements specification [25] or planning tasks [26]. The package uses a special language (named also NuSMV) to define the verified model as a set of linked finite state machines, as well as its specification in form of temporal logic formulas. The model submitted to the verification tool must be manually coded in NuSMV language or generated from another language amenable to state transition system, e.g a state charts [27] or reachability graphs of Petri nets [28].

## III. ArchiMate

ArchiMate [5] is a contemporary, open and independent language intended for description of enterprise architectures. The definition of ArchiMate has been accompanied by an assumption, that in order to build an expressive business model, it is necessary to use the relationships between completely different areas, starting from business motivation to business processes, services and infrastructure.

The language comprises five main modeling layers shortly characterized below.

- *Business* layer includes business processes and objects, functions, events, roles and services.
- *Application* layer contains components, interfaces, application services and data objects.
- *Technology* layer gathers such elements as artifacts, nodes, software, devices, communication channels and networks.
- *Motivation* layer allow to express business drivers, goals, requirements and principles.
- *Implementation&Migration* layer contains such elements, as work package, deliverable and gap.

ArchiMate allows to present the architecture in the form of views, which, depending on the needs, can include only items from one layer or can show vertical relations between elements belonging to different layers, e.g.: a relationship between a business process and a function of the component software.

ArchiMate provides a small set of constructs that can be used to model behavior. It includes *Business Processes*, *Functions*, *Interactions*, *Events* and various connectors (*Junctions*), which can be attributed with a logical operator specifying, how inputs should be combined or output produced. According to language specification casual or temporal relationships between behavioral elements are expressed with use of *triggering* relation. On the other hand, ArchiMate models frequently use *composition* and *aggregation* relations, e.g. to show that a process is built from smaller behavioral elements (subprocesses or functions).

In should be also noted that *Business Activity* present in ArchiMate 1.0 specification was removed in version 2.0. Instead, an atomic process should be used.

Although the set of behavioral elements seems to be very limited when compared with BPMN [2], after adopting a certain modeling convention its expressiveness can be similar [29]. An advantage of the language is that in allows to comprise in a single model a broad context of business processes including roles, services, processed business objects and elements of lower layers responsible for implementation and deployment.

Another process modeling notation that can be almost directly mapped on ArchiMate constructs is *Event-driven Process Chain* (EPC) [3], [4]. All behavioral elements of both languages are exactly the same: events, functions (or processes in ArchiMate) and various joins and splits (XOR, OR and AND).

## IV. Model generation

This section discusses language patterns that can be used to model ArchiMate elements in NuSMV, as well as details of the translation procedure.

### A. ArchiMate model

The internal structure of an ArchiMate model constitutes a graph of nodes linked by directed edges. Both nodes and edges are attributed with information indicating the type of element or relation. While generating NuSMV code describing behavioral aspects of ArchiMate model, we focus on components of the *Business layer*: processes (interactions, functions), events and various junctions.

It should be noted that ArchiMate behavioral constructs have no precisely defined semantics. In fact, translation from ArchiMate specification to NuSMV assigns a semantics, which, although arbitrarily selected, follows a certain intuition, e.g. how to interpret an activity or an event.

**Definition 1** (ArchiMate model)**.** ArchiMate model $AM$ is a tuple $\langle V, E, C, R, v, e\rangle$, where

- $V$ is a set of vertices,
- $E \subset V \times V$ is a set of edges,
- $C$ is a set of ArchiMate element types,
- $R$ is a set of relations,
- $vt\colon V \to C$ is a function that assigns element types to graph vertices
- $et\colon E \to R$ assigns relation types to edges.

As we focus on business layer elements that are used to specify behavior, it is assumed that $C = \{$*Process*, *Function*, *Interaction*, *Event,Junction*, *AndJunction*, *OrJunction*, *Other*$\}$ and $R = \{$*triggering*, *association*, *composition*, *other*$\}$.

We will discuss the procedure of NuSMV model generation on a small process example presented in Fig. 2. The whole process is activated upon occurrence of the event *Start*. Then the subprocess *P1* is launched, which is followed by *P2*. If *P2* terminates correctly, a decision is made whether additional subproces *P3* should be executed or the control flow leads to event *End* directly. However, execution of *P2* can be interrupted by the event *Interrupt*, which redirects back to the process *P1*.

### B. NuSMV model

The basic structural unit in NuSMV language is a *module* understood as a set of variables and statements that assign initial values to variables and define a transition relation. Depending on the module definition, we may distinguish input variables corresponding to stimuli, internal state variables and output variables (actions).

Definition of a module introduces a new type that can be instantiated. Hence, it is possible to declare a variable of a module type and bind it during declaration resembling a constructor call to a number of input variables. Subsequent variables definitions may reference outputs of other modules instances as their inputs. This allows to define a system of communicating state machines of desired complexity, which propagates input stimuli to its components causing subsequent state changes and generation of output signals. Typically, the model integration is achieved within the special *main* module, however, it can be distributed among lower level modules, which are referenced from *main*.

Fig. 3 shows the structure of NuSMV model corresponding to the process in Fig. 2. Although the structure of the presented process is clearly sequential, it is realized by a number of concurrent state machines (modules) linked by their output

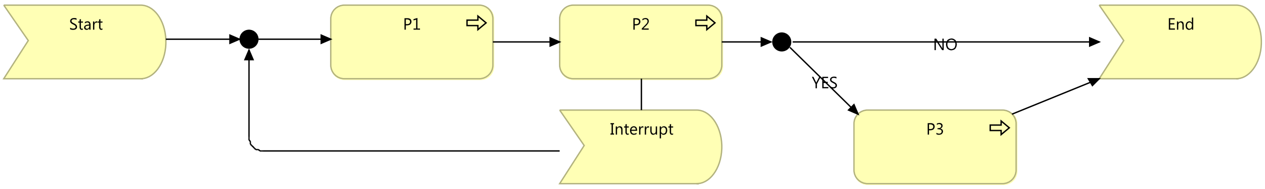


Fig. 2. Sample ArchiMate process specification

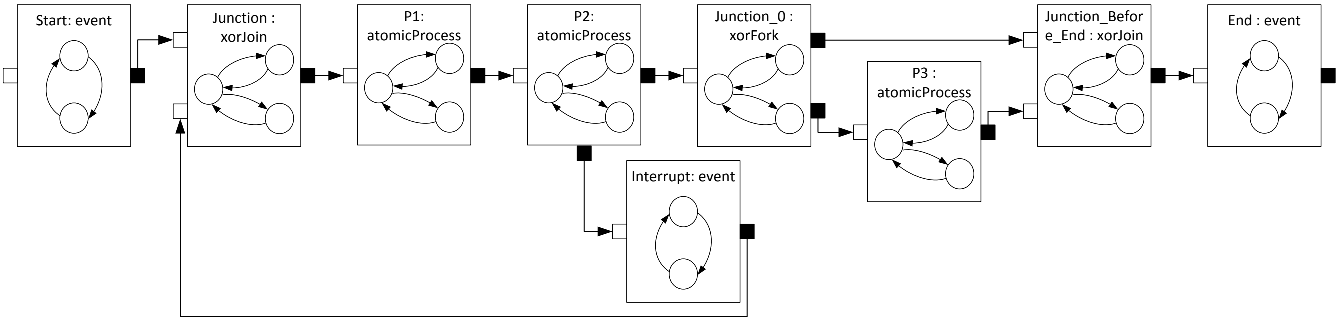


Fig. 3. Linked NuSMV modules used to model the process in Fig. 2

and input variables. The reusable modules correspond to the language constructs: processes, events, forks, joins, etc.

After conducting an analysis of components used to describe ArchiMate processes the following basic modules were identified and implemented:

- $atomicProcess_n$: $n$-ary atomic process has exactly one input, one primary output and $n$ additional outputs, which can be activated if one of $n$ exceptions occurs. The exception should be modeled in ArchiMate as an event linked with the process by the association relation.
- $event$: has only one input and one output (a boolean flag). Multiple recipients may use this flag as trigger.
- $andFork$: used to model AndJunction in Archmate. The module construction is analogous to event.
- $andJoin_n$: $n$-ary andJoin produces output signal, if all $n$ inputs are set to TRUE.
- $xorFork_n$: $n$-ary xorFork have one input and $n$ outputs. Upon module activation, only one among possible outputs will be triggered.
- $xorJoin_n$: $n$-ary xorJoin has $n$ inputs and sets the output flag if any of them is set. Moreover it tracks the number of inputs, e.g. if two from $n$ inputs are activated, the output flag will be set twice.

Fig. 4 shows the state diagram of the module `atomicProcess1`. The number "1" appearing in the module name indicates the number of additional outputs, which can be set as a result of exception occurrence. The process is activated by the input signal *trigger*. Upon the signal arrival it changes the state from *idle* to *started*. Then a choice can be made between the states *finished* and *interrruptedl*. Synchronously, the corresponding output variable is set: either *outflag* or *exccptflag1* to *TRUE*. The output variable, whichever is set, will be cleared during the transition to *idle* state.

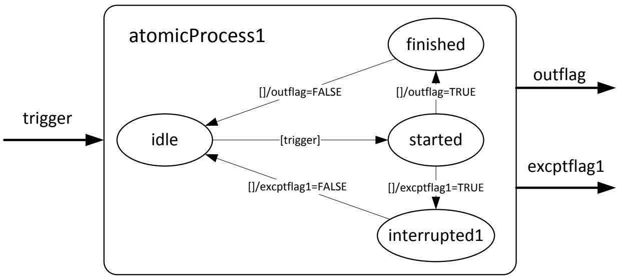


Fig. 4. State machine modeling an atomic process

The NuSMV code for the module is given in Fig. 5. It should be mentioned, that in the case of a process having $n$ exceptional outputs, we generate module `atomicProcess_n` with states *interrupted1*,. . . , *interrruptedn* and $n$ output flags *exceptflag1*,. . . , *exceptflagn*.

Basically, all modules corresponding to ArhiMate language elements were implemented as state machines, which receive input(s), change their internal states and produce outputs. However, due to performance issues we provided also alternative synchronous implementations, which immediately compute output values based on inputs. An example of synchronous process implementation is given in Fig. 6. Its definition uses an invariant that restricts acceptable combinations of variables instead of a transition relation. Basically, synchronous implementations allow to reduce interleaving, what makes the internal model representation in NuSMV smaller and speeds up the verification process. However, synchronous implementations are rather intended to be used

```
MODULE atomicProcess1(trigger)
VAR
  state : {idle,started,finished,interrupted1};
  outflag : boolean;
  excptflag1 : boolean;
ASSIGN
  init(state) := idle;
  next(state) :=
    case
      state = idle & trigger: {started};
      state = started : {finished,interrupted1};
      state = finished & !outflag : idle;
      state = interrupted1 & !excptflag1 : idle;
      TRUE : state;
    esac;
  init(outflag) := FALSE;
  next(outflag) :=
    case
      state = finished : TRUE;
      state = idle : FALSE;
      TRUE : outflag;
    esac;
  init(excptflag1) := FALSE;
  next(excptflag1) :=
    case
      state = interrupted1 : TRUE;
      state = idle : FALSE;
      TRUE : excptflag1;
    esac;
SPEC
  AG (trigger = TRUE -> AF (outflag = TRUE | excptflag1 = TRUE))
```

Fig. 5. NuSMV code of the module `atomicProcess1` (the number 1 indicates number of exceptional outputs)

for such elements as business events, forks or joins, than processes, for which activation and termination should be modeled as two distinguishable events separated by time.

```
MODULE atomicProcessSynchro1(trigger)
VAR
  outflag : boolean;
  excptflag1 : boolean;
INVAR
  (!trigger & !outflag & !excptflag1) |
  (trigger & outflag & !excptflag1) |
  (trigger & !outflag & excptflag1)
```

Fig. 6. Synchronous process implementation, a simplification of the state machine in Fig. 4

It should be mentioned that in the generated code is compatible with with the input language of nuXmv [30], the NuSMV succesor released at the end of 2014. In particular, the `process` keyword is not used and the choice between synchronous and asynchronous execution of model elements is entirely done within the generated model.

### C. Generation procedure

The generation procedure consists of the following stages:

1) *Refactoring.* With relation to the numbers of inputs and outputs, it is expected that elements fall into one of two classes: $1 : m$ (one input and $m$ outputs) or $n : 1$ ($m$ inputs and one output). Hence elements with the arity $n : m$ are replaced by two two elements: the first is an appropriate xorJoin or andJoin of arity $n : 1$. The second is an atomic process, event or fork of arity $1 : n$.
2) *Assigning representation.* For each ArchiMate element an appropriate NuSMV module type is selected and configured based on element type and numbers of inputs/outputs. Only required modules are generated, e.g. if the specification uses only processes with one and three exceptional outputs, only modules defining `atomicProcess1()` and `atomicProcess3()` will be generated.
3) *Main module generation.* This step comprises declaration of variables and linking them. For roots (modules without inputs) appropriate initial variables and transitions are added as well.
4) *Generation of specification.* The implemented procedure analyses the graph of elements and generates CTL specifications. See Section IV-D.

The generated NuSMV code for the *main* module is presented in Fig. 7. It can be noticed that variables definition are unordered and the code contains forward references, e.g. the output variable `Junction.output` is referenced before `P1` definition. The event *Stop* has two inputs. As the result of model refactoring an *OrJunction* (variable `Junction_Before_End`) was introduced into the model. For the event *Start* constituting a root element, the boolean variable `Start_trigger` with corresponding transition was added. The initial value of `Start_trigger` is FALSE, and the next value is chosen from {FALSE,TRUE}.

```
MODULE main
VAR
  P1 : atomicProcess(Junction.output);
  P2 : atomicProcess1(P1.outflag);
  P3 : atomicProcess(Junction_0.outflag2);
  Start : event(Start_trigger);
  End : event(Junction_Before_End.output);
  Junction : xorJoin(Interrupt.outflag,Start.outflag);
  Junction_0 : xorFork(P2.outflag);
  Interrupt : event(P2.excptflag1);
  Junction_Before_End : xorJoin(P3.outflag,Junction_0.outflag1);
  Start_trigger : boolean;
ASSIGN
  init(Start_trigger) := FALSE;
  next(Start_trigger) := {FALSE,TRUE};
```

Fig. 7. Generated NuSMV main module code for the process in Fig. 2

### D. Generation of specification

As a specification language we use CTL, which allows to formulate properties applying to a tree of computations (paths) starting from a given state. As the tree defines a set of imaginable futures, CTL is called the branching time logic. CTL formulas are combinations of two types of operators *path quantifiers* and *linear-time operators*.

The path quantifiers are:

- $A\,p$ – $p$ holds for every path in a tree and
- $E\,p$ – there exists a path in a tree, for which $p$ holds

Temporal operators include:

- $F\,p$ – $p$ holds true sometime in the future,

- $G\,p$ – $p$ holds true globally in the future,
- $X\,p$ – $p$ holds true next time and
- $p\,U\,q$ - $p$ holds true until $q$ holds true.

Usually a specification formally describing requirements is entered by a user. However, we tried to derive some *liveness* requirements based on control flows within ArchiMate model (see Definition 1).

The implemented procedure generating a set of specifications comprises four steps:

1) Build a set of paths $\Pi = \{\pi_i\}$ within the ArchiMate model,
2) Restrict elements in $\pi_i$ to events only (elements from the set $Evt$)
3) Build a partial mapping $R\colon Evt \to 2^{Evt}$
4) Generate the specification for each pair $(e_i, R(e_i))$ in $R$

In the first step (1) a depth-first search starting from *roots* (ArchiMate elements having no predecessors) is performed. It returns a set of paths $\Pi = \{\pi_i\}$ comprising ArchiMate elements linked by control flow relation. For a path $\pi_i = (e_{ib}, \ldots, e_{ie})$, its last element $e_{ie}$ is either a final element in the model (without successors) or a branching element (already present in $\pi_i$). The set of obtained paths reflects only topological relations within the process model. The procedure does not attempt to interpret the model according to any behavioral semantics. This is left to the verification tool.

For the example presented in Fig. 2 the set of paths $\Pi$ comprises three elements:
$\pi_1$ = (Start, Junction, P1, P2, Junction_0, Junction_Before_End, End)
$\pi_2$ = (Start, Junction, P1, P2, Junction_0, P3, Junction_Before_End, End)
$\pi_3$ = (Start, Junction, P1, P2, Interrupt, Junction)

Any requirements specification must reference terms, in which the model is expressed. We decided to focus on elements of *Event* type, which in business process definitions are typically used to mark important process states (e.g. initial, final and intermediate events).

In the step (2) the paths from $\Pi$ are restricted to ArchiMate elements being events. For the discussed example the restricted set of paths $\Pi^r$={(Start, End), (Start, Interrupt)}.

In the next step (3) a partial mapping $R\colon Evt \to 2^{Evt}$ is built. The mapping $R$ assigns all (potentially) reachable events to first events appearing in paths from $\Pi$. Continuing the example from Fig. 2, the mapping $R$ contains only one pair: (Start, {End,Interrupt}).

Finally, in the step (4) for each event $e \in \operatorname{dom} R$, a pair $(e, R(e))$ is converted into a set of specifications taking the form of (1), where $\mathcal{G} = \{AG, EG\}$, $\mathcal{F} = \{AF, EF\}$ and $\mathcal{O} = \{\bigvee, \bigwedge\}$.

$$\mathcal{G}((f \to \mathcal{F}(\mathop{\mathcal{O}}_{l_i \in R(f)} l_i))) \tag{1}$$

```
SPEC
    AG( Start.outflag
    -> AF(Interrupt.outflag & End.outflag))
SPEC
    AG( Start.outflag
    -> AF( Interrupt.outflag | End.outflag))
SPEC
    AG( Start.outflag
    -> EF( Interrupt.outflag & End.outflag))
SPEC
    AG( Start.outflag
    -> EF( Interrupt.outflag | End.outflag))
SPEC
    EG( Start.outflag
    -> AF( Interrupt.outflag & End.outflag))
SPEC
    EG( Start.outflag
    -> AF( Interrupt.outflag | End.outflag))
SPEC
    EG( Start.outflag
    -> EF( Interrupt.outflag & End.outflag))
SPEC
    EG( Start.outflag
    -> EF( Interrupt.outflag | End.outflag))
```

Fig. 8. Generated CTL specifications for the process in Fig. 2

Fig. 8 gives specifications generated for the process from Fig. 2. The specification AG (Start.outflag -> EF (Interrupt.outflag | End.outflag)). is equivalent to the statement: *for every path, starting with Start event, it is possible to reach a state, where End or Interrrupt events occur.* This requirement is obviously true for the discussed example. On the other hand specifications, where conjunction of reachable events occurs are false. An example false specification is AG( Start.outflag -> AF(Interrupt.outflag & End.outflag)), what was justified by a counterexample trace comprising 20 elements produced by NuSMV. The types of generated specification are controlled by program parameters. In particular, generation of specifications using conjunctions of reachable events can be switched off.

## V. Experiments

The goal of the conducted experiments was to assess the efficiency of the workflow shown in Fig. 1, in which:

1) An ArchiMate specification is prepared with Archi tool.
2) With "one-click" a corresponding NuSMV model is generated.
3) A set of specifications for the obtained model is checked by calling NuSMV.

In particular we were focused on the time efficiency of the third step, as it seemed to be crucial for the presented approach. During the experiments NuSMV was launched as an external process in the interactive mode, then commands to load models, report numbers of variables and check automatically generated specifications were submitted. The NuSMV output was grabbed and information on models processed, as well as the execution times were collected.

### A. Artificial test cases

Tests reported in this section aimed at assessing the relationship between a process complexity and the time required to check CTL specifications with NuSMV. Obviously, process specifications can form very diverse structures, however, we assumed that the key feature characterizing the process complexity is the numbers of branches and loops in the control flow. Hence, the basic process pattern, which was analyzed, comprised several branches and loops placed between two events *Start* and *Stop*. Fig. 9 shows an example, in which four subprocesses are arranged to form two branches and two loops. We have prepared a number of test cases being variations of this pattern and the results were summarized in Table I.

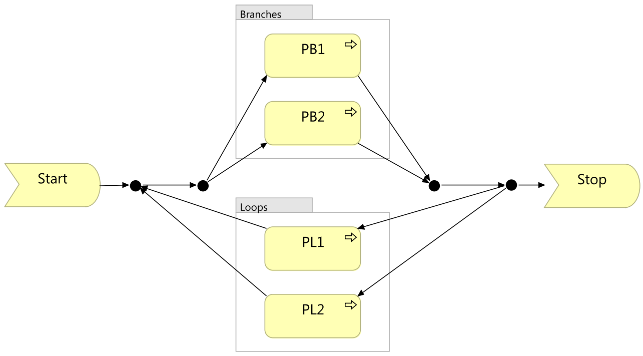


Fig. 9. Test case 4p2b2l: a process consisting of four subprocesses forming two branches and two loops

Each case name in Table I is encoded as a string: $n\,\mathrm{p}\,m\{\mathrm{b}|\mathrm{ab}\}\,k\,\mathrm{l}$, where $n$–number of processes, $m$–number of branches and $k$–number of loops. Symbol 'b' appearing in the case name means that the XOR branches (*xorJoins*) were used, wheras 'ab' corresponds to parallel branches placed between two *AndJoins*.

The column marked as $CC$ gives the value of cyclomatic complexity, a commonly used measure that can be interpreted as the number of independent paths in the specification [31]. It is calculated according to following formula: cc = e-n+2p, where where $e$ is the number of edges, $n$ is the number of nodes and $p$ is the number of connected components. The formula is applicable to sequential processes. For processes containing parallel activities the number of independent paths can be smaller, as activities starting and finishing synchronously actually build up a single control flow paths. Corrected values of cyclomatic complexity are added in parentheses.

Subsequent columns give numbers of state variables, total numbers of states and numbers of reachable states. *Dia* is a system diameter, i.e. the longest path from the initial state.

Columns *T* and *TPS* give total execution times and the execution time divided by the number of specifications (For each case 8 specifications were automatically generated and tested.)

NuSMV uses internally OBDD as a representation of a state transition system. It is well known, that the size of OBDD depends heavily on the variable ordering [8], [32], [33]. Selecting an optimal ordering is a NP-hard task, however several heuristics can be used to improve the ordering and in consequence reduce the model size, as well as the processing time. The corresponding option offered by NuSMV is called: *dynamic variable ordering*. The column *TD* gives the total processing time for dynamic variable ordering (the sift algorithm described in [34] was applied), whereas *TDPS* shows *TD* divided by the number of specifications.

The test cases are arranged into three groups. The first comprises process specifications with various numbers of branches, however without any loops. As it can be observed, in spite of growth of problem size, the numbers of reachable states and processing times remain relatively small.

The second group contains hard cases, i.e. those with several OR branches and loops. This makes both the numbers of reachable states very high and for the most complex case 24p12b12 (24 processes, 12 branches and 12 loops) the time spent to check one specification ranges to 80 seconds.

The third group includes cases with parallel branches that are synchronized at an *AndJoin*. This makes the space of reachable states much smaller, as well as keeps execution times relatively small (up to 560ms for the case 24p12ab12l). It can be noticed that for the cases 7p2ab5l, 8p3ab5l, 9p4ab5l and 10p5ab5l the numbers of reachable states are equal. This is a natural consequence of the model structure. Regardless of the number of parallel subprocesses, they all start and finish in the same states synchronized by the *AndJoin*.

Table I shows that applying dynamic variable ordering can be beneficial for complex models, e.g. for 16p8b8l, 20p10b10l and 24p12b12l it allowed to reduce the processing time up to 90%. For smaller models it was inefficient.

### B. Business process example

In this section we present results of performance tests for a medium size business process from the banking domain. The process evaluates a credit order sent over Internet and issues an approval or a rejection decision. Its description was prepared with Archi based on the EPC model published in [35] (on the page 44).

The process definition comprises five views presented in Fig. 10, Fig. 11 and Fig. 12. Following the modeling approach typical for the EPC, multiple events defining process states or conditions are used. Moreover, the events mark boundaries of subprocesses represented by the views. They bind them into a coherent model comprising branches or loops, whose endpoints are indicated by the referenced events. Due to limited space the presented models are restricted to behavioral elements, i.e. they do not comprise information on engaged roles, involved systems and processed data, which can be specified in ArchiMate.

The process is divided into three stages. The first, presented in Fig. 10, aims at collecting customer data for the received credit order. The subprocess handles two cases: when a customer is new or already present in the database of the banking system, what may influence the credit decision.

The next stage shown in Fig. 11 ends with three conditions: the credit order is accepted (green credit decision), rejected (red credit decision) or inconclusive (gray credit decision).

The final stage (Fig. 12) comprises activities, whose goals are to prepare the credit offer (in the case of green credit

TABLE I
RESULTS OF PERFORMANCE TESTS.

| Name | CC | Vars | Total states | Total states | Reachable | Dia | $T$ [ms] | $TPS$ [ms] | $TD$ [ms] | $TDPS$ [ms] |
|---|---|---|---|---|---|---|---|---|---|---|
| 2p2b0l | 2 | 15 | $2^{12} \cdot 3^2 \cdot 4^1$ | 147456 | 910 | 14 | 24.47 | 3.06 | 40.14 | 5.02 |
| 3p3b0l | 3 | 19 | $2^{15} \cdot 3^3 \cdot 5^1$ | 4423680 | 1830 | 14 | 38.18 | 4.77 | 86.03 | 10.75 |
| 4p4b0l | 4 | 23 | $2^{18} \cdot 3^4 \cdot 6^1$ | $1.27402 \cdot 10^8$ | 3118 | 14 | 50.37 | 6.30 | 110.30 | 13.79 |
| 5p5b0l | 5 | 27 | $2^{21} \cdot 3^5 \cdot 7^1$ | $3.56726 \cdot 10^9$ | 4810 | 14 | 57.27 | 7.16 | 205.65 | 25.71 |
| 8p8b0l | 8 | 39 | $2^{30} \cdot 3^8 \cdot 10^1$ | $7.04482 \cdot 10^{13}$ | 12670 | 14 | 135.37 | 16.92 | 355.86 | 44.48 |
| 12p12b0l | 12 | 55 | $2^{42} \cdot 3^{12} \cdot 14^1$ | $3.27222 \cdot 10^{19}$ | 30990 | 14 | 357.63 | 44.70 | 663.00 | 82.87 |
| 4p2b2l | 4 | 27 | $2^{21} \cdot 3^4 \cdot 4^1 \cdot 5^1$ | $3.39739 \cdot 10^9$ | 91926 | 25 | 1032.42 | 129.05 | 596.09 | 74.51 |
| 6p3b3l | 6 | 35 | $2^{27} \cdot 3^6 \cdot 5^1 \cdot 6^1$ | $2.93534 \cdot 10^{12}$ | 342294 | 25 | 1025.12 | 128.14 | 1356.48 | 169.56 |
| 8p4b4l | 8 | 43 | $2^{33} \cdot 3^8 \cdot 6^1 \cdot 7^1$ | $2.36706 \cdot 10^{15}$ | 941906 | 25 | 3808.37 | 476.05 | 2895.52 | 361.94 |
| 10p5b5l | 10 | 51 | $2^{39} \cdot 3^{10} \cdot 7^1 \cdot 8^1$ | $1.8179 \cdot 10^{18}$ | 2153330 | 25 | 11216.15 | 1402.02 | 7116.50 | 889.56 |
| 16p8b8l | 16 | 75 | $2^{57} \cdot 3^{16} \cdot 10^1 \cdot 11^1$ | $6.82405 \cdot 10^{26}$ | $1.36819 \cdot 10^7$ | 25 | 74301.78 | 9287.72 | 7552.02 | 944.00 |
| 20p10b10l | 20 | 91 | $2^{69} \cdot 3^{20} \cdot 12^1 \cdot 13^1$ | $3.21085 \cdot 10^{32}$ | $3.44569 \cdot 10^7$ | 25 | 240468.40 | 30058.55 | 22965.01 | 2870.63 |
| 24p12b12l | 24 | 107 | $2^{81} \cdot 3^{24} \cdot 14^1 \cdot 15^1$ | $1.43403 \cdot 10^{38}$ | $7.47279 \cdot 10^7$ | 25 | 640779.35 | 80097.42 | 70495.57 | 8811.95 |
| 7p2ab5l | 7(6) | 37 | $2^{29} \cdot 3^7 \cdot 8^1$ | $9.39309 \cdot 10^{12}$ | 109038 | 25 | 978.63 | 122.33 | 1428.27 | 178.53 |
| 8p3ab5l | 8(6) | 39 | $2^{30} \cdot 3^8 \cdot 8^1$ | $5.63586 \cdot 10^{13}$ | 109038 | 25 | 1011.67 | 126.46 | 1195.66 | 149.46 |
| 9p4ab5l | 9(6) | 41 | $2^{31} \cdot 3^9 \cdot 8^1$ | $3.38151 \cdot 10^{14}$ | 109038 | 25 | 1340.40 | 167.55 | 1362.44 | 170.30 |
| 10p5ab5l | 10(6) | 43 | $2^{32} \cdot 3^10 \cdot 8^1$ | $2.0289 \cdot 10^{15}$ | 109038 | 25 | 1059.06 | 132.38 | 1811.73 | 226.47 |
| 16p8ab8l | 16(9) | 61 | $2^{44} \cdot 3^{16} \cdot 11^1$ | $8.33015 \cdot 10^{21}$ | 265530 | 25 | 6205.62 | 775.70 | 4423.14 | 552.89 |
| 20p10ab10l | 20(11) | 73 | $2^{52} \cdot 3^{20} \cdot 13^1$ | $2.0414 \cdot 10^{26}$ | 418078 | 25 | 3257.10 | 407.14 | 7084.61 | 885.58 |
| 24p12ab12l | 24(13) | 85 | $2^{60} \cdot 3^{24} \cdot 15^1$ | $4.88429 \cdot 10^{30}$ | 614578 | 25 | 4484.91 | 560.61 | 12463.37 | 1557.92 |

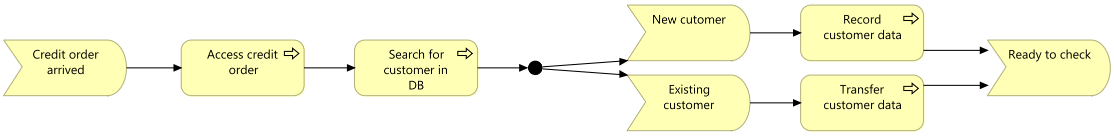


Fig. 10. Collect data

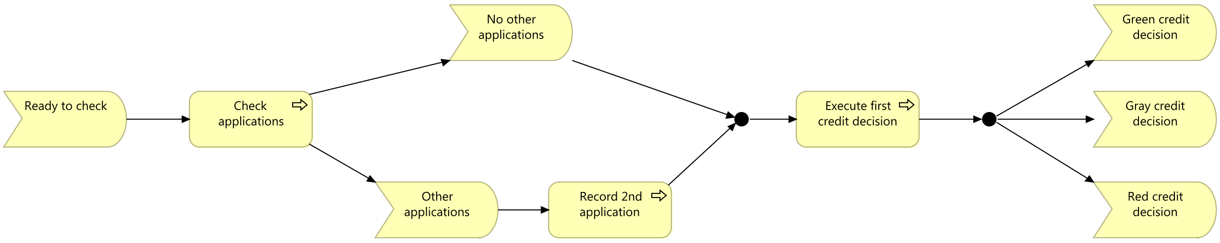


Fig. 11. First stage of the credit decision

decision), process credit order rejection for the red credit decision or reevaluate the submitted documents, if the decision was inconclusive (gray). The third case loops back the control flow to the results of the second stage.

Based on this specification the NuSMV model was generated. During the refactoring phase several joins were added before events: *Ready to check*, *Green credit decision*, *Gray credit decision* and *Red credit decision* events. An automatically generated specification checked during the experiment was the following:

```
AG( Credit_order_arrived.outflag ->
AF( Credit_offer_sent.outflag |
Contentual_problems.outflag |
Credit_order_rejected.outflag )).
```

It expresses the requirement that each credit order ends with a conclusive decision (an offer is sent or the order is rejected) or there are some problems related to the collected documents (conentual problems) that need further processing.

The cyclomatic complexity of the whole model was equal $cc = 55{-}42{+}2{\cdot}1 = 15(14)$. The value in the parentheses takes into account parallel tasks: *Create and send correspondence* and *Archive documents*.

Performing model checking for interleaving semantics of events, forks and joins failed. Without dynamic variable or-

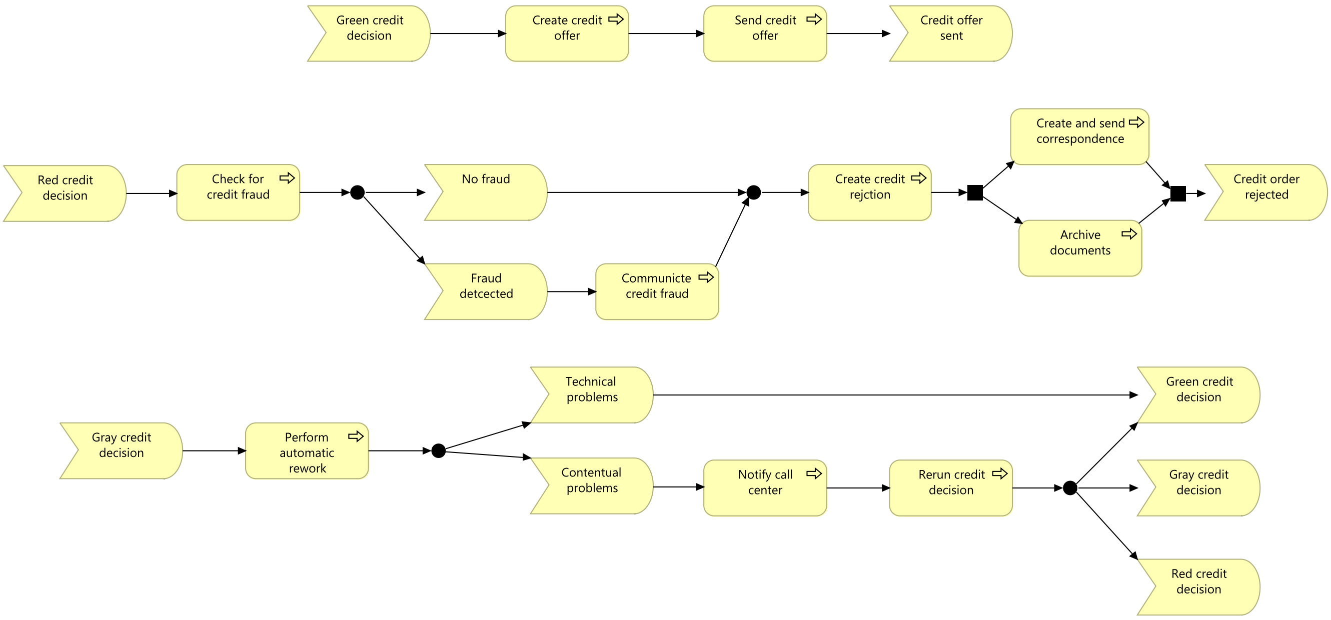


Fig. 12. Final stages of the credit decision

dering NuSMV consumed about 6GB memory. After applying dynamic ordering, the memory footprint was smaller (less then 500kB). However, in both cases the verification processes did not terminate in an acceptable time. This result was somehow disappointing, because based on the cyclomating complexity value, we were expecting that the verification is feasible.

For the synchronous semantics of events, binary forks and joins the model verification succeeded. The total number of states was equal $2^{53} \cdot 3^{17} \cdot 5^{2} \approx 2.90798 \cdot 10^{25}$. The number of reachable states was equal $3.05508 \cdot 10^{9}$ (i.e. greater by 2 orders of magnitude than the most complex processes in Table I). Without dynamic variable ordering the procesisng time was really long: 768522.57 ms = 12.5 min. However, applying dynamic variable reordering (sift method) reduced it to 41706.78 ms. We consider this result acceptable.

## VI. Conclusions

This paper investigates the problem of automatic verification of behavioral specification embedded within ArchiMate models. We propose an approach consisting in fully automatic translation of ArchiMate specification into a model in NuSMV language and then verifying it with the NuSMV model checker. Requirements specification in form of CTL formulas can be entered by user, but the implemented tool is capable of generating specifications based on analysis of control flows.

The main concern of our work was the time efficiency of the verification process. We tested it on a set of of artificial business process specifications, as well as on a real business process example.

The results show, that the described approach can be applied in an interactive verification tool, however, due to state space explosion problem, verification of complex business process specifications can still be a challenge for symbolic model checkers. Hence, dedicated model generation techniques focusing on keeping models compact, e.g. simplifying the models, avoiding interleaving and generating partial models, should be employed.

Although the presented considerations are related to processes defined ArchiMate language, the results of tests are applicable to process models defined in other languages including BPMN and EPC.

## References

[1] S. Morimoto, “A survey of formal verification for business process modeling,” in *Proceedings of the 8th international conference on Computational Science, Part II*, ser. ICCS ’08. Berlin, Heidelberg: Springer-Verlag, 2008. doi: 10.1007/978-3-540-69387-1_58. ISBN 978-3-540-69386-4 pp. 514–522.

[2] OMG, “Business Process Model and Notation (BPMN) version 2.0,” OMG, Tech. Rep., January 2011. [Online]. Available: http://www.omg.org/spec/BPMN/2.0

[3] A. Scheer, *Aris - Business Process Modeling*, ser. ARIS - Business Process Modeling. Springer, 1999, no. v. 2. ISBN 9783540644385

[4] A.-W. Scheer and M. Nüttgens, “ARIS architecture and reference models for business process management,” in *Business Process Management*. Springer, 2000, pp. 376–389.

[5] The Open Group, *Open Group Standard. Archimate 2.1 Specificattion*. Van Haren Publishing, Zaltbommel, 2013. ISBN 978 94 018 0003 7

[6] E. M. Clarke and J. M. Wing, “Formal methods: State of the art and future directions,” *ACM Computing Surveys (CSUR)*, vol. 28, no. 4, pp. 626–643, 1996.

[7] E. M. Clarke, W. Klieber, M. Nováček, and P. Zuliani, “Model checking and the state explosion problem,” in *Tools for Practical Software Verification*. Springer, 2012, pp. 1–30.

[8] R. E. Bryant, “Symbolic boolean manipulation with ordered binary-decision diagrams,” *ACM Computing Surveys (CSUR)*, vol. 24, no. 3, pp. 293–318, 1992.

[9] R. Huuck, "Formal verification, engineering and business value," in Proceedings First International Workshop on *Formal Techniques for Safety-Critical Systems,* Kyoto, Japan, November 12, 2012, ser. Electronic Proceedings in Theoretical Computer Science, P. C. Olveczky and C. Artho, Eds., vol. 105. Open Publishing Association, 2012. doi: 10.4204/EPTCS.105.1 pp. 1–4.

[10] P. Beauvoir, "Archi, archimate modelling tool," 2015, [Online; accessed March 2015]. [Online]. Available: http://www.archimatetool.com/

[11] A. Cimatti, E. Clarke, E. Giunchiglia, F. Giunchiglia, M. Pistore, M. Roveri, R. Sebastiani, and A. Tacchella, "NuSMV 2: An opensource tool for symbolic model checking," in *Computer Aided Verification*. Springer, 2002, pp. 359–364.

[12] P. Szwed, "Verification of ArchiMate behavioral elements by model checking," in *Computer Information Systems and Industrial Management*, ser. Lecture Notes in Computer Science, K. Saeed and W. Homenda, Eds. Springer International Publishing, 2015, vol. 9339, pp. 132–144. ISBN 978-3-319-24368-9. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-24369-6_11

[13] G. J. Holzmann, "The model checker SPIN," *IEEE Transactions on software engineering*, vol. 23, no. 5, pp. 279–295, 1997.

[14] G. Behrmann, A. Cougnard, A. David, E. Fleury, K. G. Larsen, and D. Lime, "Uppaal-tiga: Time for playing games!" in *Computer Aided Verification*. Springer, 2007, pp. 121–125.

[15] K. Watahiki, F. Ishikawa, and K. Hiraishi, "Formal verification of business processes with temporal and resource constraints," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1173–1180.

[16] B. Anderson, J. V. Hansen, P. Lowry, and S. Summers, "Model checking for e-business control and assurance," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 35, no. 3, pp. 445–450, 2005. doi: 10.1109/TSMCC.2004.843181

[17] M. Mongiello and D. Castelluccia, "Modelling and verification of BPEL business processes," in *Model-Based Development of Computer-Based Systems and Model-Based Methodologies for Pervasive and Embedded Software, 2006. MBD/MOMPES 2006. Fourth and Third International Workshop on*. IEEE, 2006, pp. 5–pp.

[18] X. Fu, T. Bultan, and J. Su, "Formal verification of e-services and workflows," in *Web Services, E-Business, and the Semantic Web*. Springer, 2002, pp. 188–202.

[19] A. Deutsch, R. Hull, F. Patrizi, and V. Vianu, "Automatic verification of data-centric business processes," in *Proceedings of the 12th International Conference on Database Theory*. ACM, 2009, pp. 252–267.

[20] M. T. Wynn, H. Verbeek, W. M. van der Aalst, A. H. ter Hofstede, and D. Edmond, "Business process verification-finally a reality!" *Business Process Management Journal*, vol. 15, no. 1, pp. 74–92, 2009.

[21] W. M. Van der Aalst and A. H. Ter Hofstede, "YAWL: yet another workflow language," *Information systems*, vol. 30, no. 4, pp. 245–275, 2005.

[22] R. Klimek and P. Szwed, "Verification of ArchiMate process specifications based on deductive temporal reasoning," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, Kraków, Poland, September 8-11, 2013.*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2013, pp. 1103–1110. [Online]. Available: http://fedcsis.org/2013/

[23] R. Klimek, P. Szwed, and S. Jedrusik, "Application of deductive reasoning to the verification of ArchiMate behavioral elements," *Informatyka Ekonomiczna*, vol. 29, pp. 76–97, 2013.

[24] E. M. Clarke, O. Grumberg, H. Hiraishi, S. Jha, D. E. Long, K. L. McMillan, and L. A. Ness, "Verification of the futurebus+ cache coherence protocol," *Formal Methods in System Design*, vol. 6, no. 2, pp. 217–232, 1995.

[25] A. Fuxman, M. Pistore, J. Mylopoulos, and P. Traverso, "Model checking early requirements specifications in tropos," in *Requirements Engineering, 2001. Proceedings. Fifth IEEE International Symposium on*. IEEE, 2001, pp. 174–181.

[26] P. Bertoli, A. Cimatti, M. Pistore, M. Roveri, and P. Traverso, "Mbp: a model based planner," in *Proc. of the IJCAIŠ01 Workshop on Planning under Uncertainty and Incomplete Information*, 2001.

[27] E. Clarke and W. Heinle, "Modular translation of statecharts to smv," Citeseer, Tech. Rep., 2000.

[28] M. Szpyrka, A. Biernacka, and J. Biernacki, "Methods of translation of Petri nets to NuSMV language," in *Proceedings of the 23th International Workshop on Concurrency, Specification and Programming, Chemnitz, Germany, September 29 - October 1, 2014.*, ser. CEUR Workshop Proceedings, L. Popova-Zeugmann, Ed., vol. 1269. CEUR-WS.org, 2014, pp. 245–256. [Online]. Available: http://ceur-ws.org/Vol-1269/paper245.pdf

[29] P. Szwed, W. Chmiel, S. Jedrusik, and P. Kadluczka, "Business processes in a distributed surveillance system integrated through workflow," *Automatyka/Automatics*, vol. 17, no. 1, pp. 127–139, 2013.

[30] R. Cavada, A. Cimatti, M. Dorigatti, A. Griggio, A. Mariotti, A. Micheli, S. Mover, M. Roveri, and S. Tonetta, "The nuxmv symbolic model checker," in *CAV*, ser. Lecture Notes in Computer Science, A. Biere and R. Bloem, Eds., vol. 8559. Springer, 2014. ISBN 978-3-319-08866-2 pp. 334–342.

[31] T. McCabe, "A complexity measure," *Software Engineering, IEEE Transactions on*, vol. SE-2, no. 4, pp. 308–320, Dec 1976. doi: 10.1109/TSE.1976.233837

[32] H. R. Andersen, "An introduction to binary decision diagrams," *Lecture notes, available online, IT University of Copenhagen*, 1997.

[33] P. Szwed and A. Ligeza, "Application of OBDD diagrams in verification of tabular rule systems," *Schedae Informaticae*, vol. 14, 2005.

[34] R. Rudell, "Dynamic variable ordering for ordered binary decision diagrams," in *Proceedings of the 1993 IEEE/ACM international conference on Computer-aided design*. IEEE Computer Society Press, 1993, pp. 42–47.

[35] B. Weiß, "Business process modelingand analysis in banks," 2011, [Online; accessed April 2015]. [Online]. Available: http://www.bpm.scitech.qut.edu.au/seminars/2011/BurkhardWeissBPMSeriesTalk.pdf

DOI: 10.15439/2015F113
ACSIS, Vol. 5

# Real-time Direct Translation System for Sinhala and Tamil Languages

Rajpirathap S, Sheeyam S, Umasuthan K, Amalraj Chelvarajah
Faculty of Information Technology, University of Moratuwa, Sri Lanka
Email: nova-fit10@googlegroups.com, amalraj@uom.lk

***Abstract*—Language barriers in day to day communication are common in all countries. In Sri Lanka we have a rising need for translation for Sinhala and Tamil to reduce language barriers and the statistical machine translation approach is more suitable for the concerned languages. Statistical machine translation method is one of the most promising and efficient method to perform machine translation for Sri Lankan languages likes Sinhala and Tamil. Statistical approach is more suitable for structurally dissimilar pairs of languages and efficient solution for large text translation. Sinhala and Tamil have a similarity in grammar and statistical approach will help to obtain more accurate results. We have developed a Real-time bi-directional translation system for both Tamil to Sinhala and Sinhala to Tamil for this research. We have used the Sri Lankan parliament corpus to train the language model. We have critically evaluated the both systems with parameter optimizations and have obtained the most accurate and efficient system. We have also utilized the scoring techniques like BLEU [2, 8] & NIST [2] for the system evaluation and we have integrated the MERT technique to tune the decoder.**

***Index terms*—Statistical machine translation, Natural language processing, Sinhala, Tamil, Machine Translation**

## I. Introduction

### A. Background

A*UTOMATIC machine translation* is one of the main concepts in simulating *Human Intelligence* which has several researches going on. The functionality of machine translation is to perform translation activities on natural languages which are complex and ambiguous. Machine translation comes under the learning area of Natural Language processing which depends on the subject areas like statistics, linguistics and computer science. Machine translation can be defined as *"Automatic translation from one language to another using computing devices and algorithms"*. A translation approach is about combining many techniques into one to get the translation right. Machine Translation can be divided into approaches such as transfer approach, Interlingua approach, direct approach and corpus based approaches which has two types like Example based and Statistical based)

### B. The Statistical Approach

According to proven results statistical machine translation is one of the most efficient and effective translation approaches which is well matched with western and Indic languages. Statistical MT approach helps to finish up with mathematically easily decomposable model. It is not very complicated when compared to other rule based approaches and has been proven as the most promising approach to all-purpose text translation. SMT approach has the standard algorithms and models available which can be applied to any language pair, with large corpora with few linguistic assumptions. This help to minimize the development duration.

### C. Goal, Objective, Scope & Motivation

The main goal of this research is to develop a real-time communication system which can perform statistical machine translation for Tamil and Sinhala.

Our objective is to research on the machine translation domain and create an efficient and accurate system than existing ones. The real-time system we have developed for this research performs machine translation in a very efficient way and can be used as a core for many types of software. Other than the development part of this system we also have some sections like MERT tuning, system evaluation using BLEU and NIST metrics.

The scope of this project is to develop a real-time instant text communication tool which can translate Sinhala and Tamil bi-directionally. The translation output is based on the type of the language corpora we use to implement the system. And the training data is updated continuous data from users through the reporting system which will increase the quality of the output with many users using it. First few months has been spent on research and learning of Natural language processing concepts. Few months period has been spent for the design and Implementation of the system. Software tools namely GIZA++, Moses, IRSTLM, MERT Module, NIST & BLEU module are the components that are built-in in the system.

The motivation to try out this project is to reduce the unavailability of translation systems which can be used for instant messages especially for Tamil and Sinhala. And previous researches in this domain are discontinued or not visibly used. Languages concerned in our researches are Tamil and Sinhala. When analyzing about the two languages the sentence structure on both languages seems to be identical in many occurrences. Disregard to some insignificant variations, the two languages has the same grammatical structure which is adequate to maintain the meaning understandable in both languages.

We are using Statistical Machine translation because it gives good results for even dissimilar language pairs and using rule based approaches for languages like Tamil & Sinhala will consume a lot time, effort and the result will be error full and inaccurate.

978-83-60810-66-8/$25.00©2015, IEEE

### *D. Statistical Machine Translation*

Statistical translation systems work by learning how the grammar of source and target languages are defined. They begin to work with less number of dictionary entries and language resources. Developers can train the system by increasing the entries to handle complex and extensive translation scenarios. Google is one big player in SMT related NLP applications.

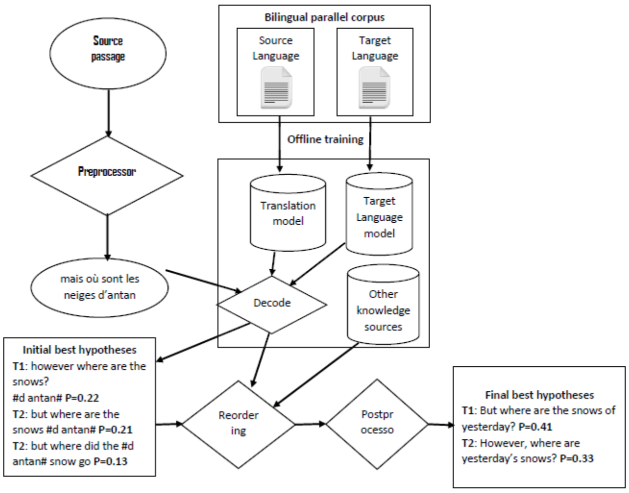

Source passage
Preprocessor
mais où sont les neiges d'antan
Bilingual parallel corpus
Source Language
Target Language
Offline training
Translation model
Target Language model
Decode
Other knowledge sources
Initial best hypotheses
T1: however where are the snows?
#d antan# P=0.22
T2: but where are the snows #d antan# P=0.21
T2: but where did the #d antan# snow go P=0.13
Reorder ing
Postpr ocesso
Final best hypotheses
T1: But where are the snows of yesterday? P=0.41
T2: However, where are yesterday's snows? P=0.33

*Diagram 01 : Statistical Machine Traslation*

Statistical MT systems usually work or get trained by breaking sentences into n-grams. Analyzing n-grams will improve accuracy and performance. A word may have many types of meanings but it will only have fewer meaning in a phrase form. Most of the SMT systems work on bi-gram and commonly tri-gram. Tri-gram simply means as three –word groups. Three-word groups are more than enough to process efficiently on the data set. Larger n-gram will require more power and time to analyze and translate. In a statistical machine translation system the common n-grams are tracked and more frequent translations are learned and used in the future translation activities. Statistical analysis of n-grams are happened which will analyze the position of the n-Gram with regard to the sentence. Mathematical translation models are updated after each and every translation which will result in accurate results. Faster processer will improve performance of training and tuning and reduce the time taken to it. Technologies such as C++, Java and Perl are used to develop NLP algorithms. As the languages are platform independent. They will enable their functionality on many Operating Systems. Training the SMT system extensively will lead to a highly accurate translation system. Compared to Rule based systems Statistical systems saves more time, money and effort. Rule-based systems for a language pair like Tamil & Sinhala will require many years and considerable amount of funding. Statistical systems can be trained to a good accuracy level in few months. This makes statistical approach less work intensive and that is the main reason we choose this approach for our short-term research purpose.

The SMT approach is more useful for corporate and government applications. One of the limitations in a Statistical machine translation approach is that it will require powerful computers for large training and accurate translation. The main concept in SMT systems is to select the target language phrase which has the maximum probability of being the translation on the source language phrase. A probabilistic model is a base for all the computations in the SMT system. Assumptions on Bayes Rules and Noisy channel model make the system less complex when generating the probabilistic model. When an input sentence is given into the system the target phrase with more probability will be selected as the output. This statistical translation approach can be briefly described in mathematical terms

$$\mathbf{t = argmax_t\ P\ (t/s)\ \ldots\ldots\ldots\ldots\ldots\ (1)}$$

Using Bayes' theorem, the value of P (t/s) can be given as

$$\mathbf{P\ (t/s) = (P\ (t)\ P(s/t))/\ (P(s))\ldots\ldots\ (2)}$$

Referring to equation (2), t can be written as,

$$\mathbf{t = argmax_t\ (P\ (t)\ P(s/t))/\ (P(s))\ldots\ (3)}$$

In equation (3), s is fixed in source language. As a result of this, P(s) can be removed from the equation when finding the sentence t. That has maximum probability. Then equation (3) becomes as equation (4)

$$\mathbf{t = argmax_t\ (P\ (t)\ P(s/t))\ldots\ldots\ldots\ldots\ (4)}$$

Element P (t) in equation 4 denotes the kind of sentences in Target language (T). P (t) is called Language Model of target language (T). The other element in equation 4, P (s/t) specifies how each sentence in target language (T) can be translated into source language sentences (s). This is called as the Translation Model. Equation (4) lays the foundation of statistical machine translation with the specification of two key components namely language model (P (t)) and translation model (P(s/t)).

Improving the language model in the SMT system is a manual job but we can use MERT techniques to train the system in minimum error rate which will result in accurate systems. MERT recommends a substitute training strategy for log-linear statistical translation models. MERT training is a straight forward method which will optimize translation quality using some automatic metric score. During the weight optimization of decoder parameters such as phrase translation table, language model, distortion, word penalty etc. MERT searches weight values that reduce translation errors. Metric scores supported by MERT are BLEU, NIST and TER. As the manual evaluation is hard and time consuming job, to make the evaluation more flexible metrics like BLEU and NIST were introduced. BLEU metric is a de facto standard which is used for Machine translation system evaluation. BLEU score gets the geometrical mean of modified precision score of the test corpora and multiply with some exponential brevity penalty factors. BLEU score increases to a higher value when the number of reference translations is increased. NIST [2] is a metric built on top of BLEU [2, 8] which enhances it with few modifications. One of the major modifications done by NIST is it assigns higher weights to rare or less occurring n-grams than regular ones. It simply applies the smoothing technique on rarely used n-grams.

## II. Related Work

There are numerous researches and projects developed on Statistical machine translation in the recent past. Sri Lankan developers have worked on many Sinhala, Tamil related NLP systems in the past few years. In the reference [3], the developers had developed statistical machine translation systems for 4 European based languages such as English, German, Spanish and French. The data they used was from a general domain and large in size. German to English SMT systems has an average BLEU score of 0.236. Spanish to English SMT systems has an average BLEU score of 0.340. French to English SMT systems has an average BLEU score of 0.316. The research [1] is a SMT implementation for English and Sinhala Languages. This research includes several refinements. This research talks about MERT inclusion, Translation Model tuning, Language Model Tuning and various word alignment and reordering techniques. The BLEU score obtain at the end of the research was around 0.1500 which is a very low value.

The research [2], which was developed by UCSC, is one of the best references for SMT implementations. This SMT system was implemented for Sinhala and Tamil Languages. This research also uses the data of specific domain which is from various public websites. The best BLEU score obtained from this research for Sinhala to Tamil system was 0.185. Even though it's not a very successful research this paper talks about some viable techniques and approaches for SMT based applications. This research uses some old tools and this research was just the beginning of a successful SMT system. The further researches by this team have given great result on evaluation. This research is one main reference for our current research. After reading all the related work we learned every strength and weaknesses of the systems and have embedded that knowledge into our research to obtain higher scores. We have used newer and updated techniques to our current research and we have also contacted the previous researchers on this domain and brainstormed ideas with them to come up with a good final output. The innovative thing we are proposing through our research is using this traditional SMT concept and implementing them into a real time communication application. This application will be less complex and will support all kind of users facing language barriers during communication.

## III. Data Preparation

In Sri Lanka when developing applications on Machine translation the only resource for datasets on Tamil and Sinhala is the parliament order papers on budget proceedings. We have used an electronic version of the parliament order papers which has parallel data on both Sinhala and Tamil. We used over 5000 phrases from each language which is totally more than 10000 sentences and more than 100000 words to train the system. One of the main reasons for the lack of development in translation systems in Sri Lanka is the unavailability of the corpus and the restrictions to obtain them. We obtained the Sinhala and Tamil Electronic version in a dirty manner where it is not aligned properly with the languages. The initial stages of the research was planned to develop a SMT systems in the chat domain for Sinhala & Tamil. After the research we found that parallel data Sinhala and Tamil in the chat domain is not available and building or creating one would require a lot of funding in the current situation. We limited ourselves to the parliament domain. The reason for using the parliament order papers is that the parliament members speak more formal languages which are not commonly used in the society. The languages are translated into more exact order in the parliament because in parliament discussion every word and phrase is important. We can 100% ensure that the meaning is not altered due to the variation in the language style. This process applies to both Sinhala and Tamil. One of the drawbacks of the system is that when we use a parliament corpus for the research the system is tuned to domain of parliament style translation so that we cannot use this system to translate non formal type of sentences or phrases. The obtained electronic versions of the parliament order papers were pre-processed before using them as inputs. Reasons for pre-processing are unwanted gaps between words, disordered words, disordering of Tamil Sentences aligned to Sinhala Sentences and some complex character are broken which makes it a garbage character. After cleaning the Sinhala and Tamil parallel data we can ensure that most of the sentences are in proper order in allowed manner, but still some issues will remain. We have to assume the pre-processed data as a perfect one. Redundant contents are removed randomly in some selected files and the final data set for the system is prepared. In a Statistical machine translation system implementation data preparation is essential and plays a major role in the final results. After the preparation of the datasets we did the selection process of the Data sets. We divided the data set into three such as

Training set, Tuning Set and Testing set. The table is given below. This table shows the number of word and phrases in the data sets.

TABLE 1: DATA SETS

| Data Sets (Words and Phrases) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Language Model | | Training Set | | Tuning Set | | Testing Set | |
| Si | Ta | Si | Ta | Si | Ta | Sin | Ta |
| 165k | 77k | 99k | 78k | 3425 | 3078 | 3110 | 3204 |
| 6550 | 6104 | 5887 | 5887 | 200 | 200 | 200 | 200 |

Addition to the above mentioned datasets, as a team we have prepared our own dataset which includes parallel sentences of normally used phrases in Sinhala and Tamil. We prepared more than 2500 parallel sentences and in our system we have added a feature where the user can report the admin with the translations. These translation sentences will be added to the system's training data and the SMT core will tuned frequently in a periodical manner. This will improve the output quality. One of the reasons we started to develop a real time SMT chat application is that we can create new and improved trained data using users' contributions. This is will lead to an improved data set creation which can be used for future projects related to SMT as well.

## IV. OUR APPROACH

The system design is the heart of all system implementation. The design section consists of three main parts such as Architecture Design, Implementation Design and Evaluation Design. Architecture design is all about explaining how the modules and components are connected and working together. The implementation design illustrates about the implementation of folder structures and file paths which are used in the SMT system. The third one, evaluation design explains about how evaluation and experiments are carried out on the system to arrive at conclusions and results.

### *A. Architecture Design*

The main Architecture of this Real-time Statistical Machine Translation system is basically not an innovative one or a new thing to be researched. It has a long standing research history in its hands. The main component of this project is the SMT module and we don't need to alter the entire design of the SMT but have to follow the best practices. SMT approach is language independent. Many of the SMT implementations follow the common architecture. In our research, the both systems use this uncomplicated architecture as its backbone. Rather than mixing all the modules together, we are defining the components explicitly and linking them according to the need. The SMT system we are developing for this research is a layered architecture. This architecture helps us to make the system uncomplicated and easy to understand. The layered architectures help to add and remove components.

In the design when considering about the data preparation component, the functionality of it is to obtain the whole data set and divide into subsets. These separate data sets are used for training, tuning and testing purposes, which will lead to efficient system evaluation. Addition to this we can do language modeling if a monolingual corpus is available. The whole data set is divided into two subsets. One is for language modeling and other one is for model evaluation. Model evaluation will not require huge data set. Before these steps the whole data set is tokenized. Formed outputs are the inputs to language modeling component. In case of Sinhala to Tamil system, Tamil language model should be produced. The un-tokenized whole data set is divided into three parts for training, testing and tuning. Training data set is the input of translation modeling component. Tuning data set is the input for the automatic tuning & decoding component and testing data set is the input for evaluation component.

The Language modeling module consists of all executable files created by IRSTLM. N-gram and n-gram count are two modules needed for LM modeling and evaluation respectively. Outputs of the language modeling component is the Language model and the perplexity scores. To obtain an optimal language model from evaluation, it is required execute the LM with various smoothing and discounting parameters and obtain the scores. Manual evaluation scores after each language modeling would help decide the optimal one. The next component is the translation model which is created using the training set. We have referred GIZA++ tool to implement this module which creates the word alignment and translation model. We can adjust the word alignment and reordering strategies to create or obtain the best Translation Model. Decoder configuration file is created at the end of this procedure so that the decoder script will start running by referring this script. The tuning component is responsible for changing decoder parameters & weights modifications which will result in change in evaluation metric scores like BLEU & NIST. The next component is the decoder executable. The inputs for this module are the test data set and the TM components such as phrase table and word alignment table and the decoder configuration file. The actual translation function happens inside this module. Source language file is translated and the translation process stops at the formation of translated output file. Prior to the automatic evaluation via three metrics, a little formatting is essential on the input data sets of metric-modules. Inputs of metric modules are translated output, reference text and source text. In Sinhala to Tamil system, reference file needs to be in Tamil which is the genuine translation of the Sinhala test data set. Scoring is restricted for only one reference file. Other than these above modules we have another two separate modules which are called as back up module and reload module. Back up module backups all system inputs, configurations and the outputs. Reload module reloads the backup data and configurations into the system. With all these components we have also having a simple client to client chat or communication application to fulfill the scope. This application uses the SMT module as the core and serves the users who want communicate in their native languages without any barriers.

### *B. Implementation Design*

For a successful research a good Design is essential. As developers we have to also focus on implementation design to achieve what we wanted. Having good design architecture isn‘t lonely enough. Implementation design consist the details of folder and directory structure of the implementation which is very important for the performance and the efficiency of the system. A proper folder structure will also help for easy maintenance which will reduce errors and unwanted exceptions.

In our SMT system we have bi-directional implementation which includes Sinhala to Tamil and Tamil to Sinhala. Both the systems share some common files in the system. Some tools are configured and made common to both systems. Interaction with each system is feasible through bash scripts. In our system we will have two clients who does the communication part. The input string is received from one client end and then the SMT core translates it into a target language. The target string is sent to the client in the other side.

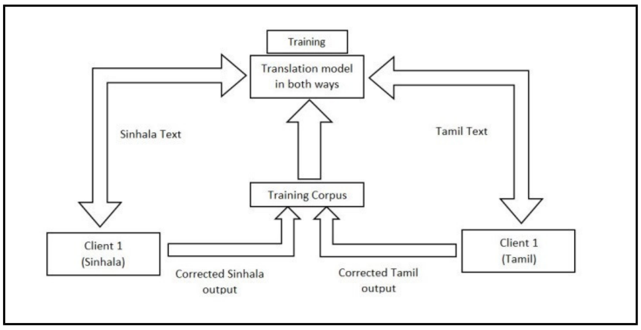


*Diagram 02 : Workflow of our Final Translation Application*

### *C. Evaluation strategy*

Evaluation of systems is an essential part of this research. A plan or strategy is developed during the development phase about how we are going to evaluate the system. The evaluation strategies we have followed here are change parameters and algorithms in different combinations and obtain BLEU and NIST score from each system. From those results we can obtain an accurate and efficient system for the languages like Tamil and Sinhala. The evaluations results are presented in a statistical form. These combinations of evaluation end in many number of system backups and as we are dealing with many evaluation techniques we can call this as a good evaluation strategy.

## V. Implementation

Major parts of our implementation are Sinhala to Tamil MT System and the Tamil to Sinhala MT System. For a successful research, excellent design architecture in hands endorses effective maintenance and help reducing mistakes. Having good design architecture isn't only enough. Implementation or the directory structure holds a place to handle executions and maintenance comprehensively. Good directory structures are formed bearing maintenance/ backup/restore simplicities in mind.

The **LM input files** directory will hold the Data sets (both Sinhala and Tamil) for language models creation and LM test files (both Sinhala and Tamil) are located here. Calculated perplexity values are written to files residing in "lm evaluation" directory. The **Input** directory will contain the inputs for both SIN-TAM and TAM-SIN system. In case of SIN-TAM, reference file (Tamil) should be the corresponding parallel corpus of test file (Sinhala). The **Output** directory will have the final de-tokenized translated output of the system. NIST, BLEU and TER scores are saved in the **Results** directory. The **Scripts** directory will contain some Bash scripts to control operations such as cleaning, backup, restore, and load backup, run components like Data model, language model, Decoder, Automatic tuning component and evaluation component of entire SMT. There is also a directory in our system called **Word** and this will store tool's output/input files.

Corpus has tokenized parallel corpora. Evaluation directory is with raw system output and formatted output files (for automatic metric evaluation). Language model is located inside "lm" directory. MERT files and tuned decoder configuration files are found in tuning directory. The same structure as that of SIN to TAM applies to TAM to SIN system also but instead of all Tamil files, Sinhala files should be there and vice versa. In backing up process, "input, output, work, lm evaluation, results and importantly all Configurations files and used commands" will be copied to a new location. This process is essential as lots of experiments need to be done and as outputs are required to be loaded again for further tests or for re-evaluation purposes.

The directories "IRSTLM, GIZA++, Moses, TER modules" have all installed files but when porting the system to another location these may be ignored as all executable are located in "executable". During our development we have used some modules from some tools such as IRSTLM, GIZA++ and Moses. The IRST Language Modeling Toolkit features algorithms and data structures suitable to estimate, store, and access very large n-gram language models. It can be used in Linux platform. IRSTLM offers the most advanced n-gram smoothing methods to estimate large LMs and approximated smoothing methods to estimate gigantic LMs. It includes methods for pruning and quantization of LMs, efficiently storing LMs on disk and in memory and it offers several language adaptation methods: linear interpolation, minimum discrimination information, and probabilistic latent semantic analysis.

Moses is one of the open source toolkits for statistical machine translation. Moses toolkit uses some tools for some of the tasks to avoid duplication such as GIZA++ for word alignments and IRSTLM for language modeling. The information about Moses was obtained from Moses documentation. The Moses toolkit has achieved some objectives such as Accessibility, Easy maintenance Flexibility, user-friendly, distributed team development and Portability. The Real-time chat application is developed using Java programming language. It is very light weight and we have developed that software in which it can be applied as a layer on top of the SMT core system and operate in good efficiency.

## VI. Evaluation

### A. Basic System Evaluation

Our evaluated system's entire language model is domain specific. And Number of 2-gram hit is relatively high in both language models. All n-gram hits in Sinhala are not as much as Tamil Language.

TABLE 2:
N-GRAM COUNTS FOR SINHALA & TAMIL

| Language Model | Sinhala | Tamil |
|---|---|---|
| 1-Gram | 4703 | 5946 |
| 2-Gram | 15337 | 16283 |
| 3-Gram | 8474 | 13280 |

TABLE 3:
AUTOMATIC METRIC SCORES FOR BOTH SYSTEMS WITHOUT MERT TUNING

| *Systems* | Sinhala - Tamil | Tamil – Sinhala |
|---|---|---|
| BLEU | 0.4277 | 0.5599 |
| NIST | 4.4090 | 4.1244 |

After the evaluation we obtained highly excellent BLEU & NIST scores in both systems and Tamil to Sinhala System gives the highest from the two systems. The BLEU score results in high values which is a proof that the system has high accuracy. In the Sinhala to Tamil system, the value of BLEU is not that best as NIST. However all these metrics favors the output of Tamil to Sinhala system in a positive manner. The both systems gives high positive values on scoring other than the existing systems developed by the past researches and experiments.

### B. Tuned system Evaluation

The good thing of the research was that we had very good score with normal settings, but more quality output can be expected after fine tuning the system. We used the Minimum Error Rate Tuning (MERT) technique to achieve more scores. After MERT [2] technique we received new scores. Below table shows the new values obtained for the both systems in a table format. The score values are in four decimal places. These scores show some great improvements.

TABLE 4:
AUTOMATIC METRIC SCORES FOR BOTH SYSTEMS WITH MERT TUNING

| *Systems* | Sinhala - Tamil | Tamil - Sinhala |
|---|---|---|
| BLEU | 0.5957 | 0.6693 |
| NIST | 4.4182 | 4.8563 |

It is very much acceptable that MERT has enhanced the scores of the both systems in a good rate. BLEU score has been improved in both systems by a good number. NIST score has been improved in the Tamil to Sinhala translation system but not much in the Sinhala to Tamil System. BLEU has been improved in the highest percentage and NIST get the least improvements. When talking about decision making of the system we have developed is that the scores are well improved values than previous research on SMT with local languages like Tamil and Sinhala. And we can mark this as a successful research. Addition to scores we have also calculated the time taken for the translations for both systems.

TABLE 5:
AVERAGE TIME TAKEN FOR TRANSLATIONS IN BOTH SYSTEMS

| *System* | *1 - sentence* | *3-sentences* | *5- sentences* |
|---|---|---|---|
| Sin - Tam | 0.778s | 1.94s | 2.052s |
| Tam - Sin | 0.035s | 0.249s | 0.299s |

## VII. Conclusion

In a summary, both translation systems have given positive results and scores. One of the beneficial things from good BLEU scores is that we can extend the research to new heights with a positive system. Word alignment algorithm we used was *grow-diag-final* and reordering algorithm was *msd-bidirectional-fe*. According to the tables mentioned in the evaluation section the BLEU & NIST scores were better than existing system and it improved even more after MERT technique. The final BLEU scores of the achieved systems are 0.595668 & 0.669333 and also achieved NIST scores are 4.4182 & 4.8563. Both systems also have very less and efficient translation times. These systems can be improved into a highly accurate system by using large dataset of training data and changing the architecture.

## VIII. Further Work

There are many future plans to improve the research and improve the system further more usable. One of the main further plans for this project is that to use a large dataset of Tamil & Sinhala phrase set to train the system. The data we used for this research for now is very minimal and with the real-time system we have developed with the reporting feature, we can hopefully prepare more amounts of translation pairs to create a translation with high quality. The limitation we faced while starting the project was the unavailability of chat data on Sinhala & Tamil. The human effort required to prepare the data is too high and costly, but we have started preparing the data using some volunteers and planning to take to the next level in the future. One step on that is the implementation of the automatic reporting system in the chat application. We will be working on the negative outcomes from this research. In the future we can differentiate the build architecture and make the translation module usable for application like Facebook and related technologies. One of our visions on this research is to make this translation functionality as a service and make it available for the developers to use it.

## Acknowledgement

The authors of this research paper wish to acknowledge the supervisors and members of Faculty of Information Technology, University of Moratuwa who helped in many ways for the research. Special thanks to UCSC, Sri Lanka for the data and the past research papers related to machine translation. The authors have adapted few technologies and research procedures for this research from their past researches.

## References

[1] J.U.Liyanapathirana, ―A Statistical Approach to English and Sinhala Translation, BSc. Thesis, University of Colombo School of Computing, Sri Lanka, July.

[2] R.Weerasinghe, ―A Statistical Machine Translation Approach to Sinhala- Tamil Language Translation.

[3] C.Callison-Burch, C. Fordyce, P. Koehn, C. Monz and J. Schroeder, ―Meta-Evaluation of Machine Translation‖, in Proc. 2nd Workshop on Statistical Machine Translation, 2007, p.136-158.

[4] Franz Josef Och, ―Minimum Error Rate Training in Statistical Machine Translation‖, in Association for Computational Linguistics.

[5] Doddington,G ―Automatic evaluation of machine translation quality using n-gram co-occurrence statistics". Proc. Human Language Technology Conference (HLT), 2002,p. 128―132

[6] R.Weerasinghe, ―A.R. bootstrapping the lexicon building process for machine translation between 'new' languages. In Proceedings of the Association of Machine Translation in the Americas Conference (AMTA), 2002.

[7] Och, F.J., Tillmann, C. and Ney, H. ―Improved alignment models for statistical machine translation. In Proceedings of the 4th Conference on Empirical Methods in Natural Language Processing (EMNLP), Maryland, 1999.

[8] K. Papineni, S. Roukos, T. Ward and W. Zhu, ―Bleu: a method for automatic evaluation of machine translation, in Proc. 40th annual meeting on association for computational linguistics -2002, 2002, pp. 311–318.

[9] P. Koehn, F. J. Och and D.Marcu, ―Statistical Phrase-Based Translation, in Proc. Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics,2002, pp. 1 – 7.

[10] Bernadette Varga, Alina Dia Trambitas-Miron, Andrei Roth, Anca Marginean, Radu Razvan Slavescu, Adrian Groza, ―LELA - A natural language processing system for Romanian tourism, in Proc. 4th International Workshop on Advances in Semantic Information Retrieval, 2014, pp. 281 – 288.

[11] Franz Josef Och, ―Minimum Error Rate Training in Statistical Machine Translation‖, in Association for Computational Linguistics, 2003, pp. 160-167.

[12] A. Birch, B. Cowan, C. Callison-Burch, M. Federico, N. Bertoldi, P. Koehn and H. Hoang, ―Moses: Open Source Toolkit for Statistical Machine Translation, in Proc. ACL 2007 Demo and Poster Sessions, 2007,pp. 177–180.

DOI: 10.15439/2015F295
ACSIS, Vol. 5

# Knowledge Management in MCDA Domain

Jarosław Wątróbski[1]  Jarosław Jankowski[1,2]

[1] West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland
Email: {jjankowski, jwatrobski}@wi.zut.edu.pl
[2] Department of Computational Intelligence, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27,
50-370 Wrocław, Poland

***Abstract*—Multi-criteria decision analysis (MCDA) methods have become increasingly popular in decision-making. Numerous methods in this field were developed to solve real-world decision problems including various engineering and scientific areas. Unfortunately, the proper use of each method is difficult due to the dispersion of the domain knowledge and lack of knowledge databases in this area.**

**The paper presents research focused on knowledge management aspects in MCDA domain. In order to achieve a high level of practicality on different levels of decision making, the ontology as a form of conceptualization is implemented.**

## I. Introduction

Together with the development of operational research, as an alternative approach evolution of MCDA methods has been observed. This applies both in theoretical studies that result in the continuous development of existing methodologies and techniques, and in the emergence of new methods as well as the application layer covering new areas of application methods in business practice.

In each case of individual specification of environmental decision-making situations, the selection of a multi-criteria method should be carried out with great care [1], and the recommendations of selection techniques for modeling and aggregation of preference should be carried out with the following considerations:

- taking into account a detailed specification of the discussed problems and pending issues, including their complex and multi-level character,
- an intuitive dialogue with the decision-maker and the user at the stage of formulating and implementing the assessment process with the subjective linguistic approach implemented,
- possible use of imprecise preference information by the evaluators, including variability of the assessment,
- the possibility of missing a situation assessment and incomparability of decision variants for multiple forms of preferential information with the deterministic, not deterministic, ordinal or fuzzy character.

The natural consequence of various areas of application of multi-criteria methods is the need for the development of dedicated approaches adjusted to the specifics of the problem. This is confirmed by a detailed analysis of the literature, where research in various scientific disciplines is effectively carried out using a number of multi-criteria methods [13] [11]. Combined with a variety of specific decision problems discussed by the authors of studies in this field, the natural direction of research can be an attempt to systematize the knowledge in this area. An additional prerequisite for undertaking research in this area is a large heterogeneity of domain knowledge including available scientific publications and the existing decision support systems.

In the literature, attempts to develop models of knowledge representation of MCDA problems and methods areas can be observed. For example, the article [6] presents an ontology designed to describe the structure of decision-making problems. It is a component of the support for the group decision-making processes. On the other hand, in [8] an ontological representation of the multi-criteria method of AHP and a set of inference rules were developed. In this way, clearly defined and formalized concepts related to the method of AHP and knowledge reusing allowed implementation in the form of ontologies. Earlier studies of systematized knowledge about various aspects of decision-making are presented [9][10]. One of the approaches is based on using ontology knowledge model integrating knowledge about decision-making process [9]. It includes such elements as the decision-making situation, the decision problem, a set of alternatives and evaluation criteria, rules, preferences and decision-makers. Then the developed model was verified within the problem of decision-making for the ERP system selection. The proposed approach was later extended by ontology components based on a generalized approach to formalization of methods of decision support [10]. The proposed ontologies focused on knowledge that makes it possible to structure knowledge about the decision problem. However, they take the problem of systematization of knowledge about the various methods of multi-criteria

978-83-60810-66-8/$25.00©2015, IEEE

decision support only to a small extent. Characterized ontologies do not include knowledge about the characteristics of the different MCDA methods and their environmental context and use cases [7].

The purpose of this article is to develop a ontology based knowledge model of MCDA methods. Taking into account the contemporary standards of knowledge engineering, it is justified to implement a repository of knowledge in the form of ontologies. In order to build the first stage of such a solution, literature related to MCDA methods was reviewed and analyzed. This formed the basis of the development of a taxonomy and ontology. Ontology as a proposed form of the conceptualization can be used as a source of knowledge that can be used repeatedly. The study was divided into two parts: a discussion of the literature and the development of a taxonomy of MCDA methods together with the ontology of MCDA methods. The article ends with the author's practical ontology verification using competency questions.

## II. Methods of multi-criteria decision support

Research on MCDA area led to the development of two main groups of methods and directions. They differ significantly from each other both in the approach to the decision situation and in the way of choosing the best alternative . These are approaches based on utility theory and outranking relations [4]. An approach based on utility theory is derived from the American MCDA school . Two kinds of relationships between alternatives are identified: indifference ($a_i$ $I$ $a_j$) and preference ($a_i$ $P$ $a_j$) of one alternative over another. The methods in this group exclude non-comparability of the decision variants and assume transitivity of preference [4]. Among the methods based on utility theory main approaches include: MAUT (Multi-Attribute Utility Theory), AHP (Analytic Hierarchy Process), and UTA (Additive Utility Theory). These methods usually do not take into account the uncertainty, vagueness and ambiguity that can occur in the data [4]. Methods based on outranking relations are derived from the European MCDA school and the outranking relationship is characterized by the lack of transitivity between pairs of decision variants.. These methods mainly include methods from ELECTRE family (ELimination Et Choin Traduisant la REalite) and PROMETHEE (Preference Ranking Organization Method for Enrichment Evaluations). Methods from this group frequently extend a set of basic preferential situations with the result that includes indifference of decision variants ($a_i$ $I$ $a_j$), weak preference one variant over another ($a_i$ $Q$ $a_j$), the strict preference of a variant of the decision-making relative to the other ($a_i$ $P$ $a_j$), and incomparability between data variations ($a_i$ $R$ $a_j$) [18]. Furthermore, couples of variants can be grouped to determine the relationship connecting the two or three basic situations. With the occurrence of the grouped relationship, it is impossible to distinguish, without additional parameters, the basic relationship of the grouped situations. Such situations are: “nonpreference” which are groups of indifference and incomparability situations ($a_i$ $N$ $a_j$), “preference” which are situations of weak and strict preference ($a_i$ $L$ $a_j$), “guess preference” which combines situations of indifference and weak preference ($a_i$ $J$ $a_j$), “K-preference” which are groups of strict preferences and situations of incomparability ($a_i$ $K$ $a_j$), and “outranking” which contains the situations of indifference strong and weak preference ($a_i$ $S$ $a_j$) [12]. Due to variations of the relationship between the decision-making, two basic approaches can be distinguished to aggregate operational performance variants : (1) aggregate to a single criterion (2) aggregation by using the outranking relationship [12]. The first operational approach excludes the incomparability situation and contains only the most indifference relationships and strict preference. It is strongly associated with the American school of decision support. The second approach takes into account incomparability and outranking and is generally used in the methods derived from the European MCDA school. Based on an analysis of the literature, a complex set of available MCDA methods was identified and the general characteristics of which are shown in Table 1. The exact classification of methods along with the characteristics is presented in the rest of the article followed by a description of the process of building a taxonomy of MCDA methods.

## III. Building taxonomy and ontology of multi-criteria decision support

Ontology is treated in the literature as a set of definitions of the terms of the area and the relationship between them [3]. It is also referred to as the specification conceptualization providing a description of the concepts and relationships that occur between them [2]. The use of ontologies as a solution supporting the choice of an MCDA method is designed to assist the user in selecting the right solution for a given decision situation described using specific criteria and parameters. The ontology should also provide detailed information about the various methods for multi-criteria decision support.

The first step in the construction of an ontology is to develop a taxonomy of criteria describing the MCDA methods. Identification and analysis were based on the analysis of 25 MCDA methods for creation of a set of criteria and sub-criteria characterizing the different solutions. A total set was formed comprising four main criteria (available binary relations, linear compensation effect, the type of aggregation and the type of preferential information) and 16 sub-criteria. This collection was also

TABLE I. CHARACTERISTICS OF SELECTED METHODS OF MCDA

| Method name | Essence of the method | Reference |
|---|---|---|
| AHP | The problem is formulated in a hierarchical form. Overall rank is based on the collective aggregation of partial marks obtained in a paired comparisons in matrix. | [19] |
| Additive weight method | Aggregation based on the inadditive function. Choosing the best alternative with the highest value of the global index obtained as the sum of the partial marks for all criteria. | [20] |
| EVAMIX | Ranking of alternatives on the basis of global dominance index (ordinal indices and cardinal dominance). | [21] |
| Electre I | The aim of the method is to determine the subset of variants containing the best alternative. The procedure is based on the construction of compliance and non-compliance tests, following the structure of the outranking graph. Preferences modeling is done using the true criteria. | [22] |
| Electre II | Extension of ELECTRE I method. The essential part is the use of two outranking relations :weak and strong. | [23] |
| Electre III | Ranking of variants based on outranking relationship. Modeling preferences with pseudo criteria and weights. | [24] |
| Electre IS | The method constitutes a development of ELECTRE I with additional modeling preferences based on pseudo criteria. | [25] |
| Electre IV | Ranking of variants based on the relationship with outranking pseudo criteria. It does not apply weights to the criteria. | [26] |
| Electre TRI | Sorting variants into categories based on outranking relationship. Modeling preferences with pseudo criteria. | [1] |
| MAUT | Ranking of variants based on the aggregation of sub-additive utility function form. | [27] |
| MAVT | Ranking of variants based on the aggregation of multiplicative utility function. | [27] |
| MELCHIOR | The extension of the ELECTRE IV method. Ordinal relationship validity of the criteria is added. | [28] |
| Maximin method | The aim of the method is to choose the strongest variant of the weakest. | [20] |
| Maximin fuzzy method | The aim of the method is to choose the strongest variant of the weakest. Evaluation of alternatives has the fuzzy form. | [29] |
| Methods of extracting the minimum and maximum values of the attribute | Methods reject criterion values successively above and below the predetermined value. | [20] |
| NAIADE | Application of fuzzy measures distances and of paired comparison. Calculation of preferences as in PROMETHEE. | [30] |
| ORESTE | Uses alternatives assessment and weighting of the criteria described only on an ordinal scale. | [31] |
| PROMETHEE I | The method constitutes a development of ELECTRE methods, but expanded the number of binary relations describing preferences to six. | [32] |
| PROMETHEE II | Extension of PROMETHEE I by the global results for all variants. | [32] |
| REGIME | The method is based on a pairwise comparisons matrix. Scale {1,0,1} is used. The values of the scale correspond to the domination, equivalence and dominance. | [33] |
| Additive fuzzy weight method | The method is based on a fuzzy version of the additive weight method where weight and evaluation are modeled as fuzzy numbers. | [34] |
| Fuzzy methods of extracting the minimum and maximum values of the attribute | Methods reject variants' criterion values successively above and below the predetermined value. Take the form of fuzzy evaluation. | [35] |
| SMART | Ranking of variants based on the aggregation of partial form of additive utility function. Global evaluations are calculated as a weighted average of partial evaluations. | [36] |
| TOPSIS | Choice of the best option is based on the multidimensional evaluation of the distance from the ideal and opposite solutions. | [20] |
| UTA | The model is additive utility function. Partial utilities are determined by using principles of ordinal regression. | [37] |

the basis for the construction of taxonomies of analyzed solutions. Table 2 presents a summary of the criteria and information about them by the various methods of multi-criteria decision support.

Based on a defined set of criteria, sub-criteria, and information about fulfilling them by selected solutions, the taxonomy created solutions for specific MCDA methods. This taxonomy is presented in the ontological form. This task requires distinguishing the concept on the basis of criteria and sub-criteria and establishing their hierarchy.

Figure 1 presents a graphical diagram of a set of criteria and sub-criteria of the main built taxonomy. The taxonomy provides a set of MCDA methods shown in Table 2, with a set of differentiating criteria and a network of taxonomic relationships between concepts (relations between the different classes of instances). Using this taxonomy, there is a possibility to select methods based on selected criteria.

A detailed analysis of taxonomic relationships reveals the full features of the different MCDA methods. This is the basis for a simple reusable but structured domain

TABLE II. TAXONOMY OF SELECTED METHODS OF MCDA

| Criterion | Available binary relations | | | | | Linear compensation effect | | | Type of aggregation | | | Type of preferential information | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method name** | **I** | **P** | **Q** | **R** | **S** | **No** | **Total** | **Partial** | **Single criterion** | **Outranking** | **Mixed** | **Deterministic** | **Cardinal** | **Non-deterministic** | **Ordinal** | **Fuzzy** |
| AHP | Y | Y | | | | | | Y | Y | | | Y | Y | Y | | |
| Additive weight method | Y | Y | | | | | Y | | Y | | | Y | Y | | | |
| EVAMIX | Y | Y | | | | | | Y | Y | | | Y | Y | | Y | |
| Electre I | | | | Y | Y | | | Y | | Y | | Y | Y | | Y | |
| Electre II | | | | Y | Y | | | Y | | Y | | Y | Y | | Y | |
| Electre III | | | | Y | Y | | | Y | | Y | | Y | Y | | Y | |
| Electre IS | | | | Y | Y | | | Y | | Y | | Y | Y | | Y | |
| Electre IV | | | | Y | Y | | | Y | | Y | | Y | Y | | Y | |
| Electre TRI | | | | Y | Y | | | Y | | Y | | Y | Y | | Y | |
| MAUT | Y | Y | | | | | | Y | Y | | | | Y | Y | | |
| MAVT | Y | Y | | | | | | Y | Y | | | Y | Y | | | |
| MELCHIOR | | | | Y | Y | | | Y | | Y | | Y | | | Y | |
| Maximin | Y | Y | | | | Y | | | Y | | | Y | Y | | Y | |
| Maximin fuzzy method | Y | Y | Y | | | Y | | | Y | | | Y | Y | Y | Y | Y |
| Methods of extracting the minimum and maximum values of the attribute | Y | Y | | | | Y | | | | | Y | Y | Y | | Y | |
| NAIADE | | | | Y | Y | | | Y | | Y | | Y | Y | Y | Y | Y |
| ORESTE | Y | Y | | Y | | | | Y | | Y | | Y | | | Y | |
| PROMETHEE I | Y | Y | | Y | | | | Y | | Y | | Y | Y | | Y | |
| PROMETHEE II | Y | Y | | | | | | Y | | Y | | Y | Y | | Y | |
| REGIME | | | | Y | Y | | | Y | | Y | | Y | | | Y | |
| Additive fuzzy weight method | Y | Y | Y | | | | Y | | Y | | | | Y | Y | Y | Y |
| Fuzzy methods of extracting the minimum and maximum values of the attribute | Y | Y | Y | | | Y | | | | | Y | | Y | Y | Y | Y |
| SMART | Y | Y | | | | | | Y | Y | | | Y | Y | | | |
| TOPSIS | Y | Y | | | | | Y | | Y | | | Y | Y | | | |
| UTA | Y | Y | | | | | | Y | Y | | | Y | | | Y | |

knowledge area. Based on preset criteria a user can receive detailed information about the satisfying method (methods) with its specific taxonomic characteristics. An example set of graphical results is shown in Figure 2, illustrating a method (here ELECTRE Tri) which met the criteria for the query: binary relations R and S, the partial effect of linear

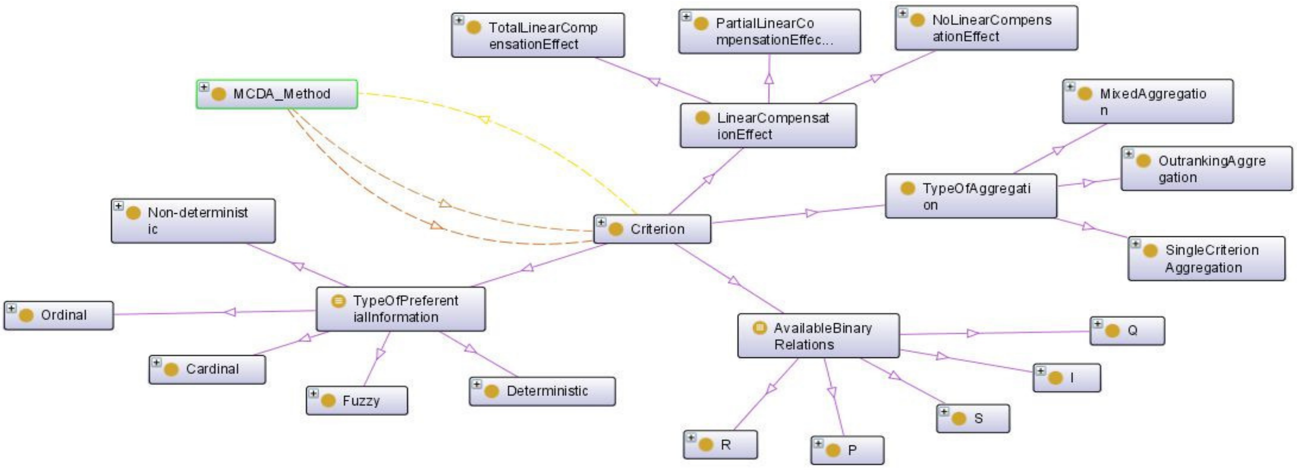


Fig 1. Elements of MCDA method taxonomy - set of criteria

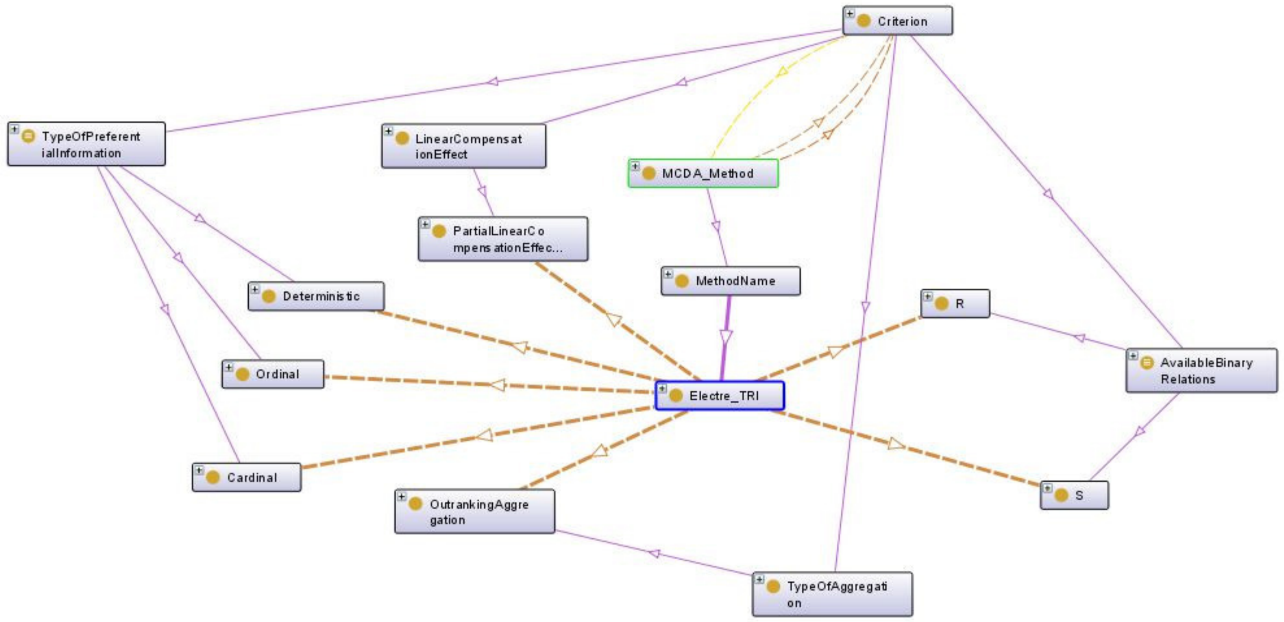


Fig 2. Graphical representation of competence query

compensation, aggregation using outranking relations, the type of preferential information – order.

## IV. Conclusion

This article discusses the problem of the construction of a taxonomy of MCDA methods and reference examples. The paper presents characteristics of identified MCDA methods. Based on the following analysis an identified set of criteria and sub-criteria characterizing the different solutions was presented. The results formed the basis for the construction of taxonomic relationships between the different MCDA methods and a complete taxonomy of MCDA methods and their use cases.

The results confirmed the possibility of the conceptualization of knowledge in the area of MCDA methods. The use of the proposed taxonomy supports the decision-maker's correct choice of multi-criteria method and allows for full domain knowledge about each one. It should be noted that the standard used for the construction of the ontology (OWL) ensures full compliance with current international semantic standards. This allows further use of the developed solution as well as its connection to other ontologies in various fields within the rapidly growing trend of knowledge engineering.

Further research should be supplemented by ontology of reference cases of the application of each method in various areas (management, logistics, environment, medicine, etc.) and reference publications characterizing the different MCDA methods. For ontology, additional multi-criteria methods can be attached, as well as criteria characterizing the various methods and the environmental context of their use. It makes possible the greater use of the adequacy of the reasoner and requests the use of various methods in decision problems using SWRL language rules.

## Acknowledgment

The work was partially supported by Fellowship co-Financed by European Union within European Social Fund, by European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE].

## References

[1] D. Bouyssou, B. Roy, Aide Multicritere a la decision: Methodes et Cas. Paris: Economica, 1993.

[2] W. Gliński, "Kwestie metodyczne projektowania ontologii w systemach informacyjnych," in Strategie informatyzacji i zarządzanie wiedzą, Z. Szyjewski, Ed. Warszawa:WNT, 2004, pp. 201-212.

[3] T.R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, vol. 5, no. 2, pp. 199–220, 1993.

[4] T. Trzaskalik, Metody wielokryterialne na polskim rynku finansowym. Warszawa: Polskie Wydawnictwo Ekonomiczne, 2006.

[5] http://protege.stanford.edu/

[6] J. Chai J., J.N.K. Liu, "An Ontology-driven Framework for Supporting Complex Decision Process," World Automation Congress (WAC), 2010.

[7] J. Wątróbski, J. Jankowski, Z. Piotrowski, "The Selection of Multicriteria Method Based on Unstructured Decision Problem Description," Lecture Notes in Artificial Intelligence, vol. 8733, pp. 454-465, 2014, http://dx.doi.org/10.1007/978-3-319-11289-3_46

[8] X.Y. Liao, E. Rocha Loures, O. Canciglieri, H. Panetto, "A Novel Approach for Ontological Representation of Analytic Hierarchy Process," Advanced Materials Research, vol. 988, pp. 675-68, 2014, http://dx.doi.org/10.4028/www.scientific.net/AMR.988.675

[9] E. Kornyshova, R. Deneckere, “Using an Ontology for Modeling Decision-Making Knowledge,” Frontiers in Artificial Intelligence and Applications, vol. 243, pp. 1553-1562, 2012.

[10] E. Kornyshova, R. Deneckere, “Decision-Making Ontology for Information System Engineering,” Lecture Notes in Computer Science, vol. 6412, pp. 104-117, 2010.

[11] P. Ziemba, M. Piwowarski, J. Jankowski, J. Wątróbski, “Method of Criteria Selection and Weights Calculation in the Process of Web Projects Evaluation,” Lecture Notes in Artificial Intelligence, vol. 8733, pp. 684-693, 2014, http://dx.doi.org/10.1007/978-3-319-11289-3_69

[12] B. Roy, Multicriteria Methodology for Decision Aiding. Dordrecht: Springer, 1996.

[13] M. Velasquez, P.T. Hester, “An Analysis of Multi-Criteria Decision Making Methods,” International Journal of Operations Research, vol. 10, no. 2, pp. 56-66, 2013.

[14] B.C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, U. Sattler, “OWL 2: The Next Step for OWL,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 6, no. 4, pp. 309-322, 2008.

[15] O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez, A. Lopez-Cima, “Building Legal Ontologies with METHONTOLOGY and WebODE,” Lecture Notes in Computer Science, vol. 3369, pp. 142-157, 2005.

[16] F. Baader, I. Horrocks, U. Sattler, “Description Logics,” in Handbook On Ontologies. Second Edition, S. Staab, R. Studer, Ed. Berlin: Springer, 2009, pp. 21-43.

[17] E. Della Valle, S. Ceri, “Querying the Semantic Web: SPARQL,” in Handbook of Semantic Web Technologies, J. Domingue, D. Fensel, J.A. Hendler, Ed. Berlin: Springer, 2011, pp. 299-363.

[18] B. Roy, “The Outranking Approach and the Foundations of Electre Methods,” Theory and Decision, vol. 31, no. 1, pp 49-73, 1991.

[19] T. Saaty, The Analytic Hierarchy Process. New York: McGraw-Hill, 1980.

[20] C.L. Hwang, K. Yoon, Multiple Attribute Decision Making: Methods and Applications. Berlin: Springer, 1981.

[21] H. Voogd, “Multicriteria evaluation with mixed qualitative and quantitative data,” Environment and Planning B, vol. 9, pp. 221–236, 1982.

[22] B. Roy, “Classement et choix en presence de points de vue multiples (la methode Electre),” Revue Francaise d'Informatique et de Recherche Operationnelle, vol. 8, pp. 57-75, 1968.

[23] B. Roy, P. Bertier, “La methode Electre II - une application au media planning,” in Operational Research OR'72, 1973, pp. 291-302.

[24] B. Roy, “Electre III: un algorithme de classement fonde sur une representation floue des preferences en presence de criteres multiples,” Cahiers du CERO, vol. 20, pp. 3-24, 1978.

[25] B. Roy, J.M. Skalka, “Electre IS - aspects methodologiques et guide d'utilisation,” Document du LAMSADE 30, 1984, 125 p.

[26] B. Roy, J.C. Hugonnard, “Ranking of suburban line extension projects on the Paris metro system by a multicriteria method,” Transportation Research, vol. 16A, pp. 301-312, 1982.

[27] R.L. Keeney, H. Raiffa, Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York: Wiley, 1976.

[28] J.P. Leclercq, “Propositions d'extensions de la notion de dominance en presence de relations d 'ordre sur les pseudo-criteres: Melchior. Revue BeIge de Recherche Operationnelle," de Statistique et d'Informatique, vol. 24, pp. 32-46, 1984.

[29] R.E. Bellman, L.A. Zadeh, “Decision-Making in a Fuzzy Environment,” Management Science, vol. 17B, no. 4, pp. 141-164, 1970.

[30] G. Munda, Multicriteria Evaluation in a Fuzzy Environment, Heidelberg: Physica-Verlag, 1995.

[31] M. Roubens, “Preference relations on actions and criteria in multicriteria decision making,” European Journal of Operational Research, vol. 10, pp.51-55, 1982.

[32] J.P. Brans, B. Mareschal, P. Vincke, “Promethee: A new family of outranking methods in multicriteria analysis,” in Operational Research OR'84, 1984, pp. 408-421.

[33] E. Hinlopen, P. Nijkamp, P. Rietveld, “Qualitative discrete multiple criteria choice models in regional planning,” Regional Science and Urban Economics, vol. 13, pp.77-102, 1983.

[34] D. Dubois, H. Prade, “The use of fuzzy numbers in decision analysis,” in Fuzzy Information and Decision Process, M.M. Gupta, E. Sanchez, Ed. Amsterdam, 1982, pp. 309-321.

[35] D. Dubois, H. Prade, C. Testemale, “Weighted fuzzy pattern matching,” Fuzzy Sets and Systems, vol. 28, no. 3, pp. 313-331, 1988.

[36] W. Edwards, J.R. Newman, Multiattribute Evaluation. Beverly Hills, CA:Sage, 1982.

[37] E. Jacquet-Lagrèze, J. Siskos, “Assessing a set of additive utility functions for multi-criteria decision making: The UTA method,” European Journal of Operational Research, vol. 10, pp. 151-164, 1982.

DOI: 10.15439/2015F294
ACSIS, Vol. 5

# Methodological Aspects of Decision Support System for the Location of Renewable Energy Sources

Jarosław Wątróbski
West Pomeranian University of Technology in Szczecin,
Żołnierska 49, 71-210 Szczecin, Poland
Email: jwatrobski@wi.zut.edu.pl

Paweł Ziemba
The Jacob of Paradyż University of Applied Sciences in Gorzów Wielkopolski, Chopina 52, 66-400 Gorzów Wielkopolski, Poland
Email: pziemba@pwsz.pl

Waldemar Wolski
University of Szczecin,
Mickiewicza 64, 71-101 Szczecin, Poland
Email: wwolski@uoo.univ.szczecin.pl

*Abstract*—**The aim of this article is to present the methodological framework for Decision Support System for the process of selecting the location of renewable energy sources. For this purpose, the paper presents the methodology and MCDA methods sequentially structuralizing and modeling the decision space and exploiting the developed model. The detailed area of research and verification of proposed approach are related to the problem of offshore wind farm localization. Proposed framework defines input information together with methodological background required for decision support processes.**

## I. Introduction

In recent years, there has been increasing interest in renewable energy sources (RES). The conditions for such a situation are related to the development of new technologies that look for ways of making the national economies of many countries more independent from conventional energy sources. Additionally, the continuing decline of natural energy resources, while their prices increase in the global market, has forced changes in macro and micro strategies to find new sources of energy. Renewable energy sources can be used almost anywhere in the world. The main problem is establishing the correct economical, technological, environmental and social justifications for the location and infrastructure construction using new technologies and resources. In terms of this study, the above issues were narrowed by focusing on the use of offshore wind farms. Decision related to selection of renewable energy sources should be taken carefully due to high impact on environment and community. Big number of factors and preferences create area for the use of Decision Support Systems [13]. Due to the complex set of factors, determining the multi-criteria profitability of this class of investments in wind energy for further study the MCDA methods were used as a methodological basement. This approach is also justified by the analysis of previous studies related to selecting and evaluating renewable energy installations.

## II. Renewable energy sources

According to the classification of the World Energy Council, wind energy is one of the various types of renewable energy sources [1]. This type of energy is used to produce electricity, utilizing wind farms composed of devices using wind turbines to convert wind energy into mechanical energy, which is then converted into electrical energy [3]. The effective operation of a wind farm is crucially influenced by the choice of location. This choice determines the efficiency of its operation and also has an impact on the costs, the benefits, and the environment. For obvious reasons, potential areas for the implementation of this type of investment must be assessed in terms of numerous factors: wind, water depth, and possible conflicts or formal legal exclusions (for example, if the potential area is located in a protected area [4]). In addition, there is a complex set of factors determining the success of this type of investment. Detailed analyses include determining the impact on the marine environment of the investment area and existing users of those areas. The fact that offshore wind farms generate some noise and have an impact on the landscape, which in turn affects the level of acceptance from local communities, must also be considered. That is why it is important to maintain an adequate distance from areas where there are marine mammals, fish species particularly sensitive to noise, birds, marine space, and fishing areas. Improperly located farms can be a source of negative environmental and social impacts. Another extremely important consideration is therefore the appropriate selection of criteria, which determines the accuracy of the entire decision-making process. The multitude of often conflicting criteria makes it methodologically formulated as a multi-criteria decision problem. Among the many available methods of research, the AHP method and Promethee II were chosen for this study; therefore, the article presents the concept of multicriteria model for a decision support system for choosing an investment location.

Review of the existing literature indicates the possibility of using multi-criteria decision methods for problems of selection for different types of economic activities. For

978-83-60810-66-8/$25.00©2015, IEEE

example, Forte [16], in the assessment of land management options, used the method of REGIME, NAIADE, AHP and FLAG. In [20], the region in the context of the use of various installations of renewable energy sources using the AHP method was assessed. Similar problems were also solved by Burton and Hubacek [14], who used the method of MACBETH and assessed the potential locations of such facilities. Eleftheriadou and others [19] used the PROMETHEE II method to evaluate the different variants in the foundation of wind turbines. Such a problem is presented also in the work of Cavallaro and Ciraolo [15], where the authors used the NAIADE method. This method was also proposed in the work of Gamboa and Munda [17] to assess the potential for wind farms in Spain. In contrast, the work in [19] was devoted to the evaluation of alternatives that consider different possibilities for the deployment of wind farms in one of the provinces in China, with the help of the AHP method. The work of Georgopoulou [18] is also relevant, as that research integrated different types of RES in Greece using the ELECTRE III method. The synthesis of research relating to the use of MCDA methods in problems of RES locations are given in Table 1.

## III. The use of multi-criteria decision support methods in chosing the location of an offshore wind farm

The procedure presented in this paper is in accordance with the guidelines contained in [8] and includes five successive stages:

- defining a set of decision variants,
- structuring the decision problem,
- setting priority vectors using the AHP method,
- exploitation of model and ranking of the decision variants using the Promethee II method,
- interpretation of the model with Plane Gaia and a sensitivity analysis of the set of decisions.

The first step in obtaining solutions is to define a set of decision variants for the problem of choosing the location of offshore wind farms in the Baltic Sea. Based on the map developed by the Maritime Institute in Gdansk [5] [6], including a list of potential sites for the location of offshore wind farms, 6 areas were selected. On this basis, a 6-piece set of variants for decision-making was established: W = {W1, W2, W3, W4, W5, W6}, where Wn is one of the locations on the Baltic Sea. The locations are shown in Figure 1.

Table I. Research relating to the use of MCDA methods in problems of RES locations

| No. | Assessment | Category | Main criteria | | | Method | Number of criteria | Type of data | Ref. |
|---|---|---|---|---|---|---|---|---|---|
| | | | Spatial | Environmental | Economical | | | | |
| 1 | Selecting site location | TS + SL | L | SA | IC, MC | AHP | 9 | qualitive | [25] |
| 2 | Selecting site location | SL | L | EI | IC, MC, EC | Macbeth | 8 | qualitive | [14] |
| 3 | Selecting site location | SL | L | SA | IC | Promethee II | 7 | qualitive | [19] |
| 4 | Selecting site location | SL | U, L | SA | IC, MC | Naiade | 9 | mixed | [15] |
| 5 | Selecting site location | SL | U, L, T | WN | IC | Naiade | 10 | mixed | [17] |
| 6 | Selecting site location | SL | U | SA | IC, MC | AHP | 12 | mixed | [19] |
| 7 | Selecting of renewable energy power plant technologies | TS | L | EI | IC, MC | Fuzzy DEA | 7 | qualitive | [26] |
| 8 | Selection of suitable electricity generation alternatives | TS | U, L | WN | IC | Promethee | 5 | mixed | [27] |
| 9 | Selecting of renewable energy power plant technologies | TS | U, L | SA | IC, MC | Electre III | 8 | quantative | [18] |
| 10 | Renewable energy sources project selection | SL | ED, U, L | WN, SA | IC, MC | Electre | 8 | mixed | [28] |
| 11 | Derive wind farm land suitability index and classification | SL | U | WN | IC | AHP | 10 | mixed | [29] |
| 12 | Derive wind farm land suitability index and classification | SL | L | EI, SA | IC, MC | Fuzzy AHP | 9 | mixed | [31] |
| 13 | Assessment of land management options | TS | U, L | WN | IC, MC | AHP | 12 | mixed | [20] |
| 14 | Define energy indicators used in the assessment of energy systems which meet sustainability criterion | TS | L | SA | IC, EC | weighted arithmetic mean | 5 | quantative | [30] |
| Abbreviations: TS – Technology selection; SL – Site location; L – Land use; U – Urban area; T – Tourism; ED – Ecological degradation; SA – Social acceptability; WN – Wildlife and natural reserves; EI – Environmental impact; IC – Investment cost; MC – Maintanace cost; EC – Electricity cost | | | | | | | | | |

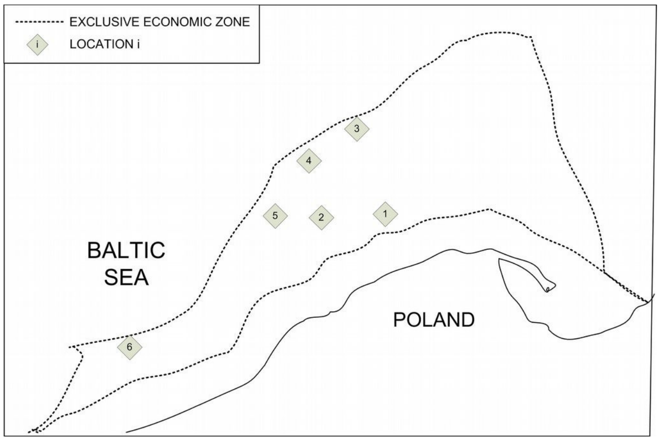


Fig 1. Map of potential locations

Selected locations have similar wind conditions, but vary in size, which determines the number of possible wind turbines that can be installed.

In the next step, the structure of the decision problem was identified. Analysis of work related to wind energy, including [2] [17] [3] [19] was used as a foundation to build a set of criteria: K = {K1, K2, K3}, where: K1 - spatial factors, K2 - economic factors, K3 - the social and environmental risks. As part of a set of criteria, three subsets were extracted: K1 = {K11, K12, K13, K14}, K2 = {K21, K22, K23}, K3 = {K31, K32, K33}, where: K11 - the average depth of the basin [m] , K12 - the distance from the shoreline [km], K13 - distance to connection NEN (National Energy Network) [km], K14 - the type of seabed; K21 - the cost of the investment, K22 - payback time, K23 - annual energy production; K31 - conflict with fisheries, K32 - the risk of navigation safety, and K33 - the impact on protected areas. The detailed specifications are shown in Table 2. The values of the criteria for each decision variant were determined using the reference literature contained in Table 2, and their resultant values are shown in Table 3.

Given the above results from individual investigations, the criteria were defined. Due to the fact that the different sets of criteria are related to different perspectives of looking at the same decision-making problem, it is difficult to clearly and objectively determine their weight. Consequently, it was assumed that the weight of each of the sets are mutually equal: K1 = K2 = K3. On the other hand, to determine the weights of the more specific criteria within each set, the AHP method was used [21]. This method was used because it is difficult to determine the absolute values of the weights of the individual criteria. However, the AHP method allows for the determination of the relative weights of criteria by

TABLE II. DETAILED SPECIFICATIONS OF EVALUATION CRITERIA

| Criteria | | | Type of criterion | Description of criterion | Reference literature |
|---|---|---|---|---|---|
| K1 | | Spatial factors | | | |
| | K11 | Average depth of the basin [m] | cost | Was determined in each location on the basis of the available bathymetry data. | [5] |
| | K12 | Distance from the shoreline [km] | cost | Distance from the coast is measured in a straight line. The increase in the distance from the edge causes a significant increase in the cost of building the farm (transport equipment, longer build time). | [5] |
| | K13 | Distance from the NEN connection [km] | cost | Distance of each farm to the nearest NEN port. | [7] |
| | K14 | Type of seabed | profit | Quality rating of the seabed for marine construction: rocky bottom - the best, grainy - very good, silty-sandy bottom - good, muddy - moderately good. | [6], [12] |
| K2 | | Economic factors | | | |
| | K21 | The investment cost [million PLN] | cost | Investment costs were estimated assuming a 7 MW turbine power. | [7], [9] |
| | K22 | Payback time [years] | cost | The calculation of payback periods makes possible to obtain annual profit. This value is determined using the unit price of electricity, the unit price of certificates of origin, and operating costs. Based on the calculated profit for the year, as well as taking into account the specified operating baskets payback time. | [9], [3] |
| | K23 | Annual energy production [GWh] | profit | Annual energy production is based on information concerning the annual average wind speed and wind turbine performance. | [7], [9], [1] |
| K3 | | Environmental and social risk | | | |
| | K31 | Conflict with fisheries | cost | Conflict of interest with the marine fisheries sector was estimated at a 9 points scale with 9 as the biggest conflict). | [6] |
| | K32 | Threat to navigation safety | cost | Based on the traffic map of water, determining how much influence it can have on a location for sailing in terms of possible dangers of ship collisions with wind farms. | [10] |
| | K33 | Influence on the protected areas | cost | The proximity of Natura 2000 protected areas of the potential offshore location. Determines the possible impact of investment on the protected areas on a nine-element scale, assuming that 1 means the least impact on protected areas, while 9 is the biggest. | [11] |

TABLE III. VALUES OF CRITERIA FOR DECISION VARIANTS

| **Criteria** | | | **Variants** | | | | | |
|---|---|---|---|---|---|---|---|---|
| K1 | Spatial factors | | **W1** | **W2** | **W3** | **W4** | **W5** | **W6** |
| | K11 | Average depth of the basin [m] | 40 | 31 | 29 | 62 | 51 | 35 |
| | K12 | Distance from the shoreline [km] | 34,7 | 45,6 | 86,3 | 77,1 | 63,1 | 44,9 |
| | K13 | Distance from the NEN connection [km] | 31 | 45 | 82 | 79 | 61 | 41 |
| | K14 | Type of seabed | very good | good | very good/good | moderately good | very good | good |
| K2 | Economic factors | | | | | | | |
| | K21 | The investment cost [million PLN] | 9040 | 9023 | 11231 | 10602 | 7870 | 7324 |
| | K22 | Payback time [years] | 9 | 10 | 11 | 15 | 14 | 12 |
| | K23 | Annual energy production [GWh] | 2803 | 2432 | 3132 | 3415 | 2132 | 1897 |
| K3 | Environmental and social risk | | | | | | | |
| | K31 | Conflict with fisheries | 8 | 5 | 9 | 4 | 5 | 6 |
| | K32 | Threat to navigation safety | 2 | 1 | 5 | 1 | 5 | 2 |
| | K33 | Influence on the protected areas | 2 | 8 | 1 | 1 | 4 | 1 |

mutually comparing their validity in a pairwise comparisons matrix, and then aggregating the results of these comparisons in the vector of preference. This is more effective for the decision-maker than determining the absolute weights [22], as one of the criteria for significance here is the point of reference in determining the validity of the other. The arrays of paired comparisons within the validity criteria sets of criteria K1 = {K11, K12, K13, K14}, {K2 = K21, K22, K23}, {K3 = K31, K32, K33} are contained in Tables 4, 5 and 6.

TABLE IV. PAIRWISE COMPARISONS OF VALIDITY OF THE "SPATIAL FACTORS" CRITERIA

| **CR=0,004** | **K11** | **K12** | **K13** | **K14** | **Weights (K1)** |
|---|---|---|---|---|---|
| **K11** | 1 | 1/2 | 1/2 | 2 | 0,189 |
| **K12** | 2 | 1 | 1 | 3 | 0,351 |
| **K13** | 2 | 1 | 1 | 3 | 0,351 |
| **K14** | 1/2 | 1/3 | 1/3 | 1 | 0,109 |

TABLE V. PAIRWISE COMPARISONS OF VALIDITY OF THE "ECONOMIC FACTORS" CRITERIA

| **CR=0,024** | **K21** | **K22** | **K23** | **Weights (K2)** |
|---|---|---|---|---|
| **K21** | 1 | 3 | 1/2 | 0,32 |
| **K22** | 1/3 | 1 | 1/4 | 0,122 |
| **K23** | 2 | 4 | 1 | 0,558 |

TABLE VI. PAIRWISE COMPARISONS OF VALIDITY OF THE "SOCIO-ENVIRONMENTAL HAZARDS" CRITERIA

| **CR=0,009** | **K31** | **K32** | **K33** | **Weights (K3)** |
|---|---|---|---|---|
| **K31** | 1 | 1/2 | 2 | 0,297 |
| **K32** | 2 | 1 | 3 | 0,54 |
| **K33** | 1/2 | 1/3 | 1 | 0,163 |

Tables 4, 5, 6 also show a right-hand value in each eigenvector matrix. This vector contains the aggregate weight of each criterion.

Assuming different weights of the individual sets of criteria (K1 = K2 = K3), the global weights of the criteria were determined. They are presented in Table 7.

TABLE VII. GLOBAL WEIGHTS OF CRITERIA

| **Set of criteria** | **Criterion** | **Global weight** |
|---|---|---|
| K1 | K11 | 0,063 |
| | K12 | 0,117 |
| | K13 | 0,117 |
| | K14 | 0,036 |
| K2 | K21 | 0,107 |
| | K22 | 0,041 |
| | K23 | 0,186 |
| K3 | K31 | 0,099 |
| | K32 | 0,180 |
| | K33 | 0,054 |

For ranking decision variants, the multi-criteria decision support Promethee II method was selected. It uses the outranking relation to the decision to choose the best alternative. This method uses the positive and negative flows of preferences specifying how much one alternative is greater than the other, and how much it is surpassed by other variants. Promethee II solves the sorting problem, and delivers the ranking of variants and indicate the best of them (in terms of Pareto evaluation). Using the Promethee II method, a decision maker can choose between six preference functions: simple criterion, the quasi criterion, the criterion with linear preference and preference level to the level of equivalence and preferences, the criterion with linear preference and indifference area, Gaussian criterion [23].

Preference functions based on the Promethee II method are shown in Figure 2.

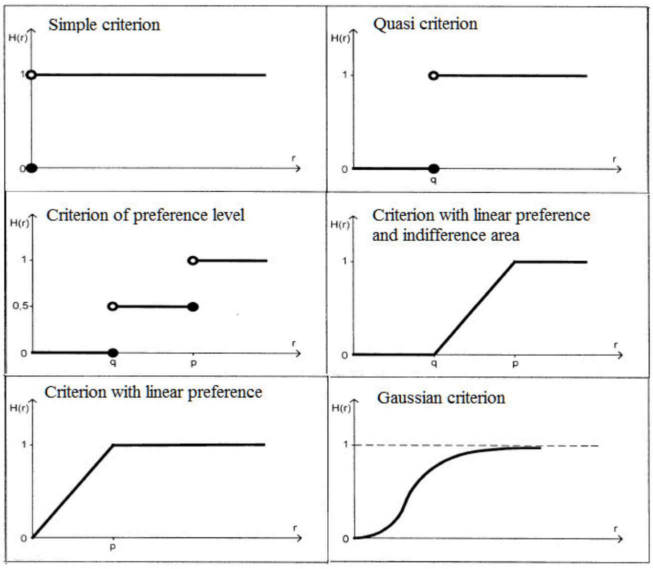

Simple criterion
Quasi criterion
Criterion of preference level
Criterion with linear preference and indifference area
Criterion with linear preference
Gaussian criterion
H(r)
1
0,5
0
q
p
r

Fig 2. Preference functions used in the Promethee II method

Rankings of alternatives were determined using ordinary criterion as a function of preferences. The results of evaluating the options and their positions in the ranking are presented in Table 8 and Figure 3. In line with the weights of criteria as the best option, the location of W1 was selected. When analyzing Fig. 3, the various variants of the specific form of solutions due to the applied criteria and their weights should be noted. Two solutions are ranked as the best (W1 and W2), two as good (W4 and W6) and two as bad (W5 and W3).

The presented ranking of decision variants is not final. The subjective nature of the vector of introduced priorities and the need to examine the strength of fixation in the ranking of the different alternatives provide a basis for decision-making executed in the next stage sequentially involving the analysis of sensitivity of this decision-making model.

TABLE VIII. PREFERENCE FLOWS IN ALTERNATIVES AND THEIR POSITION IN THE FINAL RANKING

| **Decision variant** | **Preference flow Φ+** | **Preference flow Φ-** | ***Preference flow Φ netto*** | ***Ranking*** |
|---|---|---|---|---|
| W1 | 0,5967 | 0,36 | *0,2367* | *1* |
| W2 | 0,5724 | 0,3645 | *0,2079* | *2* |
| W3 | 0,2906 | 0,6517 | *-0,3611* | *6* |
| W4 | 0,5297 | 0,4125 | *0,1172* | *3* |
| W5 | 0,3362 | 0,6007 | *-0,2645* | *5* |
| W6 | 0,4994 | 0,4356 | *0,0638* | *4* |

## IV. SENSITIVITY ANALYSIS OF THE RESULTING SOLUTION TO THE PROBLEM OF CHOOSING A WIND FARM LOCATION

The Promethee II method also enables a broad analysis of theresults , including a sensitivity analysis, and also provides an analytical tool, GAIA (ang. Geometrical Analysis for Interactive Assistance). GAIA aims to provide a complete graphical representation of the decision problem, so it allows the analysis of "goodness"-obtained solutions and provides directions for its possible improvement. The GAIA methodology information concerning the k-criteria of the decision problem in the k-dimensional space that is projected onto a sphere, so that a part of the information is lost, is presented, among others, including the Л vector, indicating the weights assigned to each criterion [24]. Alternatives are represented by points and criteria preferences are symbolized by vectors. If these vectors are oriented in the same direction, this means that they are not represented by the criteria in a similar way and it affects the global assessment of variants. The length of the vector indicates the strength of a given criterion to assess the global options. The closer a vector is to the end of a particular alternative, the more the vector supports this alternative, resulting in its ranking [23]. Examination of the results was conducted through a GAIA analysis representing individual criteria, which is shown in Figure 4.

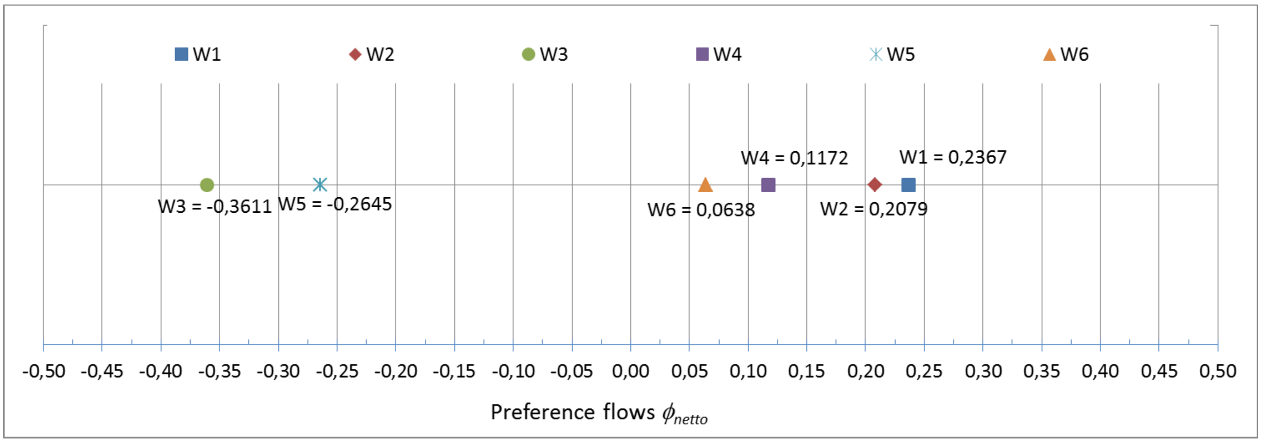


Fig 3. Netto preference flows in variants

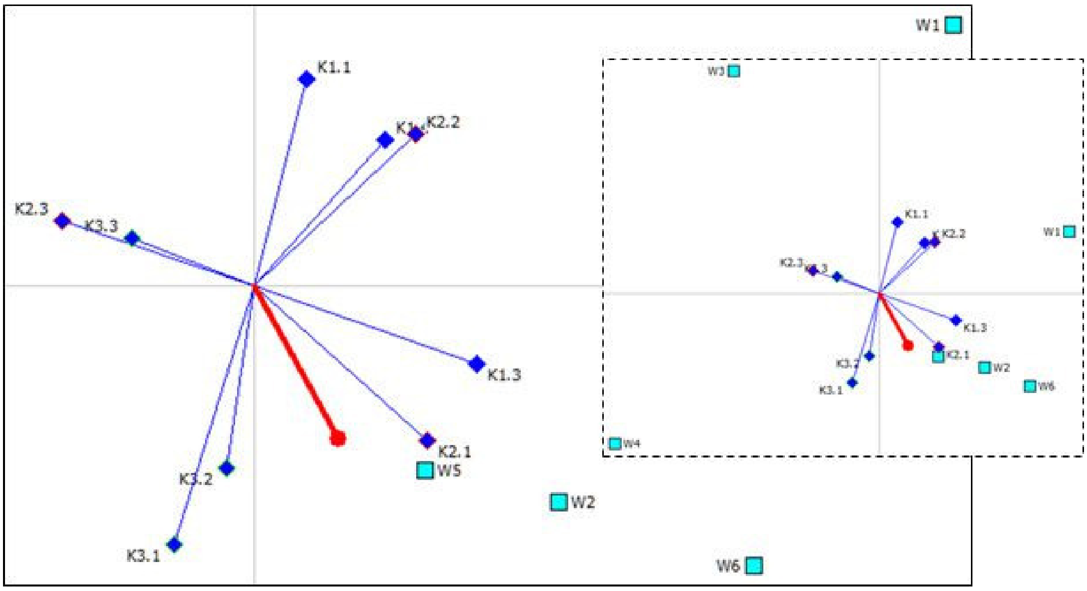


Fig 4. GAIA sphere representing individual criteria

When analyzing Fig. 4, it can be concluded that all the criteria have a similar effect on the acquired global assessment of variations, since the lengths of their vectors are all similar. However, there is a slight difference between the lengths of the K31 and K32 vectors, resulting from the difference of the weights given to these criteria. Shown here are the criteria coalitions and conflicts. For example, the criteria for K31 and K32 are compatible in the sense that, for most embodiments, a high value of K31 is associated with a high value of K32. However, these criteria are in conflict with K11; for variants with a high level of K31 and K32, the value of K11 is usually low. K14 and K22 support variant W1 because they are turned in his direction, and K13 and K21 criteria support the variants W1, W2, W5, W6, undermining the global assessment of the W3 and W4 variants. Due to the fact that the analysis of a multi-criteria decision problem using the GAIA strategy is projected onto a sphere, some information is lost. An example could be a vector compromise that seems to aim towards W5, while the highest rating is given to variants W1 and W2. The study also obtained solutions through the analysis of the GAIA sphere, which represented the sets of criteria. This sphere is shown in Figure 5.

Seeing The GAIA sphere from the perspective of criteria sets, it can be seen that the criteria related to spatial factors support the W1, W2 and W6 variants the most, since the end of the vector representing this collection lies closest to the specified options. The socio-environmental criteria strengthen the position of variants W2 and W4. The strength of the impact of these two sets of criteria for the solution of the decision problem is similar and is much larger than the impact of economic factors. Compromise vector reveals that variants of W1, W2, W4 and W6 are a much better choice than W3 and W5 variants. It should also be noted that the various sets criteria are not opposed to each other, and to some extent, the economic criteria are consistent with the environmental and social criteria.

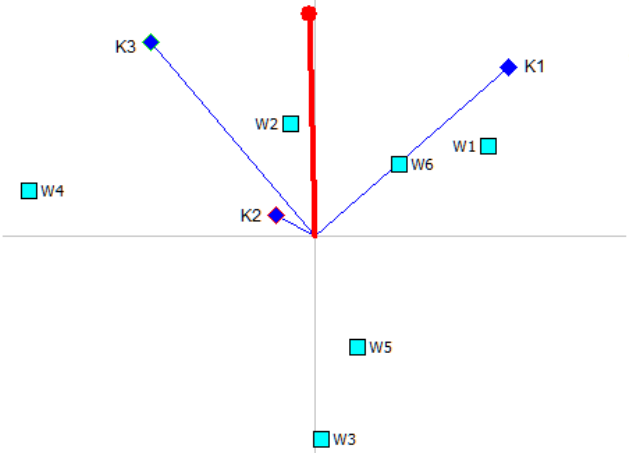


Fig 5. The GAIA sphere sets of criteria

The next step in the study was the analysis of the results to determine the stability of the sensitivity of the solution obtained, in terms of changing the weights of criteria. Due to the fact that, when looking at the problem from different perspectives (economic, localization, social and environmental), the relevance of the different sets of criteria may be seen as different, the weight of the sets of criteria was considered in this analysis. The results of the sensitivity analysis were included in Figure 6.

A sensitivity analysis shows that the solution is resistant to changes in sets of criteria weights in the range of 3%. If the environmental and social risks were much more important than other factors (weight K2> 45%), or if the weight of the economic factors was more than 75%, then the best option would be the W4 variant of the decision problem. The

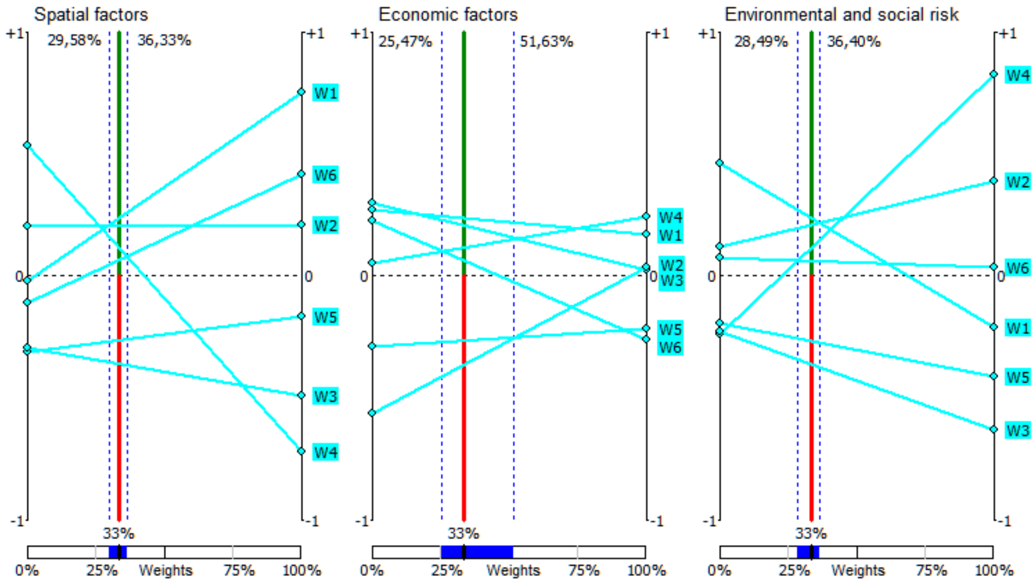


Fig 6. Sensitivity analysis of obtained solutions and weights

weight of a small range of socio-economic risks (36.4% - 45%), means that the best solution is also becoming a variant W2.

As part of the verification of the results the problem of ranking variants was determined by using a linear preference function instead of the true criterion. In this case, as a preference threshold for each of the criteria the standard deviation of variants with respect to a given criterion was computed. The netto preference flows and the rankings obtained in this study are included in Table 9.

Using the linear preference, there only one change of position of the resulting ranking; an exchange of the positions of the W4 and W6 variants was observed. As for the Pareto optimal solution, variant W1, using the linear preference, still received a higher than previously aggregated overall rank. The results of the sensitivity analysis for the solution obtained by using a linear preference function is shown in Figure 7. It shows the ranking of higher stability generated by applying the criterion with linear preference, as compared to the ranking obtained using the true criterion. Option W1 is the best at any weight and increase collections of spatial and economic criteria. For a group of socio-environmental criteria it remains in the first position in the ranking and increasing the importance of this group of up to nearly 12%. It can be concluded that the variant of the wind farm location W1 is a rational solution to the decision problem.

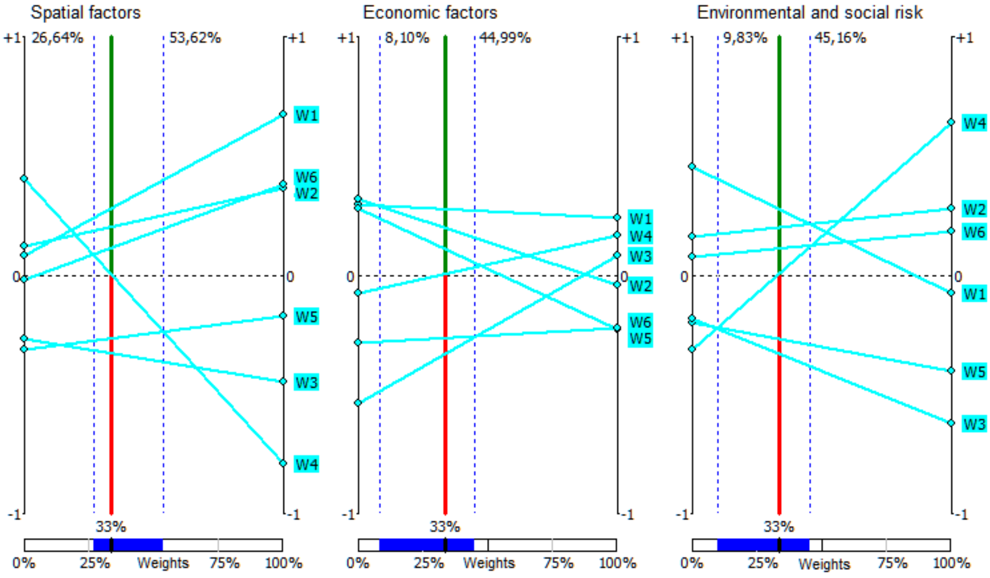


Fig 7. Sensitivity analysis of solutions obtained using the linear preference function

TABLE IX. PREFERENCE FLOWS IN ALTERNATIVES AND THEIR POSITION IN THE RANKINGS

| Preference criterion | Simple criterion | | Criterion with linear preference | |
|---|---|---|---|---|
| **Decision variant** | **Preference flow Φ netto** | **Ranking** | **Preference flow Φ netto** | **Ranking** |
| W1 | 0,2367 | 1 | 0,2773 | 1 |
| W2 | 0,2079 | 2 | 0,2001 | 2 |
| W3 | -0,3611 | 6 | -0,33 | 6 |
| W4 | 0,1172 | 3 | 0,0056 | 4 |
| W5 | -0,2645 | 5 | -0,2649 | 5 |
| W6 | 0,0638 | 4 | 0,1121 | 3 |

## V. SUMMARY

Design of Support system for selection of the location of renewable energy sources requires proper analytical methods. A practical application of multi-criteria decision support methods, as a maethodological basement for this system such as AHP and Promethee II was presented in this paper. It was verified for assessment of the potential location for offshore wind farm in Polish maritime areas. The possibility of linking these methods was presented and an advantage of the proposed approach was demonstrated. The pairwise comparisons, associated mainly with AHP, were used for the priorities to determine the vector which was used during the usage of Promethee II calculation method. The Promethee II method allowed for a comprehensive analysis of the resulting solution of the decision problem by using the GAIA sphere and sensitivity analysis. In addition, the solution obtained using the real criteria was confirmed by using a linear preference function. It should be noted that there were some limitations of the analysis conducted. Data representation is restricted to the field of crisp numbers, but each of the presented methods allows the use of fuzzy logic rules. In the future work it is assumed to extend the discussion about the use of an assessment methods based on, for example, the impact matrix implemented within the NAIADE method.

## REFERENCES

[1] M. Kaltschmitt, W. Streicher, A. Wiese, “Renewable Energy: Technology, Economics and Environment,”Springer Science & Business Media, 03.06.2007 – 596, http://dx.doi.org/10.1007/3-540-70949-5

[2] E. Eleftheriadou, D. Haralambopoulos, H. Polatidis, “A Multi-Criteria Approach to Siting Wind Farms in Lesvos, Greece,” http://www.srcosmos.gr

[3] R. Gasch and J. Twele (Eds.), “Wind Power Plants: Fundamentals, Design, Construction and Operation”, Springer, Berlin, Heidelberg 2012, http://dx.doi.org/10.1007/978-3-642-22938-1

[4] International Renewable Energy Agency (IRENA), 2015, www.irena.org

[5] http://www.transport.gov.pl/2-4e393a7f7308f.htm

[6] http://morskiefarmywiatrowe.pl/strefa-wiedzy/polska

[7] Z. Lubośny, “Wind Turbine Operation in Electric Power Systems. Advanced Modeling”, Springer-Verlag, Berlin Heidelberg New York, 2003.

[8] J. Wątróbski, J. Jankowski, Z. Piotrowski, “The Selection of Multicriteria Method Based on Unstructured Decision Problem Description,” Lecture Notes in Artificial Intelligence, vol. 8733, pp. 454-465, 2014, http://dx.doi.org/10.1007/978-3-319-11289-3_46

[9] R. Pesta, “Analiza opłacalności budowy farmy wiatrowej o mocy 40MW,” Rynek Energii, nr 1/2009, http://www.cire.pl/pliki/2/analizabudowyfarmy.pdf

[10] http://maps.helcom.fi

[11] http://www.ine-isd.org.pl

[12] L.J. Kaszubowski, R. Coufal, “Wstępny podział geologiczno-inżynierski dna polskiej części Morza Bałtyckiego,” Inżynieria Morska i Geotechnika, nr 3, 2010.

[13] P. Ziemba, M. Piwowarski, J. Jankowski, J. Wątróbski, “Method of Criteria Selection and Weights Calculation in the Process of Web Projects Evaluation,” Lecture Notes in Artificial Intelligence, vol. 8733, pp. 684-693, 2014, http://dx.doi.org/10.1007/978-3-319-11289-3_69

[14] J. Burton, K. Hubacek, “Is small beautiful? A multicriteria assessment of small-scale energy technology applications in local governments,” Energy Policy, vol. 35, pp. 6402–6412, 2007, http://dx.doi.org/10.1016/j.enpol.2007.08.002

[15] F. Cavallaro, L. Ciraolo, “A multicriteria approach to evaluate wind energy plants on an Italian Island,” Energy Policy, vol. 33, pp. 235–244, 2005.

[16] F. Forte, P. Nijkamp, F. Torrieri, “Shared Choices on Local Sustainability Projects: A Decision Support Framework,”, 2001, ftp://zappa.ubvu.vu.nl/20010024.pdf

[17] G. Gamboa, G. Munda, “The problem of windfarm location: A social multi-criteria evaluation framework,” Energy Policy, vol. 35, pp. 1564–1583, 2007, http://dx.doi.org/10.1016/j.enpol.2006.04.021

[18] E. Georgopoulou, D. Lalas, L. Papagiannakis, “A Multicriteria Decision Aid approach for energy planning problems: The case of renewable energy option,” European Journal of Operational Research, vol. 103, pp. 38-54, 1997.

[19] A.H.I. Lee, H.H. Chen, H. Kang, “Multi-criteria decision making on strategic selection of wind farms,” Renewable Energy, vol. 34, pp. 120–126, 2009, http://dx.doi.org/10.1016/j.renene.2008.04.013

[20] K. Nigim, N. Munier, J. Green, “Pre-feasibility MCDM tools to aid communities in prioritizing local diable renewable energy sources,” Renewable Energy, vol. 29, pp. 1775–1791, 2004.

[21] T.L. Saaty, “How to make a decision: The Analytic Hierarchy Process,” European Journal of Operational Research, vol. 48, pp. 9-26, 1990.

[22] T.L. Saaty, “Why the Magic Number Seven Plus or Minus Two,” Mathematical and Computer Modelling, vol. 38, pp. 233-244, 2003.

[23] J.P. Brans, B. Mareschal, “Promethee Methods,” in Multiple Criteria Decision Analysis, J. Figueira, S. Greco, M. Ehrgott, Ed. Boston: Springer, 2005, pp. 163-195.

[24] G. Janssens, M. Pangilinan, “Multiple Criteria Performance Analysis of Non-dominated Sets Obtained by Multi-objective Evolutionary Algorithms for Optimisation,” Artificial Intelligence Applications and Innovations, vol. 339, pp. 94-103, 2010, http://dx.doi.org/10.1007/978-3-642-16239-8_15

[25] T. Kaya, C. Kahraman, “Multicriteria renewable energy planning using an integrated fuzzy VIKOR & AHP methodology: The case of Istanbul,” Energy, vol. 35, pp. 2517-2527, 2010, http://dx.doi.org/10.1016/j.energy.2010.02.051

[26] M. Baysal, A. Sarucan, C. Kahraman, O. Engin, “The selection of renewable energy power plant technology using fuzzy data envelopment analysis,” in Proc. the World Congress on Engineering, vol. II, London, July 2011,.

[27] Y. Topcu, F. Ulengin, “Energy for the future: An integrated decision aid for the case of Turkey,” Energy, vol. 29, pp. 137-154, 2004.

[28] P. Haurant, P. Oberti, M. Muselli, “Multicriteria selection aiding related to photovoltaic plants on farming fields on Corsica island: A real case study using the ELECTRE outranking framework,” Energy Policy, vol. 39, no. 2, pp. 676-688, 2011, http://dx.doi.org/10.1016/j.enpol.2010.10.040

[29] S. Al-Yahyaia, Y. Charabi, A. Gastli, A. Al-Badi, “Wind farm land suitability indexing using multi-criteria analysis,” Renewable Energy, vol. 44, pp. 80-87, 2012, http://dx.doi.org/10.1016/j.renene.2012.01.004

[30] N. Afgan, M. Carvalho, “Multi-criteria assessment of new and renewable energy power plants,” Energy, vol. 27, pp. 739-755, 2002

[31] A. Sagbas, A. Mazmanoglu, “Use of multicriteria decision analysis to assess alternative wind power plants,” Journal of Engg. Research, vol. 2, no. 1, pp. 147-161, 2014.

DOI: 10.15439/2015F303
ACSIS, Vol. 5

# Integration of B2B system that supports the management of construction processes with ERP systems

Monika Łobaziewicz
OPTeam SA, Tajęcina 113,
36-002 Jasionka
Email: mlobaziewicz@opteam.pl

***Abstract*— B2B is not only a model of cooperation between enterprises. It may also be implemented in companies with a dispersed organizational structure to which construction companies belong. It turns out that B2B systems contribute to an effective implementation of the investments located even in distant geographical locations. The use of Internet technology contributes greatly to the successful exchange of information between the ERP system and other IT systems or applications. Thus, the management of complex processes is facilitated by automating the flow of information and data integration. In construction companies, each investment may be treated as a large and complex project. Therefore, the ERP systems are very useful there. Combined with Web Services with advanced B2B supporting construction processes, they are a powerful tool for managing a construction business.**

**The purpose of this article is to present the results of an analysis of the idea to integrate ERP systems with OPTIbud system based on research conducted under the second stage of the project *"The prototype of an innovative and technologically advanced OPTIbud B2B platform that supports the management of construction processes through the integration of data and information from multiple sources."***

## I. INTRODUCTION

IT TOOLS are used in construction companies all the time, before the beginning of the investment process, during its implementation, and after its completion. In the construction industry, the IT solutions support not only the management of business, financial, and human resources processes, but also the processes typical for the industry, such as tendering, cost estimating, designing, monitoring, standardization, project management, construction works, etc. In the construction industry, the most frequently used systems or applications facilitate design, such as Computer Aided Design (CAD) that helps the architects and designers to quickly and efficiently create a project that requires to make advanced calculations, taking into account investor requirements, different quality standards presented in technical and engineering norms, wide range of building materials and works.

From the effective construction projects management point of view, the system that support the investment process is equally important, in terms of the management and the flow of materials, goods, finances, and human resources. For this purpose, ERP systems integrated with sophisticated IT applications are often used.

Then, the key for the enterprise is whether the data on the progress of implementation of the investment are available in real time. This is important from the decision making point of view. In ERP systems there are mechanisms capable of simulating potential actions and analysing their effects, including analysing financial outcome. This allows, among other things, forecasting, planning, testing, and comparison of possible decision-making activities [1].

ERP systems have a beneficial effect on the functioning of companies. They allow to collect and process data, to make analyses necessary for the development of the whole company as well as for a specific construction project. The author's experience suggests that such a solution proves to be efficient in practice in any company with a complex organizational structure and, in particular, with dispersed structures or organizations that use the model of a remote work, characteristic for construction companies.

ERP systems with dedicated modules or trade applications operating on the basis of a central database connected by mechanisms integrating with other systems, simplify the communication between the construction company and its subcontractors and other business partners or between the headquarter of a construction company and the building site, contribute to cost reduction and the integration of construction processes and the improvement of the quality of services rendered to investors.

For the construction industry, the key aspect is the best combination of the company activities with the implementation of investment projects, often in geographically remote locations, with an intensive rotation and exchange of employees working on construction sites. The right project implementation also requires coordination with many subcontractors and the use of building equipment. Therefore, it turns out that the specific applications or systems supporting construction processes should be integrated with ERP systems. One of the possibilities is B2B system making the use of data stored in ERP system database, processes them, and exports the information back to the base of the system.

In the following section of this article, first the research methodology, then the most popular ERP systems together with the results of the market research are presented. The next sections present the outcome of an integration of ERP systems with other IT systems and applications, methods, architectures, and tools related to ERP systems.

978-83-60810-66-8/$25.00©2015, IEEE

Each section has some relevant literature and includes conclusions. Finally, the most important recommendations as a result of the study are presented.

## II. Research Methodology

The purpose of this article is to present the results of an analysis of the idea to connect B2B OPTIbud with ERP systems based on research conducted under the second stage of the project entitled "*The prototype of an innovative and technologically advanced OPTIbud B2B platform that supports the management of construction processes through the integration of data and information from multiple sources.*"[2][1]

To complete this purpose, the following methodology has been adopted. As part of the first stage of the project, the existing computer systems that are being used in the construction industry have been analysed together with indicating the directions of their development in terms of managing the construction processes and technology in which they were created. Then, at this stage, there was conducted in-depth research that covered 10 construction companies. The tool used in the research was a questionnaire consisting of 60 questions regarding the detail lements of a construction process, such as: budgeting, construction manager panel, equipment and transport base, tenders and quoting.

As a result of the research, the designed B2B OPTIbud system should handle the following internal and external processes:

- organization of tenders, including their recording and documents flow,
- budgeting of projects in relation to the most popular programs for cost estimation, such as: Winbud, Rodos, Zuzia, Norma,
- scheduling of construction works,
- supporting of the work of a construction manager with the application of data exchange with the ERP system in on-line way,
- settlement of work time of equipment, people as well as fuel usage and building materials,
- handling of the equipment and transport base.

The outcome of the research led to assumptions that are now the basis for the development of the functional architecture of the designed B2B OPTIbud system.

Comparative analysis between the functional expectations of construction companies regarding a dedicated system and the functionalities of computer systems available on the market allowed developing a list of functionality gaps that should be complimented and implemented in the B2B OPTIbud system so that it would create the competitive advantage compared to other solutions.

Based on the results of the first part, the second stage involved an analysis of ERP systems that are most frequently used by enterprises from the construction industry, paying special attention to their architecture, functionalities, mechanisms, tools, and capabilities, bearing in mind the aforementioned possibility to integrate the computer solutions with the developed B2B OPTIbud system.

## III. ERP Systems

### *A. The market of ERP systems in Poland and in the world*

According to Gartner, on the global market of ERP systems the leaders are [3]:SAP, Oracle, Sage ERP X3, Microsoft Dynamics AX, Microsoft Dynamics NAV, Infor ERP. In Poland, the market leaders of ERP systems are SAP, Oracle, Comarch, IFS, and BPSC.

According to IDC, in 2012 more than 40% of market share belonged to SAP [4]. The Comarch was at the second place (12.5%), followed by Oracle with 11.7% of the share. The next two places were occupied by IFS (4.8%) and BPSC (4.0%). Analysis of data from previous periods shows that the percentage share of the aforementioned companies has oscillated around these values for several years. Similarly, the position of leading suppliers of ERP systems on the Polish market has not been changed for several years.

Apart from the above mentioned leaders, on the domestic market, there are ERP products of such companies as Asseco Business Solutions, Unit4 Poland, and SIMPLE.

### *B. The types of implemented ERP systems and methods of their use*

Research conducted by Panorama Consulting Solutions [5] showed that the vast majority of enterprises in 2014 used traditional ERP systems (85%), and the remaining (15%) used ERP in a cloud or in SaaS formula. In comparison to 2013, there was a significant decrease in choosing cloud ERP software. In 2013, 26% of companies have chosen ERP software in the cloud or SaaS compared to 15% in 2014. Two most important reasons why companies did not implement cloud systems is a lack of knowledge about the technology (45%) and the risk and fear of breaches of data security (30%). The research also shows that the ERP cloud providers usually offer secure and reliable solutions that should be very important for managers in the process of choosing the software.

The research shows that smaller companies want ERP systems to used them in the most important areas of activity, while medium-sized companies expect integration of software with specialized applications in an integrated structure [6].

The important role in ERP systems is played by platforms and technologies for conducting electronic business, and mechanisms that optimize processes within the logistics supply chain [7]. One of the directions of the development of modern ERP systems is taking into account solutions important from the point of view of systems reserved for industries that were previously handled by independent, specialized software. This is primarily about construction

[1] The project, implemented by OPTeam SA is financed from INNOTECH program, HI-TECH path, carried out by The National Centre for Research and Development. The author is the chief scientific officer of the project.

companies, banks, insurance companies, or companies related to food processing on a small scale [8].

The designers of IT systems are facing new challenges involving the creation and implementation of solutions that support the integration of business processes, including processes specific to a given branch or industry, which would serve as an extension to the functionality of standard ERP management systems [9], [10]. For such solutions to be effective in practice, there must be a suitable mechanism designed for an access to multi-faceted data and information.

## IV. INTEGRATION OF SYSTEMS AND APPLICATIONS WITH ERP SYSTEMS

### *A. System integration*

Practice shows that the right combination of IT systems in an enterprise is, in many cases, necessary [11].

The system integration has helped to create standards that currently are tools that facilitate the programming works in this scope. High-level programming languages allow to launch created processes with their use programs without modifying various hardware and system platforms. The key role, thanks to its "portability," is played by XML, which is applicable in the domain of data structure description.

Correctly implemented system integration automatizes and improves the processes of the introduction and transfer of data. It is based on making it possible to enable an efficient cooperation of two completely different systems based on different technologies. The answer to these requirements is *Electronic Data Interchange (*EDI) and B2B solutions.

### *B. Integration of applications*

In turn, the integration of applications may be analysed in two ways. The first is based on an interaction between applications on the level of understanding the transmitted data bits. Over time, standards were developed which specified higher levels of communication model known as *Open System Interconnection Reference Model* (OSI RM). These included, respectively, the data-link layer (Ethernet) network (IP), transport (TCP, UDP), and higher layers. This enabled companies that created IT applications to gradually become independent of the elements of communication between the systems, so that they could focus on the semantics of data exchange between integrated elements. At the same time, standards for describing aspects of cooperation between applications were developed. An example of efforts to integrate applications and data within an enterprise that make the sharing of data possible between many heterogeneous computer systems and the integration of a distributed within the enterprise business processes into one coherent set is *Enterprise Application Integration* (EAI) [12]. On its basis, the *Service Oriented Architecture* (SOA) pattern was defined that describes integrated systems as independent services cooperating with each other [13],[14].

The attempts to use architectures based on SOA led to the creation of, among other things, the Distributed Component Object Model (DCOM) and Common Object Request Broker Architecture (CORBA). These technologies had a widespread applicability. However, the implementation of them with the use of SOA with technology proved to be impractical because of the dependence on a particular platform (DCOM) and high complexity (CORBA).

### *C. SOA model as an integration tool*

The dynamic development of the Internet has led to the need for integration both between applications inside organizations and between different companies by means of widely used communication protocols. In the construction industry, both types of integration exist. The author's experience leads to the conclusion that the diversity of computer system implementation environments and the pursuit of the reduction of costs associated with the implementation of solutions make the integration an expensive process.

SOA is a widely accepted programming, publishing, searching, and services launching standard. Its basic components include the following [15]:

- Service Provider - responsible for installing the service on the server, publishing its description in the registry and access control;
- Service requester - an application used by the user that discovers services in the registry and orders a demand for their start-up; and,
- Service Registry / Service Broker makes the provision of services possible.

SOA is the architecture for business applications created as a set of stand-alone components, arranged so as to provide services, operating according to certain criteria, and supporting the implementation of business processes [16]. From the end user's perspective, it is a set of services that support the implementation of business processes, and from the technological point of view, it is an infrastructure necessary to provide services.

SOA categorizes relationships between service providers and their customers represented by software components that implement complex business processes. It provides the reuse of software components, the encapsulation of functionality, the precise definition of interfaces, and the flexibility of applications created in a form of composing.

Commonly used element in the implementation of business services based on SOA are Web Services, which are understood as applications that make the business logic available outside the closed area of the network of organizations and which communicate through interfaces used in the Internet and communications protocols.

Web Services [17]:

- are autonomous applications available on the network through their URL (web address),
- have described interfaces and the way of using them by means of structures expressed in XML (Extensible Markup Language),
- are launched by other applications,

- perform heterogeneous operations ranging from simple responses to user requests to complex business processes,
- use standard communication protocols used on the Internet, and
- provide the possibility to launch and connect with other services providing new functionalities.

Web Services architecture refers to SOA. Therefore, it is based on the following principles:

- Message-oriented;
- Modularity of protocols - the use of the blocks that make up infrastructure protocols, which can then be used in almost any combination;
- The autonomy of services - allowing for independent construction, development, management, progress of versions, and securing endpoints;
- Transparency - control which aspects of endpoint are (or are not) seen by external services;
- Integration based on protocols.

To sum up, the process of integration is based on the interaction of individual systems and applications on the physical or functional level. SOA introduces a model of building computer systems based on communication between separated services. By creating dispersed applications, the individual functions of the program are distinct and separated into separate modules. Subsequently, these segments are merged to form a coherent and dispersed application.

Internet services based on Web Services enabled the effective implementation of the SOA concept in practice. Properly defined and simultaneously unambiguous specification of Web Services that includes, among other things, a description of the communication protocol SOAP and the language of interface description WSDL based on XML, mechanisms of semantic cataloguing and searching for providers caused that this standard quickly gained popularity and is now widely used in dispersed applications. At the same time, thanks to the support of Web Services by coding, language application services have been made available that are inconsistent with SOA through the establishment of appropriate adapters.

Web Services have caused people to move away from the paradigm of the classic integration of applications to the benefit of existing services in processes and their modelling, depending on changing needs.

## V. METHODS, ARCHITECTURE AND INTEGRATION TOOLS OF ERP SYSTEMS

### *A. SAP and MS Dynamics*

In the case of SAP, B2B system integration may be realized through the following: SAP .NET Connector, Microsoft BizTalk and Duet Enterprise.

The most beneficial tool is SAP.NET, because it may communicate with SAP system in two ways as:

- Client which queries SAP system (server) - client calls functions of SAP system and as a reply it receives results,
- Client playing the role of a server, receives calls from SAP while NET program adopts the role of RFC. (The SAP system treats the .NET application code as if it handled a different SAP system. Then, SAP sends a function call to .NET applications and receives results.)

Microsoft Dynamics operates in two versions, as Microsoft Dynamics AX and Microsoft Dynamics NAV. Regardless of which version we are dealing with, the system is delivered to the purchaser, together with its full source code and a set of tools for its development fully integrated with Visual Studio. Integrated Development Environment in MS Dynamics is MorphX that helps to create new modules, windows forms, reports, menus, queries, etc. The application is fully adapted to work with .NET platform. MS Dynamics allows for flexible development and software configuration according to the needs of an enterprise. This gives unlimited possibilities for adapting the system to any organization as well as integration with B2B systems.

The main advantages of the integration of MS Dynamics with other environments are the following:

- Using other technologies to develop the whole enterprise system is possible. Instead of code written in an internal X ++ language, it is possible to program functionalities in C # and have integration of Microsoft .NET with Microsoft Dynamics;
- MS Dynamics can be coordinated with external applications;
- There is the interaction functionalities and data exchange.

There are a few possibilities to integrate Microsoft Dynamics with other systems:

• The services and application integration platform - Application Integration Framework (AIF) and its services make business logic available to external business systems. MS Dynamics platform supports the integration of the AIF platform, using a services programming model. It exposes functionalities by means of services based on Windows Communication Foundation (WCF) from which code may use both MS Dynamics and external programs. WCF services in conjunction with the AIF provide a programming model, tools, and a complete infrastructure to integrate and exchange data using XML.

• .NET Business Connector allows external applications to access Microsoft Dynamics data and using its business logic.

• .NET Interop. allows to cooperate with .NET platform in both directions. It allows X ++ code to use the methods managed by the runtime environment of the .NET platform (Common Language Runtime) (CRL). It also provides solutions of proxy classes (C #, Visual Basic) for Class X ++ MS Dynamics generated in Visual Studio environment, from which C # code may directly benefit.

Business logic of MS Dynamics is made available through websites that use the latest Windows

Communication Foundation (WCF) rendered by the Application Object Server (AOS).

The configured pre-processor of requests (Request), AIF, captures all requests. MS Dynamics service runs the necessary business logic to process a request. A similarly configured postprocessor of response (Response) takes over the message of the post-processing response. Then it returns the response to the user.

The best way to integrate the ERP system with external applications is to use services that provide the business logic written in an internal X ++ language. In the Microsoft Dynamics environment, the programmer may build, modify, and publish services. Creating a service is based on defining, according to certain rules of X ++ class and website interface, and then their linking [18].

Making the services available through the Internet is made through Internet Information Services (IIS), which are equivalent to Web Services, and which redirect requests to the AOS regardless of whether they come from the Internet or intranet. AOS returns responses for the recipient of the service through IIS. The exchange configured to the use of Web Services is not queued, since it is supported synchronously.

AIF provides an extensible platform for the exchange of documents with external systems using XML technology. Both the synchronous and asynchronous transmission is possible. In the case of the synchronous exchange, the requests are paired with the answers, which means that the same connection it is used and that the demanding party does not continue work but just waits for a response. AIF immediately processes requests and sends responses. In asynchronous mode, the requests are queued, and the answers are sent with a time delay. The advantage is a more efficient processing of a large number of messages.

AIF supports the exchange in both directions: incoming and outgoing. Incoming exchange means that the external system, for example, B2B may send commands. On the other hand, outgoing is receiving data coming from the MS Dynamics system and sending them to the requesting user.

### *B. CDN XL*

The Comarch ERP XL system allows an interactive work with the user and also acts as an application server that provides external applications' built-in functionalities. This is provided by using a function mechanism, called the Application Programming Interface (API).

The external application may connect to CDN XL application server by means of API and use mechanisms available to the ERP system. This provides the openness of the system, making the integration possible with standard Microsoft Visual Studio programming tools and any external system. The CDN XL system provides an access to a dozen of API functions stored in the CDN_API.dll library supplied with the system.

The XL-API mechanism is used by IT companies who form their own solutions that cooperate with the Comarch ERP XL system. It allows for simplicity and security through the use of direct record in database and independence from the system and database platform, on which Comarch ERP XL runs.

The CDN XL system, while implementing the security policy, logs the changes to the database via the API functions. If an error occurs causing the termination of the application that calls out API functioning, there is a re-enactment of a database to the state before running the API function.

Approval of the transaction occurs upon the termination of entering the object or log- ut function (XL Logout).

XL API functionality with the development of CDN XL system is constantly being expanded. In order to ensure compatibility with previous API versions, there was introduced a versioning mechanism. It serves as a translator between API calls from external applications and structures used inside CDN XL system. For each version of the system, there is a corresponding version of the API libraries that allows the integration with it.

The programmer using the API functions defines its version based on which its structure is recognized and the transfer of the value to the internal CDN XL structure is made. In the case of non-compliance of structures, unused fields are filled with default values. The versioning mechanism is bi-directional. Data structures used by API also allow the return of data from CDN XL using the translation procedures.

## VI. RECOMMENDATIONS CONCERNING INTEGRATION

Based on the research of ERP systems and standards, integration tools, the following actions are recommended for the B2B OPTIbud platform:

- Implementation of mechanisms enabling a direct connection to the database of the ERP system,
- The use of Web Services available for ERP systems,
- Direct use of different ERP systems with a business logic characteristic in construction processes,
- Using an interface that is made available by the ERP system,
- The transfer of files with data using XML formats.

Moreover, there were identified the following activities characterising the impact of integration complexity requirements and architecture (Fig. 1).

The scale of the flow of various data and information is so large that it is difficult for companies to function on autonomous systems or computer applications. Integration complexity positively influences the optimization of business processes, enables a seamless cooperation of organizational units, and gives a possibility to operate on a one database to all participants of the processes work on the same actual data.

In many construction companies, heterogeneous applications coexist. The technological progress and market demands caused that construction companies carry out their operational activities through the Internet, and that is why the ERP software must cooperate with many applications. Considering the above, one of the key conditions for the

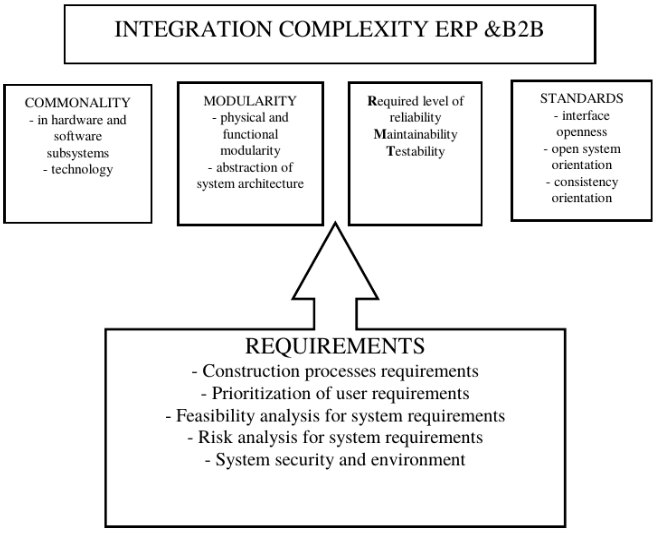

INTEGRATION COMPLEXITY ERP &B2B
COMMONALITY
- in hardware and software subsystems
- technology
MODULARITY
- physical and functional modularity
- abstraction of system architecture
Required level of reliability
Maintainability
Testability
STANDARDS
- interface openness
- open system orientation
- consistency orientation
REQUIREMENTS
- Construction processes requirements
- Prioritization of user requirements
- Feasibility analysis for system requirements
- Risk analysis for system requirements
- System security and environment

Fig. 1. Attributes having the impact on ERP system with B2B OPTIbud integration complexity

effectiveness of B2B OPTIbud is its integration with ERP systems end points. In this case, the interaction in the scope of data, information, and construction processes takes place on many levels.

The OPTIbud B2B system will be based on the philosophy of integrated systems that guarantees the flow of information between the individual modules in the ERP system, so the user would not have to repeatedly enter the same data in different functional areas of the system, but he will use the data already entered, transforming them into information necessary to make decisions and execution of construction processes. Thus, the B2B OPTIbud system, because of its autonomy may be able to be integrated with various ERP systems operating on the market. Thanks to built-in mechanisms, it would be able to communicate automatically with the database of the ERP computer system of a construction company.

## REFERENCES

[1] W.M. Grudzewski, I. K. Hejduk, *Methods of management systems design*. Warszawa: Difin, 2004.

[2] OPTeam SA, The prototype of an innovative and technologically advanced OPTIbud B2B platform that supports the management of construction processes through the integration of data and information from multiple sources, Project documentation.

[3] The market of ERP class systems ERP. http://issuu.com/benchmarkpl/docs/bbm_nr5, p.21 (Retrived 10.04.2015)

[4] http://biznes.benchmark.pl/artykul/rynek-systemow-klasy-erp-raport/strona/230 (Retrived 30.03.2015)

[5] Panorama Consulting Solutions „2014 ERP Report”, http://panorama-consulting.com/resource-center/2014-erp-report/ (Retrived 8.04.2015)

[6] http://www.erp-view.pl/erp/koniec_erp.html (Retrived 5.04.2015)

[7] B. Wieder, P. Booth, et. al, “The impact of ERP systems on firm and business process performance", Journal of Enterprise Information Management, Vol. 19 Iss: 1, 2006, pp.13–29, http://dx.doi.org/10.1108/17410390610636850.

[8] A. Chandrasekaran, G.Elias, R.Cloutier, R.Jain. ”Exploring the Impact of Systems Architecture and Systems Requirements on Systems Integration Complexity”, IEEE Systems Journal, vol. 2, no. 2, pp. 209-223, June 2008

[9] J. Ram, D. Corkindale, M.L. Wu, „Implementation critical success factors (CSFs) for ERP: Do they contribute to implementation success and post-implementation performance?”, International Journal of Production Economics, Volume 144, Issue 1, July 2013, pp. 157–174, doi:10.1016/j.ijpe.2013.01.032.

[10] J. May, G. Dhillon, M. Caldeira, “Defining value-based objectives for ERP systems planning”, Decision Support Systems, Volume 55, Issue 1, April 2013, pp. 98–109, doi:10.1016/j.dss.2012.12.036.

[11] S. Durvasula, M. Guttmann, A. Kumar, J. Lamb, T. Mitchell, B. Oral, Y. Pai, T. Sedlack, H. Dr Sharma, S. Ram Sundaresan, “SOA Practitioners’ Guide. Part 1: Why Services-Oriented Architecture?”, Oracle,2006.

[12] R.E Giachetti, “A Framework to review the Information Integration of the Enterprise”, International Journal of Production Research, vol. 42,no.6,2004, pp.1147–1166, doi: 10.1080/00207540310001622430.

[13] R. Jardim-Goncalvesa, A. Grilob, A. Steiger-Garcaoa, “Challenging the interoperability between computers in industry with MDA and SOA”, Computers in Industry, Volume 57, Issues 8–9, December 2006, pp. 679–689, doi:10.1016/j.compind.2006.04.013.

[14] T. R. Soomro, A. H. Awan, “Challenges and Future of Enterprise Application Integration”, International Journal of Computer Applications (0975 – 8887), vol. 42, no.7, March 2012.

[15] S. Gandhi, “A Service-Oriented Approach to B2B Integration Using Web Services”, White Paper Published for Dreamscape Media, http://www.featbooks.com/read-online/a-service-oriented-approach-to-b2b-integration-using-web (Retrived 20.03.2015)

[16] J. Łagowski, SOA – ideology, not technology, XV Conference PLOUG, Kościelisko, 2009, pp.180-192.

[17] M. Pielecka, „Integration of computer systems – interorganisational information exchange ”, Journal of Management and Finance, vol. 28, no. 4, part 1., 2013.

[18] http://msdn.microsoft.com/en-us/library/aa877498.aspx (Retrived 10.04.2015).

# 10th Conference on Information Systems Management

THIS event constitutes a forum for the exchange of ideas for practitioners and theorists working in the broad area of information systems management in organizations. The conference invites papers coming from two complimentary directions: management of information systems in an organization, and uses of information systems to empower managers. The conference is interested in all aspects of planning, organizing, resourcing, coordinating, controlling and leading the management function to ensure a smooth operation of information systems in an organization. Moreover, the papers that discuss the uses of information systems and information technology to automate or otherwise facilitate the management function are specifically welcome.

## Topics

The areas and topics of interest include, but are not limited to two groups:

- Management of Information Systems in an Organization:
  - Modern IT project management methods
  - User-oriented project management methods
  - Business Process Management in project management
  - Managing global systems
  - Influence of Enterprise Architecture on management
  - Effectiveness of information systems
  - Efficiency of information systems
  - Security of information systems
  - Privacy consideration of information systems
  - Mobile digital platforms for information systems management
  - Cloud computing for information systems management
- Uses of Information Systems to Empower Managers
  - Achieving alignment of business and information technology
  - Assessing business value of information systems
  - Risk factors in information systems projects
  - IT governance
  - Sourcing, selecting and delivering information systems
  - Planning and organizing information systems
  - Staffing information systems
  - Coordinating information systems
  - Controlling and monitoring information systems
  - Formation of business policies for information systems
  - Portfolio management,
  - CIO and information systems management roles

## Event Chairs

**Arogyaswami, Bernard,** Le Moyne University, USA

**Chmielarz, Witold,** University of Warsaw, Poland

**Karagiannis, Dimitris,** University of Vienna, Austria

**Kisielnicki, Jerzy,** University of Warsaw, Poland

**Ziemba, Ewa,** University of Economics in Katowice, Poland

## Program Committee

**Antosova, Maria,** Technical University of Košice

**Bialas, Andrzej,** Institute of Innovative Technologies EMAG, Poland

**Christozov, Dimitar,** American University in Bulgaria, Bulgaria

**Csikosova, Adriana,** The Technical University of Košice, Slovakia

**DeLorenzo, Gary,** California University of Pennsylvania, United States

**Dima, Ioan Constantin**

**Dudycz, Helena,** Wrocław University of Economics, Poland

**Espinosa, Susana de Juana,** University of Alicante, Spain

**Gafni, Ruti,** The Academic College Tel-Aviv-Yaffo, Israel

**Geri, Nitza,** The Open University of Israel, Israel

**Grabara, Janusz,** Czestochowa University of Technology, Poland

**Jelonek, Dorota,** Czestochowa University of Technology, Poland

**Kersten, Grzegorz,** Concordia University, Montreal, Poland

**Kobyliński, Andrzej,** Warsaw School of Economics, Poland

**Kohun, Frederick,** Robert Morris University, United States

**Korczak, Jerzy,** Wrocław University of Economics, Poland

**Lasek, Mirosława,** University of Warsaw, Poland

**Levy, Yair,** Nova Southeastern University - Graduate School of Computer and Information Sciences (GSCIS), United States

**Miliszewska, Iwona,** University of Canberra, Australia

**Modrak, Vladimir,** The Technical University of Košice, Slovakia

**Niedźwiedziński, Marian,** University of Lodz, Poland

**Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland

**Pańkowska, Małgorzata,** University of Economics in Katowice, Poland

**Pastuszak, Zbigniew,** Maria Curie-SKlodowska University, Poland

**Phusavat, Kongkiti,** Kasetsart University in Bangkok, Thailand

**Rizun, Nina,** Alfred Nobel University, Dnipropetrovs'k, Ukraine

**Rouibach, Kamel,** Kuwait University, Kuwait

**Ruzic-Dimitrijevic,** Ljijana, Higher Education Technical School of Professional Studies, Novi Sad, Serbia

**Schroeder, Marcin,** Akita International University, Japan

**Skovira, Robert,** Robert Morris University, United States

**Stanek, Stanisław,** The General Tadeusz Kościuszko Military Academy of Land Forces in Wrocław, Poland

**Świerczyńska-Kaczor, Urszula,** Jan Kochanowski University in Kielce, Poland

**Travica, Bob,** University of Manitoba, Canada

**Wielki, Janusz,** Opole University of Technology, Poland

DOI: 10.15439/2015F257
ACSIS, Vol. 5

# Exploring Determinants of Adoption and Higher Utilisation for E-Government: A Study from Business Sector Perspective in Saudi Arabia

Saleh Alghamdi
University of Sussex,
Informatics Department
Brighton, UK
Email: sa434@sussex.ac.uk

Natalia Beloff
University of Sussex,
Informatics Department,
Brighton, UK
Email: N.Beloff@sussex.ac.uk

***Abstract*—Providing e-Government services to business sector is a fundamental mission of governmental agencies in Saudi Arabia. The adoption of e-Government systems is less than satisfactory in many countries, particularly in developing countries.One pertinent, unanswered question is what are the key factors that influence the adoption and utilisation level of users from business sector. This paper utilised e-Government Adoption and Utilisation Model (EGAUM) proposed in our previous work in order to analyse determinants of higher level of e-Government adoption and usage. The study involved 48 participating business entities from two major cities in Saudi Arabia, Riyadh and Jeddah. The descriptive analysis is presented in this paper and the results indicated that all the proposed factors have degree of influence on the adoption and utilisation level. *Perceived Benefits, Functional Quality of Service, Previous Experience, Perceived Simplicity, Accessibility and Regulations & Policies* factors were found to be the significant factors that are most likely to influence the adoption and usage level of users from business sector.**

## I. Introduction

E-GOVERNMENT refers to the utilisation of various Information and Communication Technologies (ICTs) for facilitating communication between the government and the stakeholders; citizens, businesses and governmental agencies, providing effective, efficient and integrated e-Services that enhance the interaction between the government and the stakeholders through multiple and flexible channels that lead to an increased engagement. Adoption and utilization level is fundamental in terms of measuring the successfulness of the implemented e-Government systems. As governments develop e-Government systems to provide e-Services to stakeholders, the adoption and usage level is still low especially in developing countries [1][2][3]. Successful implementation of ICTs in government units and satisfactory usage level by all government stakeholders are the main goals of e-Government. Thus, analysing the significant factors that influence the adoption and utilisation of e-Government is becoming a necessity.

One of the main targeted stakeholders when providing e-Government services is business sector. The business sector, or what so called private sector in Saudi Arabia, is growing considerably in the recent years. The growth rate of the private sector is the highest between the three main sectors in Saudi Arabia, namely, governmental sector, private sector and oil sector [4]. Therefore, facilitating the communication and interaction between government agencies and business sector (business firms) is very important especially in the current advanced IT era. Many online e-Services have been provided to the private sector in the recent years in Saudi Arabia and there are more e-Services under development. Besides the need to increase the adoption and utilization level of the implemented e-Government systems, it is also crucial to understand the factors that can influence the adoption and usage of the new e-Services. The analysis of such factors will bring e-Government service provided to business sector to a more successful level and also will draw the path for developing new e-Government services.

In order to explore the determinant factors of high adoption and usage level of e-Government, we need to utilize a comprehensive framework. Thus, this paper used E-Government Adoption and Utilisation Model (EGAUM) see [5]. This model was developed based on critical evaluation of several common models and theories related to technology acceptance and usage including TAM and UTAUT, in conjunction with analysis of technology acceptance literature [5]. The rest of this paper will be divided as follows; the second section will explain the research methodology and the third section will present and discuss the research findings. The final section will provide a conclusion and our planned future work as this study is a part of our on-going research.

## II. Research Methodology

The study surveyed 53 business entities from different cities in Saudi Arabia, namely, Riyadh and Jeddah. Most of them were leader and large companies. Several medium and small business entities have also been involved in this study sample. The data has been collected in the period between August and October 2014. The business activities

978-83-60810-66-8/$25.00©2015, IEEE

of the participating companies were different in order to provide more comprehensive results. The sample included both business owners and employees who work in business firms and deal with government agencies with regard to their companies' transactions.

A 99-item questionnaire was distributed to the participants to respond to. The questionnaire comprised different forms of questions. Although the questionnaire was relatively long, it collected fundamental data that led to efficient and sufficient analysis. Moreover, a semi-structure short interview has been conducted with number of participants and it was optional. The questionnaire was distributed in person for several reasons including 1) ensuring high response rate, 2) clarifying questions when participant need, 3) conducting the short interview after completing the questionnaire. Users from business sector are usually busy. Therefore, meeting them in person at appropriate time to collect data is a useful method. The data has been collected from different cities in Saudi Arabia. Although this way consumed effort, cost and time, it enabled us to obtain more comprehensive and useful data.

Since the factors in the research model (EGAUM) are abstract [5], number of items was developed for each factor to measure its influence on the adoption of e-Government. Raubenheimer stated that number of items per factor is crucial specifically for scales with one factor which requires at least four items to be identified. However, most of the scales measure more than one factor. Scales that measure more than one factor, like this research scale, can be identified with as little as two items [6]. Table 1 shows the number of items for each factor in the scale developed for this study. The table exclude the personal factors since they are not measurable factors.

This research aims to study the influential factors of adopting and utilizing e-Government in the context of Saudi Arabia. Therefore, Saudi nationality was the targeted sample since the study involved some factors that aimed to be studied in the Saudi context such as socio-cultural, perceived trust and previous experience. The exclusion of non-Saudi nationalities is associated with the research goals of conducting a cross-sectional study of Saudi society. Moreover, the research instrument was designed to collect data relate to the Saudi context such as opinions about the available payment methods in Saudi Arabia, the Saudi post mail services and the influence of WASTA, which refers to the use of interpersonal relationships in transactions processing, on the adoption of e-Government. Moreover, incomplete questionnaires were also excluded since all the questionnaires' items need to be answered as they represent the research model constructs. Therefore, the total valid responses were from 48 business entities.

### *A. Reliability and validity*

Blunch stated that a research instrument is evaluated by its reliability. The reliability of a measuring instrument means its ability to provide identical results if it is repeated under identical conditions [7]. It concerns with a question "can we get the same results if we repeat the measurement test?". There are many different methods to determine the reliability of an instrument including Test-retest method, Split-Half method and Internal Consistency method. There were difficulties of using some of these methods in this research context. For example, repeating the test in different times with the same participants as in Test-retest method was difficult since the data collection phase in this study aimed to be conducted in different country, Saudi Arabia, whereas the researchers are in the United Kingdom. Moreover, Split-half method is appropriate for the long questionnaires that measure or test one construct and it is not suitable for instruments that measure several constructs as the current study [8]. Thus, this study used the internal consistency method to test the reliability with the use of Cronbach's Alpha. , see Table 1.

Internal consistency method determines how the items in a test relate to the other items. It is basically used to assess a survey items with a single construct to discover the consistency degree of the items that measure that construct. Cronbach's Alpha is the most common coefficient that can be calculated to evaluate the internal consistency of a test. It is applied to responses that have more than two options [9]. Therefore, it is applicable to this research instrument. Cronbach's Alpha reliability coefficient could range between 0 and 1 with value closer to 1 is considered to be high level of reliability. Acceptable level of reliability depends on the research purpose. The reliability of an instrument developed for research purposes is acceptable at 0.6 where in diagnostic research that the instrument is developed to make decision on individuals, such as psychological tests, need to be much higher [10].

In terms of validity, Bhattacharyya defined it as "the degree to which a test measures what it intends to measure" [11]. Validity of an instrument does not mean that the instrument is either valid or not but it is a matter of degree. The greater

TABLE 1: Internal consistency of the study instrument

| Measured variable | # of items | Cronbach's $\alpha$ |
|---|---|---|
| Perceived Benefits (PB) | 7 items | .825 |
| Socio-Cultural (SC) | 5 items | .686 |
| Awareness (AW) | 9 items | .822 |
| Functional Quality of Services (FQS) | 10 items | .800 |
| Previous Experience (PE) | 3 items | .601 |
| Perceived Simplicity (PS) | 6 items | .638 |
| Technical Quality of Service (TQS) | 5 items | .624 |
| Accessibility (ACC) | 4 items | .619 |
| Perceived Trust (PT) | 9 items | .792 |
| Regulations and Policies (RP) | 4 items | .899 |
| Intention to use e-Government (ITU) | 2 items | .659 |
| Perceived e-Readiness of e-Government (PER) | 2 items | .667 |

the evidence that an instrument produces valid results the greater the likelihood that we will get information that we need [12]. Validity cannot be calculated or measured directly, it is judged by the existing evidence [13]. There are several methods to determine or assess the validity of a research instrument and the most dominant ones include face validity, content validity, criterion validity and construct validity. This study utilised face validity and content validity methods to evaluate its validity degree.

Face Validity method provides useful information about the measure instrument and determines to what extent the instrument meets the intended purpose [12]. Test in which its purpose is clear, even for simple persons who have elementary knowledge, is judged to have high face validity. Where on the other hand, test in which its purpose is unclear would be judged to have low face validity [14]. Most of the items developed in this study were accurately face validated during the pilot study phase by the participants in accordance to the model factors' purposes, whereas some of them were not validated due to lack of understanding them clearly. Invalid items were either reworded if they were reported as unclear or deleted if they were reported as irrelevant. Furthermore, 5 academic members who are experts in interactive systems field have also reviewed the research instrument to have high content validity. Thus, it has been reviewed, tested and revised several times in order to have high validity.

## III. Research findings and discussion

This section provides an overview of the demographic characteristics of the respondents including age groups, education level, income, the use of the Internet and the use of e-Government systems. This is followed by a descriptive analysis for each factor proposed in EGAUM in order to explain their influence on the participants' adoption and utilisation of e-Government services provided to business sector.

### A. Respondents' demographic data

This section analyses demographic data obtained from the respondents (see Table 2). As per the questionnaire results, the average age group was ranging between 31 and 45 with males accounting for 89.5% of the participants and 10.4% were female. It is clear that the percentage of male participants was more than female participants. These percentages were expected for several reasons. Most of the employees who work in business sector (private sector) and use e-Government services in their jobs' activities are male employees. This is mainly because the jobs that involve dealing with government agencies to perform governmental transactions were almost exclusive to men before implementing e-Government in Saudi Arabia, therefore, they dominate these kinds of jobs due to their experience of dealing with the government services. Furthermore, it is difficult to collect data personally from female in Saudi Arabia, either business owners or employees, due to religious and cultural reasons.

TABLE 2: Demographic data

| Variables | % |
|---|---|
| **Participants gender** | |
| Male | 89.6 |
| Female | 10.4 |
| **Participants age in years** | |
| 18 - 30 | 20.8 |
| 31 - 45 | 64.6 |
| 46 - 60 | 12.5 |
| Over 60 | 2.1 |
| **Participants education level** | |
| Secondary school or less | 29.2 |
| Diploma | 20.8 |
| Bachelor degree | 41.7 |
| Master | 8.3 |
| **Proficiency of using computer** | |
| Average | 8.3 |
| Good | 41.7 |
| Excellent | 50 |
| **Internet usage rate** | |
| Several days a week | 6.3 |
| Several days a month | 2.1 |
| Everyday | 91.6 |
| **Participants relationship to the business** | |
| Owner | 27.1 |
| Representative | 18.8 |
| Employee | 54.2 |
| **Business age in years** | |
| 0 - 5 | 16.7 |
| 6 - 10 | 16.7 |
| 11 - 20 | 10.4 |
| Over 20 | 56.3 |
| **Number of employees in business** | |
| 0 - 10 | 18.8 |
| 11 - 50 | 8.3 |
| 51 - 250 | 4.2 |
| More than 250 | 68.8 |
| **Annual net profit in SAR (1 GBP $\approx$ 5.7 )SAR** | |
| 0 - 60000 | 10.4 |
| 61000 - 12000 | 4.2 |
| 121000 - 180000 | 2.1 |
| 181000 - 240000 | 4.2 |
| More than 240000 | 35.4 |
| Unknown | 43.8 |
| **Business field** | |
| Constructing and building | 14.6 |
| Restaurants | 2.1 |
| Food supply and grocery | 4.2 |
| Cars trade (sale and lease) | 12.5 |
| Communication | 6.3 |
| Health and medical supply | 14.6 |
| Other | 45.8 |

### B. Measures of central tendency

A measure of central tendency is a single value that attempts to describe a set of data. We define measures of central tendency to find some central values around which the data tend cluster or concentrate [15]. The mean is the most popular measure of central tendency and it can be computed by the following formula where $\bar{x}$ denotes to the mean, *n* is the number of values and *Xn* represents the values [16]:

$$\bar{x} = \frac{(X1 + X2 + .... + Xn)}{n}$$

The median is the middle score of the data set values that have been arranged in order of magnitude. Since

Likert response measures were used in this research, clarifications of the instrument scales will be presented in the following section. These clarifications helped us to choose the appropriate central tendency measures in the descriptive statistics and analysis.

### C. Likert-Type vs Likert Scale

Clason and Dormody identified the Likert-Type items as single questions that use some aspect of Likert responses. Whereas, Likert Scale is composed of multiple Likert-Type items that are combined into one single composite score/variable during the data analysis process. The composite score that combined multiple items represents a quantitative measure for the intended aspect that needs to be measured [17]. Likert-Type item categorized as an ordinal measurement scale and therefore, the recommended descriptive statistics are mode or median for central tendency and Inter-Quartile range (IQR) for variability. On the other hand, Likert Scale is categorized as interval measurement scale since it is represented as a composite score that is combined multiple items (either the sum or the mean of the combined items). Therefore, the recommended descriptive statistics are mean for central tendency and standard deviation (S.D.) for variability [18].

### D. Descriptive analysis of data

In this section, the collected data will be analysed in relation to EGAUM constructs [5]. The participants from business sector were asked to describe several characteristics including their attitude, behaviour and opinion towards adopting and utilizing the e-Government systems. Such characteristics were examined through number of statements using different Likert Scales. Likert Scale scores were also calculated in order to interpret the participants' responses.

***Perceived Benefits (PB)*** was measured with 7 items in order to indicate its influence on the respondents' adoption and utilization of e-Government systems. The participants were asked questions to determine their perception degree on several main benefits that they can gain from using e-Government services for their businesses. The perceived benefits that were measured included the speed of processing e-Transactions over the traditional ways, ensuring the equality of processing business entities' transactions and providing investment opportunities for the business sector. They also measured the respondents' perception of the ability of e-Government to save time, cost and effort. All the items in this factor were measured using 5-points Likert scale ranging from 'Strongly agree' to 'Strongly disagree'.

TABLE 3: The result of PB factor

| Factor | # of item | Mean | S.D. | result interpretation |
|---|---|---|---|---|
| PB | 7 | 1.37 | 0.49 | Very influential |

In respect to the perceived benefits influence, the results showed that the vast majority of the participants 97% believed that e-Government would enable them to perform their business transactions quicker than traditional ways. Moreover, around 95% of them believed that using e-Government services would save time, cost and effort. With regard to the benefits that they can gain for their businesses' entities, 89% of the respondents thought that using e-Government systems would minimize legal and regulatory violations that might occur from their businesses and 89% of them believed that using e-Government would help their businesses to comply with the governmental requirements.

The results also showed that the overwhelming majority (91%) of the respondents believed that using e-Government services would ensure equality in the dealing between governmental agencies and business entities. Furthermore, over 80% of the respondents thought that e-Government would increase the investment opportunities in business sector. It is clear from the results that the perception of the possible gained benefits from using e-Government was relatively high amongst users in the business sector. Table 3 presents the total score of this factor' items.

Perceived Benefits was found as a very influential factor with a total score of 1.37 i.e. PB has a strong impact on the adoption and utilization of users from business sector when using e-Government systems. The results suggest that the intention to use e-Government services is very likely to increase if users from business sector perceive the e-Services to be beneficial for their businesses. This indicates that to get business owners and business representatives to adopt and use e-Government services, these services must be genuinely useful for their business. They should be implemented efficiently and effectively in order to meet the needs of this category of stakeholders. These findings are in accordance with results reported in the literature, for example, [19] and [20] that have been conducted in the US. The findings also in accordance with a study conducted in Kuwait which is a close neighbour country to Saudi Arabia that have similar nature and culture. [21].

***Socio-Cultural (SC)*** was measured with 5 items using 5-points Likert scale. SC factor was concern with the influence of others, including colleagues and other business entities, on the intention to use e-Government. It also concerned with the influence of cultural norms, believes and behaviour patterns in Saudi Arabia on the adoption and use of e-Government services. For example, the respondents were asked to determine whether they feel that dealing with government agencies regarding their businesses' transactions should be real and tangible i.e. using papers not electronic means. The results showed that 70% of them disagreed with this item. Moreover, about three quarters of the respondents (75%) believed that e-Government would reduce the use of the interpersonal relationships (wasta) when processing business entities' transactions. This indicates a high perception about

TABLE 4: The result of SC factor

| Factor | # of item | Mean | S.D. | result interpretation |
|---|---|---|---|---|
| SC | 5 | 2.44 | 0.79 | Influential |

the ability of e-Government systems to eliminate or reduce this behaviour pattern. Furthermore, the results also showed that 81% of the respondents would be encouraged to use e-Government services for their businesses transactions if they know that other business entities use them. This indicates that the social influence is relatively high in terms of using e-Government services (see Table 4 for total score of SC).

The influence of some social aspects involved in this study, such as the influence of other, was not supported in other research such as a study conducted by AlAwadi and Morris [22]. However, the findings related to the influence of cultural aspects were in consistent with those reported in AlAwadi and Morris's study.

In respect to the ***Awareness (AW)*** factor, the respondents were asked two groups of questions with two different 5-points Likert scales. The first group (AW p1) was to evaluate the participants' awareness and measure the influence of the Awareness factor on their adoption and utilisation of e-Government systems. The 5-points Likert scale in this group was ranging from 'Strongly agree' to 'Strongly disagree'. The second group of questions was to measure the influence of several awareness methods that can affect the participants' awareness which in turn affect their willingness to accept and use e-Government services for their businesses' transactions. The 5-points scale used in this group was ranging from 'Very influential' to 'Very uninfluential'.

The results of the first group revealed that around 64% of the participants felt that they had good knowledge about e-Government potentials and services while 6% did not feel that. Moreover, 29% gave neutral or do not know responses. The reason of having slightly high percentage of 'Neutral or do not know' responses is likely because some participants did not know much about e-Government potentials and services due to lack of advertisements and awareness campaigns, so they did not feel that they had good knowledge about such benefits and services. Moreover, the results of the first group also showed that the majority of the respondents (89%) agreed with the item "Offering workshops and visual presentation about potentials and services provided for business sector would encourage me to attend to know more about e-Government". This indicates that such awareness campaigns would increase the adoption and usage level of the users from business sector. Table 5 presents the total score of AW p1. In the second part (AW p2), the most dominant advertising methods were introduced to the participants to determine the influence degree of such methods on their willingness to use e-Government services for their businesses' transactions. Fig 1 shows the introduced advertising methods and their influence on the participants' willingness to use e-Government services.

TABLE 5: The result of AW factor

| Factor | # of item | Mean | S.D. | result interpretation |
|---|---|---|---|---|
| AW p1 | 3 | 2.14 | 0.78 | Influential |
| AW p2 | 6 | 1.97 | 0.97 | Influential |

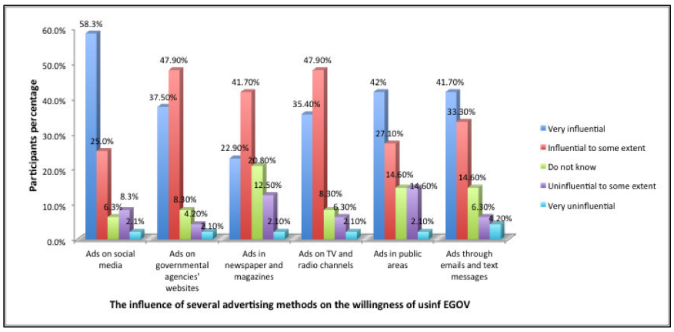


Fig. 1: The result of AW p2

It can be seen that 58% of the respondents reported that ads on social media significantly influence their willingness to use e-Government for their businesses' transactions. Albayari stated that social media revolutionized the business sector in Saudi Arabia since it creates sufficient ways to communicate with customers [23]. This means that most of business entities have accounts in social media networks and also new businesses are expected to follow the same trend of using social media. The result of this study and the status of social media spread in business sector in Saudi provide government with an excellent way of marketing e-Government services and increase the awareness about their benefits and potentials. Furthermore, 42% of the respondents reported that ads on public areas, such as billboards and banners, are very influential way that would increase their willingness to use e-Government services for their business. Advertising through emails and text messages comes third in terms of a very influential method with a percentage of 41%.

The higher three percentages of advertising methods that were reported as influential to some extent were ads on governmental agencies' websites (47.9%), ads on TV and radio channels (47.9%) and ads in newspaper and magazines (41.7%). It is likely that most of the participants are not usually exposed to such advertising; hence, they were reported as 'Influential to some extent' methods. Since the majority of the participating businesses were large companies, they usually linked to some governmental agencies electronically and thus, respondents did not see many ads on the agencies' websites. Being linked electronically means being viewed with the system directly not through website interfaces. Moreover, most of the respondents were employees in the participating businesses (70%) and it is likely that they do not usually have time to watch TV, listen to radio, read newspaper and magazines. Thus, they reported these methods as influential methods that affect their willingness to some extent. Participants who reported uninfluetial responses (either 'Very uninfluetial' or 'Uniflueti al to some extent') were relatively low in general.

The overall score for Awareness p2 was 1.97 indicating that the proposed advertising methods, which is part of the *Awareness* factor, influence to some extent the willingness to use e-Government systems for business entities' transactions. Furthermore, the overall findings of part 1 and 2 indicate the need for sufficient and efficient awareness and also indicate its influence on the adoption and usage level. These findings are similar to those revealed by AlAwadi [22], Alshihi [24] and Davies [25].

***Functional Quality of Service (FQS)*** has also been divided into two parts. The first part (FQS p1) comprised 4 items that measure the influence of the quality of service on the adoption and utilisation of e-Government systems but from functional aspect that concern with the e-Service itself. This part measures how the respondents feel about the general quality of provided e-Services through e-Government, how using e-Government would enhance the quality of their businesses' activities, what is the influence of important external factors (such as post and payment services) on the quality of services provided by e-Government. The items in this part were measured with 5-point Likert scale ranging from 'Strongly agree' to 'Strongly disagree'.

The results of the first part revealed that 48% of the respondents felt that e-Government services provided to business sector are high quality services. A slightly high proportion of respondents was neutral or did not know whether such e-Services are high quality and they represented 35%. When investigating neutral responses percentage, it was found that over 70% of them were employees. Employees who are responsible for performing e-Services in companies in Saudi Arabia usually are divided into different groups. Each group is responsible to deal with one governmental agency in terms of performing e-Transactions. Therefore, it is very likely that they cannot judge the e-Government services in general as they might be dealing with certain amount of e-Transactions. Moreover, The vast majority of the respondents (85%) agreed on that using e-Government systems would enhance the quality of their business activities.

The influence of significant external factors that are related to the functional quality of service such as the Saudi post mail services and payment options were also measured. Around 43% of the respondents agreed with that the Saudi post mail services are fast and reliable whereas 27% disagreed and 29% responded with 'Neutral or do not know' answers. Although the majority showed an agreement, the other proportions (disagree and neutral proportions) were relatively high. Those two percentages can be interpreted as that the Saudi post mail provides insufficient services for correspondence between business members and government agencies. Furthermore, it is likely that many business entities do not benefit from the Saudi post mail due to its limited services. With regards to the payment options, over three quarters (79%) of the respondents agreed with the sufficiency of SADAD system, which is the only payment method for governmental transactions in Saudi Arabia, whereas only 10% disagreed. SADAD is an intermediary system between banks and billers where it allows customers to pay for services. Although there are not many options for online payments methods in Saudi Arabia, the results showed high satisfaction level of the current payment system. Table 6 shows the total score of FQS p1.

TABLE 6: The result of FQS factor

| Factor | # of item | Mean | S.D. | result interpretation |
|---|---|---|---|---|
| FQS p1 | 4 | 2.21 | 0.66 | Influential |
| FQS p2 | 6 | 1.74 | 0.95 | Very influential |

The second part of *Functional Quality of Service* factor (FQS p2) comprised 6 items. They measured to what extent certain features, that are related to the functional quality aspect, are important to the respondents when using e-Government services for their businesses' e-Transactions. The measured features include the ability to track transactions online, the ability to view the history of the performed e-Transactions, the ability to appeal electronically, providing effective customers services and the ability to rate the quality of the provided e-Services.

The results of FQS p2 revealed that the ability to track governmental transactions online was very important to three quarters of the respondents (75%). The same percentage also reported that the ability to appeal online in case of transaction rejection was highly important to them. The vast majority of the respondents (80%) agreed on that the ability to view the history of all performed transactions online is of high importance degree. Over 70% of the respondents considered the feature of communicating with them regarding to the status of their businesses' transactions as of high importance. One of the functional features that is related to the quality of the provided services is providing customer care services. The participants were asked to determine the importance degree of providing such services that are dedicated only to serve customers from business sector with using e-Government services. The majority of participants (72%) reported high importance degree of the existence of such services.

The high importance degrees that were reported to most of the proposed functional features revealed the positive influence of providing these features when implementing e-Government services for business sector. Such functional features reflect high quality of the implemented e-Services and it is very likely that they would increase the adoption

TABLE 7: The result of PE factor

| Item | Median | IQR | result interpretation |
|---|---|---|---|
| PE1 | 2 | 1, 2 | Influential |
| PE2 | 1 | 1, 1 | Very influential |
| PE3 | 1 | 1, 3 | Very influential |

TABLE 8: The result of PS factor

| Factor | # of item | Mean | S.D. | result interpretation |
|---|---|---|---|---|
| PS p1 | 3 | 1.96 | 0.63 | Influential |
| PS p2 | 3 | 1.78 | 0.89 | Very influential |

and usage level (see Table 6 for total score of FQS).

***Previous Experience (PE)*** factor comprised three items,namely, PE1, PE2 and PE3 that measured the influence of Previous Experience on the intention to adopt and use e-Government services provided to business sector. Each of which had different measurement scale with three possible answers. The answer options of this factor's items were narrowed to three because we would like to gain more accurate results for exploring the impact of previous experience.

The respondents were asked to rate their previous experience of using e-Government services for their businesses' transactions. The results showed that 43% were fully satisfied, 54% were satisfied to some extent and only 2% were not satisfied. Although the percentages of fully satisfied and unsatisfied respondents were good indication of efficiency of the current e-Government services, the high percentage of respondents who were satisfied to some extent was, on the other hand, questionable. Therefore, information about the reasons that made the previous experience of some respondents not fully satisfied has also been collected. The reported reasons included difficulty of using e-Government service, lack of clarity of the requirements and failure to obtain the expected results of using such e-Services.

The participants were also asked to indicate how the previous experience of using e-Government for their business would affect their future use in item PE2. Three quarters of them (76%) stated that it will encourage them to use such systems for all their businesses' transactions whereas 8% stated that it made them hesitant to use such systems again. The high percentage of the encouraged respondents was not necessary an indication of high satisfactory level. It is likely that the need for online governmental service was the main reason of this high percentage of encouraged respondents. The slightly high proportion of respondents who reported that there will be no effect of the previous experience (15%) support this assumption.

Another item that measure the influence of the previous experiences of using online services was item PE3. It asked the participants who have used non-governmental online transactions for their businesses whether such experience would affect their willingness to use online governmental services. The results showed that over half of the respondents (57%) answered 'Yes, positively', only 3% answered 'Yes, negatively' and around 39% answered 'No effect'. This indicates that using online transactions and services (not governmental ones such as online banking and online purchasing) is very likely to influence the willingness of using e-Government services and transactions. All the total scores of PE items are presented in table 7.

***Perceived Simplicity (PS)*** factor comprised two groups of items and each group had different measure scale. The first group (PS p1) involved three items and they measure to what extent the perceived simplicity would influence the participants' adoption and utilisation of e-Government systems. The results of this part revealed that the overwhelming majority of the respondents (85%) believed that e-Government services provided to business sector are easy to use. However, over half of them (68%) were feeling that using such e-Services requires high level of concentration. This is likely because the lack of instant communication between customers and e-Services providers. In traditional ways, customers can instantly interact with the governmental employees who serve or conduct governmental transactions and thus, governmental employees can inform the customers if there is any mistakes or incomplete information and documents that make the performed transaction unaccepted or rejected. Where in online environment, such communication is limited especially when there is no sufficient online communication method such as online chat with representatives. In this case, customers need more concentration when performing e-Services to avoid rejecting them due to lack of or mistakes in the provided information and documents. Furthermore, the results also showed that if using e-Government services were difficult and complex, 79% of the respondents would hesitate to use them again. The overall score of 1.96 (see Table 8) indicated that PS factor is likely influence the adoption and utilisation level. In other word, the difficulty and complexity of using e-Government services would negatively influence the willingness of users from business sector to use such e-Services.

In PS p2, the respondents were asked to determine the importance degree of proposed important information related to the simplicity of use. The proposed information involved detailed steps on how to perform e-Services online (item PS4), texts or pictures that clarify the required information and documents (item PS5) and explanation of processing steps of the intended e-Service (item PS6). The proposed information is strongly related to the ease of use and thus, the participants were asked to specify to what extent such information is important to them. Measuring the importance

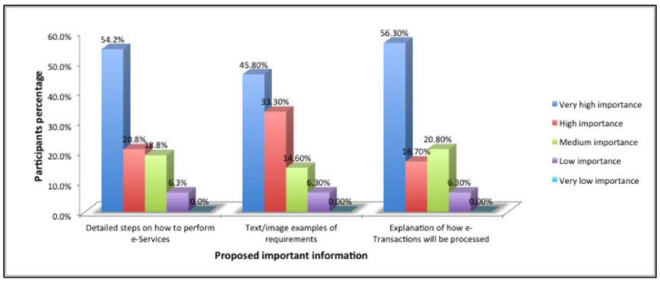


Fig. 2: The result of PS p2

of such information will give an indication on whether the simplicity would influence the adoption and utilisation of e-Services. The results of PS p2 are presented in Fig 2.

It is clear from the figure that almost half of the respondents believed that providing such information is very important to them and they represented 54% for item PS4, 45% for item PS5 and 56% for item PS6. Approximately 33% reported ’High importance’ for providing text\image examples of the requirements whereas 16% reported the same degree of importance to providing explanation of the processing steps and 20% also reported the same degree of importance to providing detailed steps on how to perform e-Services. Generally, clarifying the required information and documents with texts and pictures examples was the most important information as 79% of the respondents indicated ’Very high’ to ’High’ importance level. Remarkably, none of the participants believed that such information has very low importance degree. This indicated that all the proposed information had relatively high importance degree to the respondents (see Table 8 for the total score of PS p2) and therefore, it is very likely that providing such information would positively influence their adoption and utilisation of e-Government services.

E-Government services need to be straightforward and simple to use in order to enable all potential users to benefit from such e-Services. This simplicity is emphasised when it comes to providing e-Services to business sector where large number of e-Services and e-Transactions are needed. Other studies such as Carter and Belanger [26] and Phang et al. [27] also found that simplicity and ease of use is a significant factor that influence the intention to use e-Government.

***Technical Quality of Service*** factor was measured with five items and they were divided into two parts measured with 5-poits Likert scale (agreement scale and importance degree scale). The results of the first part revealed that 68% agreed with that the technical errors such as links not working and server errors would affect negatively their willingness to use such e-Services. This high agreement is expected since the technical aspects in e-Government systems are crucial in terms of quality of service especially when they are visible to the users. Moreover, the majority of the respondents (70%) agreed with that the interface design and layout of e-Government websites would influence their willingness to use e-Government for their business. It has been found from the results of this part (see Table 9) that the technical aspects are very important in terms of quality of service perception and thus, providing e-Services with high technical quality is fundamental. A high quality of the technical side including the design, structure and layout of e-Government portals is needed for the successful adoption and utilization [28].

TABLE 9: The result of TQS factor

| Factor | # of item | Mean | S.D. | result interpretation |
|---|---|---|---|---|
| TQS p1 | 2 | 2.40 | 0.96 | Influential |
| TQS p2 | 3 | 1.84 | 0.87 | Influential |

TQS part 2 was concern with measuring the importance degree of providing important information that is related to the technical side of e-Services (see Fig 3). Such information is important in terms of quality of service perception but from the technical side. The majority of the respondents reported that providing the proposed information is very important to them when using e-Services for their businesses. Half of the participants (50%) reported ’Very high importance’ for providing information about the expected time to complete their businesses’ e-Transactions, 45% reported the same importance degree for providing the last update time of e-Services websites and 47% also reported the same level of importance for providing the last update time of procedures and requirements that their businesses have to comply with. The overall score of 1.84 that is presented in Table 9 indicates that the proposed information had high importance degree for the users from business sector and thus, providing such information would likely increase their usage.

Furthermore, the results of TQS (part 1 and 2) indicates that this factor would influence the adoption and utilisation of users from business sector. They showed that reducing technical errors, enhancing interfaces design and providing information that reflect high technical quality of service would definitely increase the adoption and usage level of

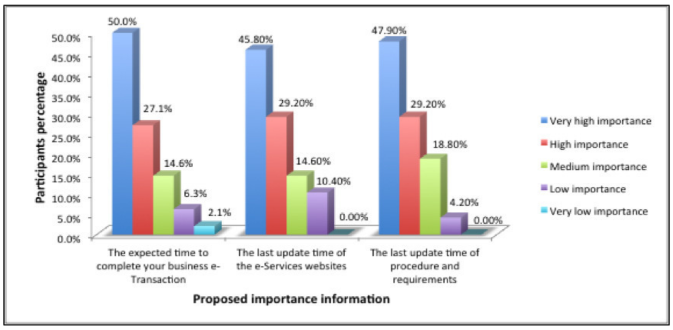


Fig. 3: The result of TQS p2

TABLE 10: The result of ACC1 factor

| Item | Median | IQR | result interpretation |
|---|---|---|---|
| ACC1 | 2 | 1, 2.5 | Influential |

TABLE 11: The result of ACC p2

| Factor | # of item | Mean | S.D. | result interpretation |
|---|---|---|---|---|
| ACC p2 | 3 | 1.71 | 0.85 | Very influential |

e-Government services provided to business sector. Therefore, paying more attention on technical aspects that appear directly to the users is important.

In terms of ***Accessibility (ACC)*** factor, three quarters of the respondents (75%) thought that the existence of authorised offices to help them to access and use e-Government services is a good idea. The authorised offices meant to be official agencies that are associated with governmental organisations to provide help to users with accessing e-Government systems, performing e-Transactions and corresponding with governmental organisations on behalf of users. This high percentage is likely because e-Government is still in early stages in Saudi Arabia and also the infrastructure of communications and post mails is in developing phases. Such help method is also useful for users in rural areas where there are no governmental agencies. This method can be introduced at lease in the developing phase of e-Services where many users are likely need help to access and use such e-Services. This result is supported by a study that has been conducted by Alsobhi et al. in Al-Madinah city, which is a large city in Saudi Arabia, to investigate the influence of intermediary agencies on the adoption and usage level. Those agencies were authorised to facilitate communication and coordination between public services providers and customers. The findings of Alsobhi's study showed that intermediaries are an extremely useful channel for improving the adoption and utilisation of e-Government [29]. Table 10 shows the overall score of this item (ACC1).

In the second part of Accessibility factor measurement, the participants were asked to indicate the importance degree of the existence of several features that are related to accessibility. These features are shown in Fig 4. It has been found that having full control of data presence in e-Government systems was highly important to three quarters (75%) of the participants. This indicate that users from business sector prefer to have a full access to their businesses data stored in e-Government systems and also to have full control over it where they can delete it or keep it stored. Moreover, accessing e-Government services provided to business sector at any time (24 hour and 7 days) was of high importance to approximately 77% of the respondents. Since most of the participating business entities (around 70%) were large companies, it is likely that they have large number of governmental transactions that they need to perform and the participants (just over 80% were employees) prefer to have flexibility to perform them at any time even out of working hours.

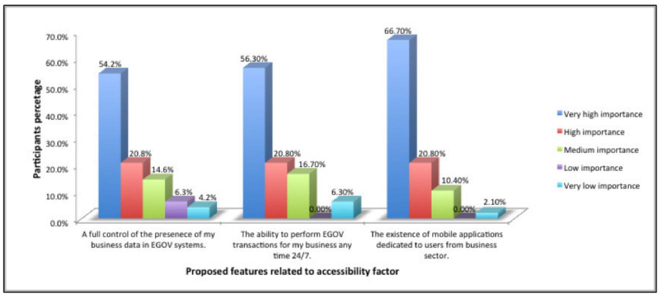


Fig. 4: The result of ACC p2

As another accessibility method that can influence the adoption and utilisation of e-Government services, providing e-Services through mobile applications was highly important to the vast majority of the participants with a percentage of 87%. This indicates that the flexibility to access e-Government systems from anywhere is very important to users from business sector and it is very likely that providing mobile application for this purpose will increase the adoption and utilisation level. Table 11 presents the overall score of ACC part 2 (1.71) which indicates the strong impact of *Accessibility* on the adoption and usage.

***Perceived Trust (PT)*** factor was measured with nine items using 5-points Likert scale ranging from 'Strongly agree' to 'Strongly disagree'. PT factor concerned with three crucial aspects, namely, trust, security and privacy. Thus, all these aspects were measured with different items. The analysis of this factor revealed that 58% of the participants agreed with that the Internet is safe to be used to perform governmental e-Transactions for their businesses whereas 27% disagreed. This indicates high security perception from the respondents towards using the online means for performing e-Government transactions although the proportion of disagreed responses was slightly high. Thus, the result was investigated further and it was found that the education level of 54% of the respondents who reported an agreement on that the Internet is not safe were secondary school or less. It is likely that the education level had an impact on the security perception when using online governmental services.

The results also indicated a relatively high privacy perception since over half of the respondents (56%) would not hesitate to provide sensitive information through e-Government systems. The same percentage of 56% also disagreed with a statement stated "Dealing with government agencies electronically online could cause invasion of my business privacy". Furthermore, 64% of the respondents

TABLE 12: The result of PT factor

| Factor | # of item | Mean | S.D. | result interpretation |
|---|---|---|---|---|
| PT | 9 | 2.34 | 0.71 | Influential |

TABLE 13: The result of RP factor

| Factor | # of item | Mean | S.D. | result interpretation |
|---|---|---|---|---|
| RP | 4 | 1.71 | 0.79 | Very influential |

believed that their business entities' data that are stored in e-Government systems cannot be used by other parties without their permission. Other parties meant to be other governmental organizations or other companies. Moreover, approximately 54% of the respondents believed that their company's data stored in e-Government systems cannot be misused. These results showed a relatively high trust, security and privacy perception.

The overall score of this factor was 2.34 (see Table 12) and this gives us an indication that Perceived Trust factor is likely influence the adoption and utilisation of e-Government from the perspective of users from business sector. Moreover, the high trust, security and privacy perceptions that have been shown in all items reflect the importance and the impact of trust for both the business owners and employees in business sector.

***Regulation and Policies (RP)*** factor is another factor that was proposed in the utilised model (EGAUM) and it was measured with four items. In these items, the participants were asked to determine the importance degree of providing important information related to the regulations and policies of using e-Government. The proposed information can be seen in Fig 5. It is clear from the figure that all the proposed information was of high importance degree for the majority of the respondents. The graph shows that around 83% reported that providing e-Version of the regulatory procedures and requirements that their business entities need to comply with was highly important to them. Such procedures and requirements are different according to the business field and activities and they are normally updated regularly by government.

Moreover, providing information security policies was also highly important to over 79% of the participants. Such policies involve security procedures applied in implementing e-Services, security procedures applied in storing users' data and security guidelines that the agencies follow when providing e-Services. Even if the agencies implement high standard policies of information security, users need to see and feel about such security matter. Furthermore, approximately 83% of the respondents reported high importance degree to the existence of information privacy policies with the highest percentage of 'Very high importance' answer.

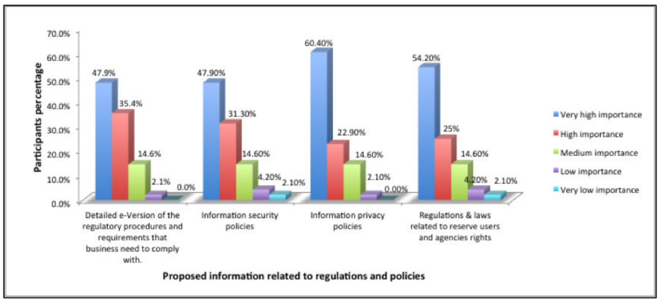


Fig. 5: The result of RP

Regulations and laws that are related to reserve users' and agencies' rights when using e-Government services were reported as of high important by more than three quarters of the respondents (79%). Such regulations and laws will protect users' rights when performing and using online e-Services for their business entities. For example, users need to know what they should do when conflict happen with e-Services provider regarding their e-Transactions. Also these regulations and laws will reserve governmental agencies rights when providing e-Services to customers and this include the cases and reasons that give the agency the right to refuse processing an e-Transaction. Remarkably, very low percentages of participants who responded low importance degree to each item.

The results revealed that applying strict, clear and detailed regulations and policies when implementing e-Government systems is very important to the users from business sector. This is likely because users from business sector are cautious and careful about their businesses' transactions. They do not want their business activities to be negatively affected due to lack of knowledge or understanding of regulations and policies especially for large business entities (around 70% of the participating business entities were large). The total score of 1.71 (see Table 13) indicates that RP is a significant factor that has a strong impact on the adoption and usage of e-Government services provided to business sector.

## IV. Conclusion

The adoption and utilisation of e-Government considered to be a challenge, and therefore, an interesting research area is introduced. This paper aimed to identify determinant of potential users' adoption and utilisation of e-Government services from the business sector perspective in Saudi Arabia. This study was based on a comprehensive framework (EGAUM) that proposed fundamental factors related to the adoption and interaction with e-Government services. The utilised model covered substantially all the aspects that have direct and indirect impact on such interaction. The results of the descriptive analysis in this study indicated that all the proposed factors have degree of influence. PB, FQS, PE, PS, ACC and RP were found to be significant factors that are very likely influence the adoption and usage level of the users from business sector.

This study is part of an on-going research and thus, the collected data will be analysed with further analysis procedures to explore the correlations between the model's variables. Although there are some limitations in this study such as the majority of male participants and the majority of large participating business firms, this study provides useful insights into the motivations underlying the intentions to adopt and use e-Government services from the business sector perspective in developing countries like Saudi Arabia.

## References

[1] E. Ziemba, T. Papaj, and R. Zelazny, *A model of success factors for E-Government adoption - the case of Poland*. Journal of Computer Information Systems, Vol. 14, Issue 2, pp. 87-100, 2013.

[2] S. AlAwadhi, and A. Morris, *Factors influencing the adoption of e-government services*. Journal of Software, 4(6), pp.584-590, 2009.

[3] M. Kunstelj, T. Jukic, and M. Vintar, *Analysing the Demand Side of E-Government: What Can We Learn From Slovenian Users?*. Electronic Government Lecture Notes in Computer Science, Vol.4656, pp 305-317, 2007. DOI: 10.1007/978-3-540-74444-3_ 26.

[4] Research and Economic Reports Unit 2015 [Growing rate of private sector in the Kingdom of Saudi Arabia]. Al-Jazirah [Online] Available from: http://www.al-jazirah.com/2015/20150315/ec6.htm (accessed in 28/04/2015) (In Arabic).

[5] S. Alghamdi, and N. Bellof, *Towards a Comprehensive Model for E Government Adoption and Utilisation Analysis: The Case of Saudi Arabia*. Federated Conference on Computer Science and Information Systems, ACSIS, Vol. 2, pp. 1217-1225, 2014. DOI: 10.15439/2014F146.

[6] J. Raubenheimer, *An item selection procedure to maximise scale reliability and validity*. SA Journal of Industrial Psychology, 30 (4), 59-64, 2004.

[7] N. Blunch, *Introduction to Structural Equation Modelling Using IBM SPSS Statistics and AMOS*. Second Edition, SAGE publication Ltd, p.p. 31-50, 2013.

[8] S. McLeod, *What is Reliability?*, 2013. [Online] Available from: http://www.simplypsychology.org/reliability.html (accessed in 28/01/15).

[9] E. Eucharia, and O. Nnadi, *Health Research Design and Methodology*. Library of Congress, CRC Press, p.p. 166-168, 1999.

[10] D. Suhr, and M. Shay, *Guidelines for Reliability, Confirmatory and Exploratory Factor Analysis*. University of Northern Colorado, Cherry Creek Schools, USA, 2008.

[11] D.K. Bhattacharyya, *Cross-cultural Management: Texts and Cases*. PHI learning private limited, New Delhi, p.p.317-319, 2010.

[12] D. Colton, and R.W. Covert, *Designing and Constructing Instruments for Social Research and Evaluation*. Jossey-Bass, Wiley Imprint, San Francisco, p.p. 64-74, 2007.

[13] B.R Worthen, W.R. Borg, and K.R. White, *Measurement and evaluation in the schools*. New York: Longman, 1993.

[14] B. Nevo, *Face validity revisited*. Journal of Educational Measurement, 22, pp. 287-293, 1985.

[15] J.A. Khan, *Research Methodology*. S.B. Nangia, APH Publishing Corporation, New Delhi, p.p. 135-143, 2008.

[16] Laerd Statistics (n.d.) *Measures of Central Tendency*. [Online] Available from: https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php (accessed in 12/04/2015).

[17] D.L. Clason, T.J. Dormody, *Analyzing data measured by individual Likert-type items*. Journal of Agricultural Education, 35(4), p.p.31- 35, 1994.

[18] H.N. Boone, and D.A. Boone, *Analyzing Likert Data*. Journal of Extension, vol. 50, No. 2. Article number: 2TOT2. West Virginia University, Morgantown, West Virginia, 2012.

[19] L. Carter, and F. Belanger, *Citizen adoption of electronic government initiatives*, 37th Hawaii International Conference on System Sciences, Hawaii, 2004.

[20] D.V . Dimitrova and Y .C. Chen , *Profiling the adopters of e-government information and services: the influence of psychological characteristics, civic mindedness, and information channels*. Social Science Computer Review,(24:2), pp.172-188, 2006.

[21] S. AlAwadhi and A. Morris, *The use of the UTAUT model in the adoption of e-government services in Kuwait*. 41st Hawaii International Conference on System Sciences, Hawaii, 2008.

[22] S. AlAwadi, and A. Morris, *Factors Influencing the Adoption of E-government Services*. Journal of Software, Vol.4, No.6, pp. 584-590, 2009.

[23] KH. H. Albayari, *Social Media Websites Create Revolution in Business Sector*, 2011. Alriyadh newspaper. [Online] Available from: http://www.alriyadh.com/674523

[24] H. AlShihi, *E-government development and adoption dilemma: Oman case study*. 6th International We-B (Working for e-Business) Conference, Victoria University, Melbourne, Australia, 2005.

[25] P . Beynon-Davis, *Constructing electronic government: the case of the UK inland revenue*. International Journal of Information Management, 25, 3-20, 2005.

[26] L. Carter and F. Belanger, *The utilization of e-government services: citizen trust, innovation and acceptance factors*. Information Systems Journal, (15:1), pp.5-25, 2005.

[27] C. W. Phang, Y. Li, J. Sutanto and A. Kankanhalli, *Senior citizens adoption of e-government: in quest of the antecedents of perceived usefulness*. 38th Hawaii International Conference on System Sciences, 2005.

[28] E. Ziemba, T. Papaj, and D. Descours, *Assessing the quality of e-government portals - the Polish experience*. Federated Conference on Computer Science and Information Systems, ACSIS. Vol.2, pp.1259-1267, 2014. DOI: 10.15439/2014F121

[29] V. Weerakkody, R. El-Haddadeh, F. Al-Sobhi, M.M. Shareef, and Y.K. Dwivedi, *Examining the influence of intermediaries in facilitating e-government adoption: An empirical investigation*. International Journal of Information Management, 33, p.p. 716-725, 2013. DOI: 10.1016/j.ijinfomgt.2013.05.001

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1481–1488

DOI: 10.15439/2015F49
ACSIS, Vol. 5

# A Platform for Context-Aware Application Development: PCAD

Ufuk Celikkan
Izmir University of Economics
Dept. of Software Engineering
Sakarya Cad. 156 35330
Izmir, Turkey
Email: ufuk.celikkan@ieu.edu.tr

Kaan Kurtel
Izmir University of Economics
Dept. of Software Engineering
Sakarya Cad. 156 35330
Izmir, Turkey
Email: kaan.kurtel@ieu.edu.tr

□

*Abstract*— **We propose a novel software platform based on the notion of context-awareness which allows rapid and easy development of context aware applications. One of the fundamental goals of the proposed platform is extensibility, allowing the platform to react to new requirements without making fundamental and substantial changes. The Context Aware Application Development Platform- PCAD- has inspired from an operating system and modeled as a layered architecture. It exhibits a plug-and-play behavior very similar to devices and device drivers found on an operating system/kernel. The platform offers functions such as data management, notification, security and privacy as services. The platform provides a public interface to application developers in their development of context-aware applications. It covers a wide range of disparate application domains such as smart city and livestock monitoring applications.**

## I. Introduction

APPLICATIONS using information and communication technologies collect and process a diverse range of data using machines connected through communication networks. This phenomenon is captured in the term the *Internet of Things*. The Internet of Things is a complex interconnection of heterogeneous devices that include sensors, cameras, micro chips and RFID based products, and which generate large amount of data obtained from various domains. For this reason platforms or architectures must be designed in anticipation of the increase in the number of devices, and the varying demands of users and applications in different contexts. In order for a platform to be successful, it is imperative that it recognizes the context in which users and applications are operating, and enable service customization for a particular user. The creation of smart applications and environments then becomes a possibility by using context-aware computing, which acquires, analyzes, and interprets relevant context information, and responds to contextual changes.

Temperature, humidity, traffic congestion, road conditions, sea pollution, and river level are some examples of context information. Such context information can be used alone or in combination within context-aware applications to provide custom services in various domains, such as transportation, health and medical systems, tracking and control of environment, energy, agriculture, industry, sport events, and tourism. However, the effective use of this context information requires its efficient and effective acquisition, storage, processing and reasoning. In this way, productivity, economic output and quality of life can be increased.

This paper describes a novel, service-based software platform proposal based on the notion of context-awareness. The platform basically follows a middleware approach, which draws on the techniques taken from operating system design. The primary goal of the proposal is to offer a platform to simplify the development of context-aware applications by relieving the applications from complex context data management issues. In an open, dynamic and continuously changing environment, context data must be acquired, managed and ultimately offered to applications which will interpret the data according to the situation. The platform separates context acquisition from the application code and handles many context data management issues on behalf of the applications. The fundamental design force of the proposal is that the platform is agile, robust, and capable of reacting to new requirements without the need for fundamental and substantial changes.

The rest of the paper is organized as follows. The motivation section explains the rationale and design issues that motivated this work. Related works section includes an overview of the existing systems based on context-aware features. Section 4 presents the technical architecture of the platform. Section 5 gives a practical scenario from a smart city parking application, and finally, Section 6 draws conclusions about the proposed system.

## II. Motivation

Dey et al. has listed three characteristics of a context aware system as (i) information and services are presented to a user, (ii) a service is executed, and (iii) context is linked to information for later retrieval [1], [2]. Therefore, a service

□ This work is supported by the Scientific and Technological Research Council of Turkey – TUBITAK under Grant 114E938.

978-83-60810-66-8/$25.00©2015, IEEE

based software infrastructure support facilitates building context aware systems as it simplifies system creation by placing common context aware computing features in the infrastructure as services. A fundamental advantage of a service-based infrastructure approach is that the resources (i.e. data, sensors, devices, services) are shared, making maintenance and evolution of the infrastructure much easier [3]. Such an infrastructure is best realized in the form of a middleware, in which the infrastructure acts as a foundation on which other systems can be built. It takes responsibility for acquisition, storage, processing and interoperability of the context data, and provides services such as alarm, notification and security to the applications.

An architecture used in building the service-based infrastructure must possess certain attributes in order to be a viable option in creating context aware systems. Notable among those attributes are extensibility, robustness, scalability, and ease of use. Extensible architectures allow the addition or removal of components without breaking or taking down the system, and facilitate the seamless deployment of new sensors, devices and services. The entire system can evolve in a step-wise fashion. Robustness requires that the infrastructure operates in the case of a malfunction and terminates peacefully when breakage is unavoidable, while scalability ensures that infrastructure is able to serve many users and collect data from a diverse range and number of sensors. The amount of administrative effort must also be minimized to maintain the efficiency of services, devices and sensors. Finally, no matter how well the architecture is designed, it must be efficient and simple to use. Efficiency is measured in terms of time and space, which are influenced by the data storage, response and real time requirements of the system.

An Operating Systems inspired approach is the most suitable proposition in designing context aware infrastructure architectures. The primary aspect of Operating System architectures is the abstraction of inner details through layering and providing a high-level interface to allow uniform access to the functions of the system contained in its layers. Two typical examples from the UNIX world are the POSIX and device driver interfaces. POSIX interface allows the applications interact with OS services such as file system, network stack and devices, and device driver interface allows the seamless integration of new devices and drivers to OS kernel.

For the reasons mentioned above, a service based middleware architecture inspired by an OS is used in the design of PCAD system. The sensors, devices, services and more importantly, the applications are separated from each other, yet interact with one another. The loose coupling of components permits plugability and reduces fragility, the two essential characteristics of an extensible and robust system.

A prototype of the platform is currently under implementation together with two sample applications that utilize the platform. One of these is a livestock monitoring application that monitors livestock physiological data and notifies the farmer and the veterinary physician of potential livestock health problems. The second application is an implementation of a smart city parking scenario described in Section 5. Both applications will allow the evaluation of the effectiveness, robustness and performance of the platform applied to diverse domains. It will also demonstrate that the platform facilitates rapid context-aware application development.

Despite being an important issue in context aware systems, we have chosen not to address context modeling. Context is defined by [2] as: “any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves”. A piece of information is considered context data only if it is interpreted, otherwise, it is simply information belonging to an environment [4]. For applications and services to adapt their behavior based on the particular context, they need to be able to do reasoning. The method chosen for modeling the context has direct influence on the kind of reasoning performed [5], [6]. Therefore it is very important to choose a modeling technique that will facilitate not only easy reasoning, but also sharing of context with others. Since our primary focus is to define the platform itself and the interactions among the components, we defer the discussion of context modeling for our future work.

## III. Related Work

Many context aware architectures proposed in the past have borrowed ideas and features from operating system designs. Gaia project [7] describes one such architecture which is heavily influenced by operating system design. It extends the OS concepts by incorporating context-awareness. It has a context file system that organizes data by context and a Unix-like signal event mechanism to create channels to connect context data suppliers to consumers through a handle called *ContextType*. Another architecture influenced by operating system design is Context Toolkit in which a widget model is adapted resembling the device driver model of an OS. In the widget model, widgets are software components that provide an interface to a hardware sensor. In other words a widget acts like a device driver to a sensor. Context Broker Architecture (CoBrA) [8] employs an intelligent context broker to provide a centralized context model shared by devices, agents and services. One of the responsibilities of the context broker is to ensure that user privacy is protected through policies by allowing the user to control how contextual information is shared. The issue of how to control access to shared information is of key importance in operating systems.

There are other architectures based on layered architecture model which have the goal of abstracting sensor data management from users. SOCAM, Service Oriented Context

Aware Middleware [9], project is a middleware architecture that supports the building of context aware services in pervasive environments. Context aware services, which may also be applications, can obtain context using the context middleware layer from context providers. As expected from a middleware-based architecture, in SOCAM, context aware application code is separated from context acquisition and reasoning. Another example of a centralized middleware approach is found in CASS project [10], where the middleware itself is a server responding to context management requests of mobile context aware applications. Context acquisition is done by the sensor nodes and context data is sent to the server. Storage, modeling and reasoning tasks are performed by the middleware server. Consequently CASS applications neither need to communicate with the context source directly, nor are they affected by context-based inferences and behaviors. Another layered architecture tailored towards mobile devices that addresses the shortcomings and special needs of mobile devices is the Hydrogen [11] project. One of its three layers, the Management layer is responsible for storing context information and sharing it using XML protocol. A second layer, the Adapter layer, delivers information obtained from context sources, i.e. sensors, to the Management layer. The Application layer, where the application code resides, can access context data synchronously or asynchronously.

A common benefit of these architectures is their capability to prevent coalescing of application code with the context management functions by decoupling context sensing and acquisition from the application code.

In addition to architectures and frameworks several context aware systems for different domains have been built or proposed in the literature. Context aware systems for pervasive environments are surveyed in [12] and web services based systems are studied in [13] and a classification of context aware systems is given in [14].

## IV. PCAD Architecture

Context-Aware Application Development Platform-PCAD- has been inspired by operating systems design and modeled as a layered architecture. PCAD's primary goal is to offer extensible, scalable, robust, secure and general purpose software architecture for use by developers in several industrial processes and sectors. The platform particularly addresses the following design issues:

1) Extensibility
2) Security and Privacy
3) Simplicity (be applicable and pragmatic)
4) Generic (not domain specific)
5) Service Based

The platform itself adapts a middleware and blackboard [4] model, providing common services to the applications. This is similar to the conceptual separation of user and kernel services in an OS. Among the services provided by PCAD are data storage, alarm, notification, a simple rule engine and reporting facilities. One can view the platform as a layer providing kernel-like services, while the applications provide user level services and the physical sensors and their software act like devices and device drivers providing data to applications. The applications subscribe themselves to the sensors through the platform, after which they are notified by the context providers when a context data becomes available. The applications and the sensor software are bound to the platform using standard protocols (e.g., REST, web sockets, and raw socket). The architecture exhibits a plug-and-play behavior similar to the devices and their associated drivers found in operating system/kernel. The sensors can be added and removed, and applications can be registered to the newly added sensor without needing to modify the platform. The platform provides a public interface to enable the rapid and easy development of applications from various domains by application developers. It relieves the applications from implementing common services and functions, thus fulfilling one of the fundamental principles of software engineering – reusability.

### A. PCAD Services

PCAD's functionality is made available to applications through services. The services of PCAD and its architectural resemblance to Operating System architecture is shown in Figure 1. The list of services available in PCAD is given in Table I and explained in detail below.

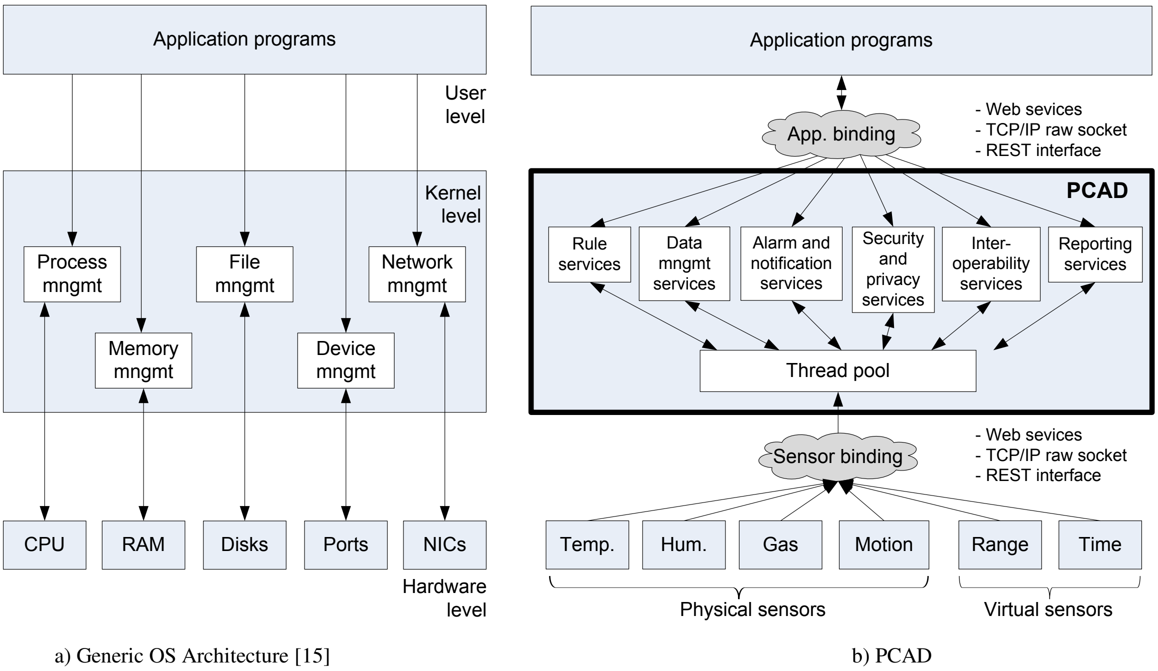
Table I.
PCAD Services

| Service name | Service description |
|---|---|
| Rule Service (RS) | Context processing |
| Data Management Service (DMS) | Stores context data and provides data upon request through a uniform interface. Underlying storage mechanism is transparent to the users. |
| Alarm and Notification Service (ANS) | Notifies the registered parties about context data. Filters data if necessary. |
| Reporting Services (RPS) | Generates detailed reports about context sources including data and status of the context source. |
| Security and Privacy Service (SPS) | Ensures only authorized parties can access to context data based on policies. |
| Interoperability and Communication Service (ICS) | Data exchange among similar platforms using standard communication protocols and message formats. |

*Rule Services*

A very simple and generic syntax will be designed in the form of if-then rules. The rules will be specified in an XML file, and will be verified against the Rule XML Schema using XSD. A generic rule will be of the following format:

```
if <cond> then action
```



a) Generic OS Architecture [15]

b) PCAD

Fig. 1 A Comparison of PCAD architecture and OS architecture

The condition in the rule can be a combination of conjuncts and disjuncts. When the condition evaluates to true, then the action is performed. This action could be further used to trigger other rules. Below, one such example is given:

```
if (raining and
    location=″main street″)
then congestion
```

When this rule is executed, firstly rain sensor data is retrieved, followed by location information. If the condition evaluates to true, which means that it is raining and the location name is *main street*, then the `congestion` rule is executed. Congestion rule simply retrieves the congestion data for the main street using the congestion sensor and returns it to the application.

The applications create a rule file in XML and hand over to PCAD for context processing. Alternatively, applications may perform their own context processing, thus opting out using the Rule service of PCAD.

*Data Management Services*

Data Management Services is responsible for storing of data received from multiple sources and making the data available to those requesting it. It provides a uniform interface for the transparent access of data by the applications. This interface let users of this service insert, update, delete and filter information. It is essential that the sensor data, whether collected synchronously or asynchronously, is stored without any loss. DMS could store data simply in a relational database, or it can use a cloud service. The underlying storage mechanism is transparent to the applications.

*Alarm and Notifications Services*

The Alarm and Notification Services (ANS) follow the paradigm of observer design pattern [16]. The applications attach themselves to a sensor of interest to acquire context information. There are two modes to query the presence of sensor data. In the asynchronous mode, ANS periodically queries the sensor and then reads and stores that data. It then immediately makes the data available to the applications that showed interest. In synchronous mode, the sensor software interrupts the ANS for a reading. Asynchronous mode requires support from the sensor software, which must notify ANS for data availability. This is generally not the conventional method of interacting with a sensor, and therefore is not available in most cases. The ANS provide sensor data to applications using either pull or push method. In the pull model, the applications receive only a minimal notification, and applications or users ask for specific details thereafter. In the push model, ANS sends applications detailed information about the sensor data.

A filtering mechanism will be employed to allow applications to register themselves to specific conditions so that they will be notified when such condition occurs. An example filter is shown in Figure 2. An application named *myApp* is only interested in data from the temperature sensor *temp001* between 9:00 AM and 12:00 PM, and only for values 39 and above.

*Reporting Services*

Detailed reports presenting sensor data and sensor availability shall be provided by the Reporting Services (RPS). Visualization of the sensor status and data provides a useful overview of the system, which is important because it is not easy to understand and interpret raw data. The data is presented to the user in different formats, such as tables, charts and text. It was decided to select open source reporting tools with an application programming interface. BIRT [17] is one such tool with several APIs to design, generate and render reports and charts.

```
<?xml version="1.0" encoding="UTF-8"?>
<Configuration name="MyApp">
<Sensor name="TME3m" id="temp001" >
<TimeFilter start="09:00:00"end="12:00:00"
      onMatch="ACCEPT" onMismatch="DENY"/>
<ThresholdFilter value="39"
      onMatch="ACCEPT" onMismatch="DENY"/>
</Configuration>
```

Fig. 2 Example filter

*Thread Pool*

The platform will ensure that the context data received from the sensors are properly stored and sent to the applications in a timely manner. A pool of threads, as shown in Figure 3, are to be created and tasks are assigned to threads whenever needed. Each thread will have two sub threads; one is responsible for real time delivery of raw data to applications, and the other responsible for the storage and processing of the context data.

*Real-time support*

A certain class of applications imposes constraints on the sensor data delivery time. Context aware systems using direct sensor access approach are better suited for meeting the hard real time requirements. Such systems collect information directly from the sensor without any intervening components, thus, time lapse in reacting to the events conforms to the real-time requirements. However, a disadvantage of these systems is their inability to re-use components due to the tight coupling of sensor device drivers with the application [11]. In contrast, systems that are built using layered architectures are considered superior because they are extensible and allow modifications to the architecture without modifying the applications. However, meeting the hard real-time requirements is difficult in layered architectures, as data flows through several layers before it reaches the application. Therefore, layered architectures are much more suited to soft real-time requirements. Soft real-time systems can tolerate larger time latencies and they still operate even when time constraints are violated, albeit with degraded performance.

PCAD addresses the soft real-time requirement by establishing a direct pipe or channel between the sensor software and the applications, as depicted in Figure 4. A dedicated thread is allocated to the channel, allowing raw sensor data to be transmitted through this channel, while a copy of the data is processed and stored by the platform concurrently.

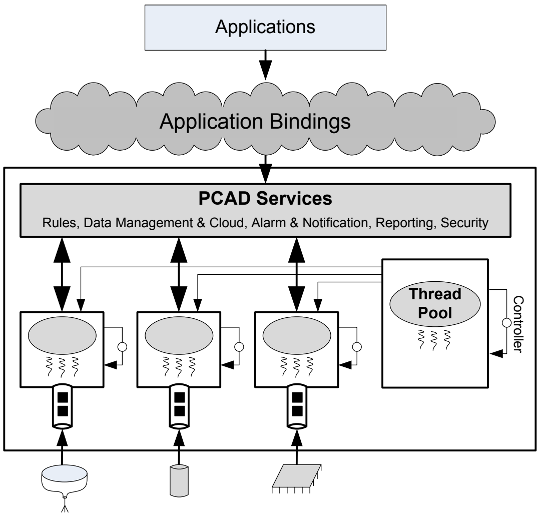

Applications
Application Bindings
PCAD Services
Rules, Data Management & Cloud, Alarm & Notification, Reporting, Security
Thread Pool
Controller

Fig. 3 Thread Pool in PCAD

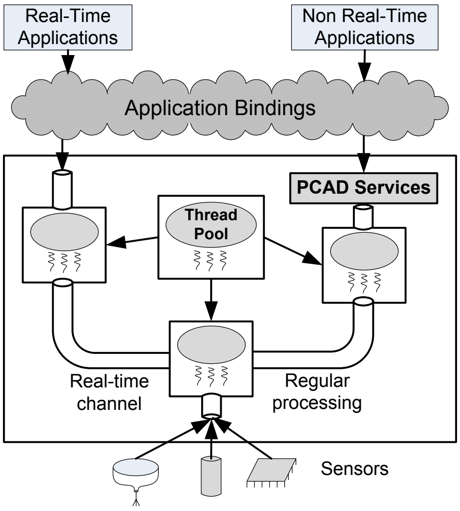


Fig. 4 Real-time support

*Security and Privacy Services*

Privacy is an important issue in PCAD as sensors will acquire information about the environment, and share it with different applications. As the sensor data is owned by the platform rather than the applications, it is the platform's responsibility to protect the privacy of the data.

A very basic Role Based Access Control (RBAC) [18] scheme mediates which sensor data is accessible to which application. An RBAC-like access control mechanism is preferred because it allows more complex policies to be accommodated in the future. The idea is to assign applications particular roles and assign access rights to roles, and then control which application can access sensor data based on those roles. Our use of RBAC will be much more closely aligned with Mandatory Access Control mechanism, where the sensor data is assigned privacy levels, and the applications are given clearance levels. The set of roles will be defined by the platform based on policy configurations. For instance, a weather application does not need to access to the transportation data.

The three basic elements of access control, Subject, Object and Access Right have the following mappings in PCAD: the subject is the software processes in which the applications run. Object is our context resources. These context resources may be physical or virtual sensors. Access Right would be the read right of the context data, or querying the status of the sensor providing the data.

Tables II-A, II-B and II-C show an example of role based access control. In this example, the subject is the application and the object is the sensor. *Application1* has been assigned roles *role1* and *role2*, which gives application the right to read *sensor2* and *sensor3*, and to query the status of *sensor1*. On the other hand *application2* can only query the status of the three sensors. RBAC makes it easy to put constraints, for instance, we could limit the number of applications collecting data from a sensor by limiting the number of applications assigned to a role.

TABLE II.
A. AUTHORIZATION EXAMPLE

| **Authorization** | **Description** |
|---|---|
| **read** | Allowed to acquire context data |
| **query** | Allowed to query the status of the context source |

B. APPLICATION ROLE MAPPING

| **Subject** | **Roles** | | |
|---|---|---|---|
| | role 1 | role 2 | role 3 |
| **application 1** | * | * | |
| **application 2** | | | * |

C. ACCESS CONTROL IN PCAD

| **Roles** | **Sensors** | | |
|---|---|---|---|
| | **sensor 1** | **sensor 2** | **sensor 3** |
| **role 1** | query | read | read |
| **role 2** | | read | |
| **role 3** | query | query | query |

Two aspects of equal importance are the authenticity of the sensor, and protection of the confidentiality and integrity of the information it provides. The field of network and information security has sufficiently matured in the provision of security features to allow combination of symmetric and asymmetric cryptography (digital certificates) to fulfill the authentication, confidentiality and integrity requirements of the platform.

### *Interoperability and Communication Services*

According to a recent survey, interoperability is listed as the number one quality driver for smart city applications [19]. Interoperability is a necessity since data is owned by other parties and managed by different platforms, similar to PCAD and also for making PCAD data available to other requestors. This creates a need for interoperability for exchanging context information among different applications. Interoperability and Communication Services are responsible for context data exchange using standard messaging format. As shown in Figure 5, PCAD uses XML as the data exchange format.

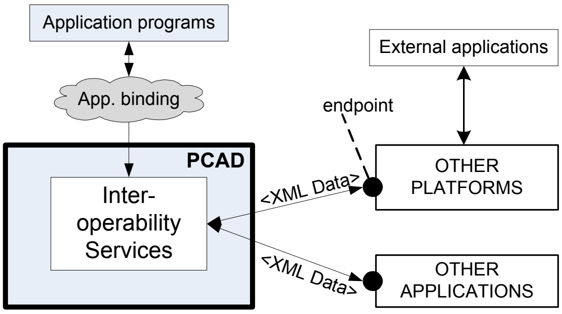


Fig. 5 Context data exchange

Applications built on top of the PCAD platform use this service to send context data to other applications. The applications can also ask the platform to receive context data from other platforms, including from another instance of the PCAD platform. In such a case, the applications must let the service know the endpoint information that has the data.

### *B. Sensor Layer*

The sensor layer consists of physical and virtual sensors. Examples of physical sensors are pH, humidity and temperature. Unlike physical sensors, virtual sensors use either Application Programming Interfaces or web services to obtain data, such as GPS position or real-time camera stream [20]. In many situations, data can be made meaningful by aggregating individual context data obtained from multiple physical and virtual sensors. This process, called context aggregation, generates combined data that is more accurate and relevant than information obtained from a single source. Context aggregation function is implemented also as a virtual sensor.

The sensor software will interact with the platform by registering to the platform and sending raw data to the platform, which is further formatted and processed by the platform. Sensor layer interacts with the platform through sensor binding layer.

### *C. Binding Layers*

For the purpose of allowing applications and sensor software to communicate with the platform, the platform contains two binding layers. The first is between the application software, and the platform and the second is between the sensor software and the platform. These binding layers are facades that simplify the application and the sensor software's interaction with the platform. The application binding layer is used by the applications when they require the services of the platform, such as receiving context data, registering and deregistering for sensor data, creating reports and receiving alarms and notifications. The sensor binding layer is used by the platform to gather data from a sensor and deliver it to the registered applications, possibly multiple times. Therefore, it is possible for many applications to access the same context data simultaneously, which may not be possible with the direct access to sensor alternative.

The implementation of the binding layer employs widely used protocols. Among the protocols supported are web services, web sockets and REST.

## V. A Scenario: Smart Parking

In this section, we present a scenario to demonstrate the interactions between an application PCAD and sensors. The scenario involves a context aware smart city application running on a mobile device developed for the purpose of finding a parking space based on certain constraints.

*One day Mr. Jones went out of the house and boarded his car. Just he was about to approach to his workplace in downtown, he remembered that he wasted time searching for a parking place the previous week. He started the application on his mobile phone and received the coordinates and the route information about the nearest parking location within a 500 meter radius on a map. He thought, "I did a good job of using this context-aware [sic] application to access to the information provided by the smart [sic] city, otherwise I would have been late for my appointment again."*

Following is the sequence of actions that take place in the application and PCAD platform:

1) User invokes the smart city context aware application on the mobile phone.
2) Smart city application starts a session with the PCAD platform.
   *A communication session is opened using WebSockets. The application and PCAD now begin to communicate via the application binding layer for fulfilling user's specific request about an available parking place.*
3) User requests route information about a parking place within the 500 meter radius of his workplace.
   *The user's operational parameters, such as the 500 meter constraint, are translated into rules and registered with the Rule Services (RS) of CAS, using the session established before. These rules get executed by PCAD later on.*
4) Smart city application sends user requests to PCAD via application binding layer.
5) PCAD periodically requests ground parking data from installed sensors, which includes sensor fingerprint (id, location, time) and parking availability.
   *Since the actual parking status changes often due to usage conditions, ground sensors are asked to send parking availability information in real-time. The frequency of sensor query is a configurable parameter. The scenario is an example of soft real time application because the latency of getting availability information is not critical to the operation of the application. PCAD collect these data via the sensor binding layer and notifies the application via application binding layer using alarm and notification services (ANS). The real time channel is used to send the data to the application. PCAD enforces data privacy by only providing parking data to the application.*
6) Smart city application determines the user's location.
7) Smart city application finds optimal parking space for the user.
   *Smart city application acquires user's location information from mobile device, and parking availability and parking location information from PCAD. The application then offers an optimal path using this information together with other constraints, such as shortest path, traffic signs or number of traffic lights etc. The application either gets some of this constraint information from PCAD or PCAD can aggregate this information from other sources, including other platforms using the interoperability and communicating services (ICS). This part directly points all other smart city applications working together.*
8) Smart city application presents the information including, coordinates and map to user.

This scenario shows how PCAD services are utilized by the application, and how the actions are performed by the actors of the system. Figure 6 shows the steps of interaction described above. The figure illustrates how sensor acquisition is decoupled from the application code.

This scenario can easily be adapted to other problems, such as finding a charging station for an electric car. A context-aware charge management system for electric vehicles is described in [21]. A more elaborate problem on parking lot management for charging stations is described in [22], while a parking reservation system based on telecommunication APIs is given in [23].

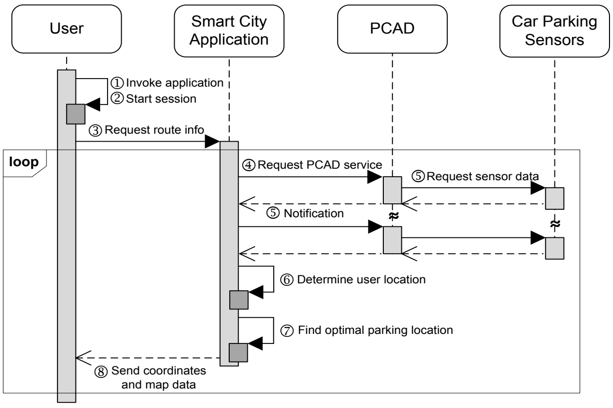


Fig. 6 Interaction diagram of the scenario

## VI. Conclusions

In this paper we propose a Platform for Context-Aware Application Development -PCAD based on operating system principles for rapid development and deployment of context aware applications. The platform has been inspired by Operating System design, and is based on a layered architecture that is modular, extensible, and follows the design pattern guidelines. The major premise of the platform was to provide a robust environment influenced by the theories of the architectural design of an operating system, and enable application developers to create context aware applications with minimum effort. The platform provides several services to both applications and sensor software. The platform has a number of benefits: it supports a diverse range of context sources, allows applications to access context sources synchronously or asynchronously using filters, decouples application code form context acquisition, and provides a simple rule engine for context processing. The platform supports soft-real time requirements by opening a direct channel between the context source and applications.

A prototype of the platform is currently under implementation along with two very distinct applications: livestock monitoring and smart city parking. These context-aware applications will verify that applications from a wide range of disparate domains can be built rapidly using the platform.

We purposefully omitted the treatment of context modeling, as our focus was to lay out the critical components of the infrastructure. Context modeling work is deferred to our future iterations of the platform. For this reason, our rule service currently provides very rudimentary services.

## Acknowledgment

The authors thank to the Scientific and Technological Research Council of Turkey - TUBITAK for their support on the 3001-Starting R&D Projects Funding Program for the project of "Architecture, Design and Implementation of Context Aware Service Platform", Grant No: 114E938.

## References

[1] A. K. Dey and G. D. Abowd, "Towards a Better Understanding of Context and Context-Awareness," in *Proc. of the Workshop on the What, Who, Where, When and How of Context-Awareness,* ACM Press, New York, 2000.

[2] A. K. Dey, "Understanding and using context," *Personal Ubiquitous Computing,* 2001, 5(1), pp. 4-7.

[3] J. I. Hong and J. A. Landay, "An Infrastructure Approach to Context-Aware Computing," *Human-Computer Interaction,* 2001, 16, pp.287–303.

[4] T. Winograd, "Architectures for Context," *Human-Computer Interaction,* 2001, 16, pp. 401–419.

[5] M. Perttunen, J. Riekki and O. Lassila, "Context representation and reasoning in pervasive computing: a review," *International Journal of Multimedia and Ubiquitous Engineering,* 2009, 4(4), pp. 1–28.

[6] T. Strang and C. Linnhoff-Popien, "A Context Modeling Survey," *First Int. Workshop on Advanced Context Modelling, Reasoning and Management at UbiComp,* 2004.

[7] M. Roman, C. Hess, R. Cerqueira, A. Ranganathan, and et al., "A middleware infrastructure for active spaces," *IEEE Pervasive Computing,* 2002, 1(4), pp. 74–83.

[8] H. Chen, "An Intelligent Broker Architecture for Pervasive Context-Aware Systems," PhD Thesis, University of Maryland, Baltimore County, 2004.

[9] T. Gu, H. K. Pung, D. Q. Zhang, "A service-oriented middleware for building context-aware services," *J. of Network and Computer Applications,* 2004, 28 (1), pp. 1-18.

[10] P. Fahy and S. Clarke, "CASS - a middleware for mobile context-aware applications," *Workshop on Context Awareness, MobiSys, 2004,* 2004.

[11] T. Hofer, W. Schwinger, M. Pichler, and et al., "Context-awareness on mobile devices – the hydrogen approach," in *Proc. of the 36th Annual Hawaii Int. Conf. on System Sciences,* 2002, pp.292–302.

[12] M. Miraoui, C. Tadj and B. C. Amar, "Architectural Survey of Context-Aware Systems in Pervasive Computing Environment," *Ubiquitous Computing and Communication Journal,* 2008, 3 (3), pp. 68-76.

[13] H. L. Truong and S. Dustdar, "A survey on context-aware web service systems," *Int. J. of Web Information Systems,* 2009, 5(1), pp 5–31.

[14] J. Y. Hong, E. H. Suh and S. J. Kim, "Context-aware systems: A literature review and classification," *Expert System with Applications,* 2009, 36(4), pp. 8509-8522.

[15] S. Stallings, *Operating Systems: Internals and Design Principles* (7. Ed.) Prentice Hall, 2012.

[16] E. Gamma, R. Helm, R. Johnson and J. Vlissides, *Elements of Reusable Object-Oriented Software.* Prentice Hall, 1994.

[17] BIRT. http://www.eclipse.org/birt/ [Accessed April 9, 2015]

[18] D. Ferraiolo, J. Cugini, R. Kuhn, "Role-based access control (RBAC): Features and motivations," *in Proc. of the 11th Annual Conf. On Computer Security Applications* (New Orleans, LA, Dec. 11-15), 1995, pp. 241–248.

[19] G. Kakarontzas, L. Anthopoulos, D. Chatzakou, D. and A. Vakali," A Conceptual Enterprise Architecture Framework for Smart Cities A Survey Based Approach," Int. Conf. on e-Busines, 2014.

[20] M. Baldauf, S. Dustdar and F. Rosenberg, "A survey on context-aware system," Int. J. of Ad Hoc and Ubiquitous Computing, 2007, 2(4), pp. 263-277.

[21] N. Masuch, M. Lutzenberger, S. Ahrndt, A. Hessler and S. Albayrak, "A context-aware mobile accessible electric vehicle management system," in *Proc. of the 2011 Federated Conference on Computer Science and Information Systems (FedCSIS),* 18-21 Sept. 2011, pp. 305-312.

[22] S. Gökay, C. Terwelp, C. Samsel, K-H Krempels K-H, S. Rabenhorst and B. Greber, "Parking Lot Management for Charging Stations," in *Proc. of the 3rd Int. Conf. on Smart Grids and Green IT Systems-SMARTGREENS 2014,* Barcelona, Spain, 2014.

[23] P. Trusiewicz and J. Legierski, "Parking Reservation – application dedicated for car users based on telecommunications APIs," in *Proc. of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS),* 8-11 Sept. 2013, pp. 865-869.

DOI: 10.15439/2015F171
ACSIS, Vol. 5

# Factors affecting the intention to use e-Government services

Prodromos Chatzoglou
Democritus University of Thrace,
Department of Production and
Management Engineering,
Vasillisis Sofias 12, 67100,
Xanthi, Greece
Email: pchatzog@pme.duth.gr

Dimitrios Chatzoudes
Democritus University of Thrace,
Department of Production and
Management Engineering,
Vasillisis Sofias 12, 67100,
Xanthi, Greece
Email: dchatzoudes@yahoo.gr

Symeon Symeonidis
Democritus University of Thrace,
Department of Electrical and
Computer Engineering,
Vasillisis Sofias 12, 67100,
Xanthi, Greece
Email: ssymeoni@ee.duth.gr

***Abstract*—Nowadays, more and more people are using Information and Communication Technologies (ICTs) in order to accommodate their daily needs. E-Government (e-Gov) adopts these technologies in an effort to provide prompt and secure services to citizens. However, the intention to use e-Government services has not yet been fully examined by the international literature. The present research attempts to bridge this gap, by examining the factors affecting citizens' intention to use e-Government services. In that direction, a conceptual framework (research model), based on an extensive review of the relevant literature, has been developed. The proposed conceptual framework has been empirically tested using a newly-developed structured questionnaire. Data were collected from a sample of 547 Greek citizens. The reliability and the validity of the questionnaire have been thoroughly examined, while the Structural Equation Modeling (SEM) technique has been used to analyze the data. Results indicate that perceived usefulness is the most important determinant of the intention to use e-Government services. Other important factors are perceived trust, internet experience, peer influence, computer self-efficacy and perceived risk.**

## I. Introduction

DURING the last 15 years, information and communication technologies (ICTs) play a central role in the global digital economy and are considered key tools of worldwide administrative reforms [1]. e-Government (e-Governance and / or e-Gov) is defined as "*an initiative aimed at reinventing how the government works and improving the quality of interactions with citizens and businesses through improved connectivity, better access, furnishes high quality services and better processes and systems*" [2]. Moreover, according to Ziemba, Papaj and Jadamus-Hacura [3], e-Government "*suggests the use of information and communication technology (ICT) to provide efficient and quality government services to employees, government units at the state and local levels, and to citizens and businesses*".

At the same time, official states (governments) are using e-Government in order to improve public services and strengthen political processes.

The intention of citizens to use e-Government services has been investigated by several previous studies [1] [2] [3]. Despite that, these studies have not reached consensus, while most of their empirical results are quite obvious (investigating already established concepts and validating ideas that have already been proven valid). Simultaneously, in the practical level, there have been many failed and costly attempts in developing citizen acceptance concerning e-Government services [1] [2].

The main objective of the present paper is to investigate the factors that have an impact on citizens' intention to use e-Government services. The proposed conceptual framework takes under consideration a bundle of factors that have never been collectively examined before. Its significance lies in the depth of its results, conclusions and empirical implications.

The following section attempts a brief literature review, while sections three and four include the presentation of the proposed conceptual framework and the research hypothesis. In the fifth section, the research methodology is being presented, while the paper is concluded with its main results and conclusions (sections six and seven respectively).

## II. Literature review

Many different definitions of e-Government can be found in the relevant literature. According to Sang, Lee and Lee [4], e-Government is based on ICTs and aims to achieve better governance. Carter and Bélanger [5] and Bekkers [6] argued that e-Government includes initiatives transforming government services and increasing their quality. Moreover, Doong, Wang and Foxall [7] underline the importance of technology in altering citizen behavior, while arguing that e-Government is a competitive tool for achieving higher efficiency of state affairs. However, these advantages can only be achieved if the technology is widely accepted [7]. Therefore, it is important to examine the determinants of acceptance and, hence, present the relevant theoretical models.

978-83-60810-66-8/$25.00©2015, IEEE

### A. Theoretical background

Diffusion of Innovation (DOI): According to DOI theory, the use of technology is a decision based on compatibility, relative advantage, social pressure and communication [8]. Phang, Li, Sutanto and Kankanhalliet [9] observed that the elderly are willing to learn and use new technologies that are accompanied by changes in the structure of the society. This is because the adoption of the technology can enhance the social status of the individual and play a positive role in assisting others to adopt the innovation themselves [2] [8].

Theory of Reasoned Action (TRA): The TRA is one of the first theories explaining the behavior, use and acceptance of computer technology. Three constructs are the main components of TRA: behavioral intention (BI), attitude (A), and subjective norm (SN). According to TRA, a person's behavioral intention depends on his attitude about the behavior, and his subjective norms (BI = A + SN) [11] [12].

Theory of planned behavior (TPB): The TPB has been proposed as an extension of TRA [13]. The elements of behavior and subjective norm are the same in TPB, as in TRA. The difference is that the TPB includes additional determinants of intention (more specifically, perceived behavioral control and self-efficacy). These modifications have increased the explanatory power of TPB.

Technology Acceptance Model (TAM) and TAM2: TAM builds upon the causal relationships of TRA in order to explain the technology acceptance behavior [10]. It suggests that perceived usefulness (PU) and perceived ease of use (PEOU) are the main determinants of technology adoption. TAM2, an extension of the initial model, incorporated additional constructs, such as social influence, subjective norms, voluntarism, social status, as well as cognitive processes, such as labor relevance, quality production and perceived ease of use [2].

Motivational model: Deci [13] focused on intrinsic motivation and argued that external motives do not have a significant persuasive power, in case the individual does not receive pleasure by the use of said technology [14].

Model of PC utilization: Thompson, Higgins and Howellet [16] argued that complexity, long-term benefits, social factors and facilitating conditions are determinants of computer use [15].

Social cognitive theory: Social cognitive theory argues that acquiring individual knowledge is directly related to observing others within the context of social interactions and experiences [4] [17] [18].

Unified Theory of Acceptance and Use of Technology (UTAUT): The UTAUT model was based on all previous theories. In comparison with previous models, UTAUT explains 70% of the variance in technology acceptance, a significant improvement from previous models (that only explained 40% of the same variance) [2]. Hence, UTAUT is considered as an enhanced model with robust features.

### B. Diffusion of innovation and e-Government adoption

The innovation diffusion theory was first systematically tested in the 1940s and has evolved exponentially since then [20]. According to Rogers [8], "*innovation is an idea, practice, or object that is considered new by an individual or another unit of acceptance*". Based on this definition, the use of e-Government services is a new practice and can be considered as an innovation for each individual user.

The diffusion of innovation theory suggests a general model including five groups of recipients, based on how early they start to use a specific innovation [21]. The five categories of recipients are: innovators, early recipients, early majority, slow majority, and laggards [21]. Each group has different characteristics.

Following the categories proposed by the diffusion of innovation theory, people who already use e-Government services can be perceived as early recipients. The diffusion of innovation theory presumes that early recipients share certain characteristics (young age, higher education and income) [21]. Previous studies conducted in the field of public administration have shown that people who use electronic public services fit this description [20] [22].

Atkin, Jeffres and Neuendorfet [23] compared internet adopters and non-adopters, using the following criteria: social status, communication needs, media use, and relation with technology. It was hypothesized that internet adopters will differ from non-adopters in specific demographics (age, education, income). Furthermore, internet adopters would be more cosmopolitan, would have increasing communication needs, and would be more interested in experimenting with new technologies. Indeed, the profile of early adopters showed that they followed the theory of diffusion of innovation, confirming all differences on demographic levels and use of technology [23].

### C. Previous studies

Lean, Zailani, Ramayah and Fernando [2] investigated the factors that influence the intention of Malaysian citizens to use e-government services. They developed a model incorporating factors from both DOI and TAM, while also taking under consideration various cultural factors, as well as trust. Their study included 150 participants. Empirical results revealed that trust, perceived usefulness, perceived relative advantage and perceived image have a direct positive effect on the intention to use e-Government services. On the other hand, perceived complexity was found to have a negative effect on intention. Lean, Zailani, Ramayah and Fernando [2] finally argued that, while online privacy has a positive impact on e-government usage, no significant effect exists between innovation factors (complexity, comparative advantage and image) and intention. Finally, they underlined that DOI has a better explanatory power than all the other employed models [2].

In another similar study, Wangpipatwong, Chutimaskuland and Papasrator [1] explored the factors affecting the intention of citizens to use e-Government websites. They used TAM as the basis of their research model, while the ability to use computers was incorporated as an additional factor affecting intention. A web-based survey was used to empirically test the proposed research model. The participants of the survey were 614, while each of them had (at least) a university degree and experience in visiting e-Government websites. Results revealed that perceived usefulness, perceived ease of use, privacy and the ability to use computers, directly enhance the intention to use e-Government websites [1]. However, it was concluded that different factors have different effect on intention. More specifically, perceived usefulness and ease of use seemed to be more important than the ability to use computers [1].

The effect of perceived risk and perceived trust on the willingness to use e-Government services was examined by Bélanger and Carter [24]. They proposed a model consisting of trust propensity, trust on the internet (TOI), trust on the government (TOG) and perceived risk. Empirical results revealed that trust propensity has a positive impact on both TOI and TOG, which in turn have a positive effect on intention to use e-Government services. Moreover, TOG negatively affects perceived risk, which in turn has a negative impact on intention. The proposed model of Bélanger and Carter [24] is a step towards identifying the unique elements of trust and risk in the e-Government literature [24].

Colesca [25] tried to identify the factors affecting public confidence towards e-government services. The study was conducted in Romania and the sample was 793 citizens. Empirical findings revealed that technical and organizational reliability, perceived quality, perceived usefulness, internet experience and trust propensity directly enhance e-Government usage. On the other hand, age and privacy concerns had a negative impact [25].

In another similar empirical study, Hung, Chang and Yu [26] examined the factors that determined the acceptance of the online submission system of tax declarations and payments (OTFPS) in Taiwan. The authors [26] developed a research model based on the theory of planned behavior. The study was conducted using a survey of 1099 citizens. Results showed that the proposed model explained 72% of the variance in usage behavior. More specifically, user acceptance of OTFPS was influenced by perceived usefulness, perceived ease of use, risk perception, trust, compatibility and peer pressure [26].

Sang, Lee and Lee [4] found that perceived usefulness, relative advantage and trust are significant predictors of e-Government adoption in Cambodia. Moreover, it was discovered that the determinants of perceived usefulness include image and output quality.

Finally, the empirical results provided by Dashti, Benbasat and Burton-Jones [27] demonstrate the significant role of trust on e-Government usage. More specifically, their study provided evidence supporting the relationships between 'felt trust' by e-Government, trust in e-Government, and intention to use e-Government websites [27].

## III. Conceptual framework

The proposed conceptual framework was developed after an extensive review of the existing relevant literature. Previous studies have been thoroughly examined and the most significant factors have been utilized in the proposed framework. Therefore, the present study offers a critical synthesis of existing empirical work on e-Government adoption. It adopts a holistic approach, unifying previous findings into an extensive conceptual framework.

The factors that have been used in the present study are briefly described in the following paragraphs.

### A. Perceived risk

For the past five decades, the theory of perceived risk is an important research agenda of consumer behavior. Perceived risk is defined as a concept of expected subjective injury [30] as well as desirable effect [31]. The literature [32] suggests that six types of perceived risk really exist. These types may vary, depending on the service or product. In the present study five types of risks are being incorporated in the proposed framework:

Performance risk: It concerns the risk of e-Government services not performing as they should be. For example, an electronic failure may result in unexpected losses [33].

Financial risk: According to Kuisma, Laukkanen and Hiltunenet [33] a transaction error or a misuse of the user's private information may bring economic consequences, thus, resulting in fear of conducting any transactions.

Social risk: It is regarded as the possibility of the purchase or use to provoke negative judgment by other members of the society (external psychological risk) [30].

Time risk: It is the risk relating to the user's loss of time during a transaction. The delay in loading the website and a disorganized site may be considered time delay causes [34].

Security risk: It is the risk of losing sensitive personal data during a transaction [35]. A breach of confidentiality is a usual fear during on-line transactions [36].

### B. Trust in e-Government

Trust is a factor that has been extensively investigated and defined in various differently ways. According to Rotter [42], trust is considered as "*the expectancy that the promise of a person or group can be relied upon*". Trust in the context of e-Government is measured through two dimensions: trust in the particular entity (which, in this case, is the government), and trust in the reliability of the enabling technology (which, in this case, is the internet) [24].

### *C. Perceived usefulness*

Perceived usefulness is the degree in which someone believes that using a particular system would enhance his or her job performance [11].

### *D. Perceived ease of use*

According to Davis [11], perceived ease of use can be defined as the degree in which someone believes that using a particular system would be free from effort [40].

### *E. Perceived quality*

Aaker [46] defined perceived quality as the consumer perception of the overall quality or superiority of a product (or a service) in relation to other alternatives. Perceived quality is different from tangible (real) quality, which is embedded in the manufacturing process [46]. No matter how high the level of every product's manufacturing quality, consumers are the ones that will determine its (perceived) quality in the marketplace.

### *F. Quality of internet connection*

Sathye [41] considered quality of the internet connection as one of the significant factors affecting the adoption of various electronic transactions.

### *G. Internet experience*

According to Eastlick and Lotz [37], the history of a user's behavior can positively affect the likelihood of adopting the same behavior in the future. Internet use is positively affected by internet experience [38].

### *H. Computer self - efficacy*

Computer self-efficacy is defined as the perception of an individual concerning his ability to use computers in order to conduct a certain task [45]. People with a high level of computer self-efficacy are able to efficiently use different software packages and computer systems, while those with a low level of computer self-efficacy perceive their potential as quite limited [45].

### *I. Self-image*

Venkatesh and Davis [10] defined 'image' as the degree to which the use of an innovation enhances the social status of an individual. A person is more likely to engage in an activity, if the activity is approved by others [5].

### *J.Peer influence*

Peer influence is the effect of the immediate environment on the behavior of the individual [14]. Peer influence (or peer pressure) encourages others to change their attitudes in order to conform to group norms.

### *K. Intention to use*

'Intention to use' is the main dependent factor of the present study, measuring the intention of citizens to use e-Government services [10].

## IV. Research hypotheses

The present study includes ten independent and one dependent factor. The relationships between these factors are analytically presented in the following paragraphs.

### *A. Perceived risk and intention to use*

Lee [49] has found that security risk (privacy risk) and financial risk have a negative effect on the intention to use online banking. In the same direction, Pires, Stanton and Eckford [50] argued that there is a negative correlation between perceived risk and intention to use web applications. Hence, it is hypothesized:

H1: Perceived risk has a negative effect on the intention to use e-Government services.

### *B. Trust in e-Government and intention to use*

According to the empirical results of a study conducted by Carter and Bélanger [5], trust is a significant predictor of the intention to use e-government services. In another similar study, the same authors [24] found out that disposition to trust positively affects trust towards the internet and trust towards the government, which in turn affect the intentions to use e-Government services. Dashti, Benbasat and Burton-Jones [27] also showed that trust for e-Government is an important factor that increases its use.

H2: Trust in e-Government has a positive effect on the intention to use e-Government services.

### *C. Perceived usefulness and intention to use*

The relationship between perceived usefulness and intention to use has been widely acknowledged by the empirical literature [20, 23, 47]. Davis [11] revealed that perceived usefulness is a key factor in the intention to use web system applications. It is only logical to assume that:

H3: Perceived usefulness has a positive effect on the intention to use e-Government services.

### *D. Perceived ease of use and intention to use*

As in the case of perceived usefulness, perceived ease of use has been found to have a positive effect on intention to use [2, 4, 10]. For example, Gefen and Straub [40] and Davis [11] argued that perceived ease of use is positively related to the use of present and future technology. Based on the above, it is hypothesized that:

H4: Perceived ease of use has a positive effect on the intention to use e-Government services.

### *E. Perceived quality and intention to use*

Colesca [25] suggests that the higher the perception of quality of an electronic service, the higher the level of trust towards that service and, consequently, the intention to use it. Moreover, Gronier and Lambert [48] argue that e-Government quality enables citizens to directly locate the services they need, thus, achieving higher levels of satisfaction and effectiveness. Therefore, it is hypothesized:

H5: Perceived quality has a positive effect on the intention to use e-Government services.

### *F. Quality of internet connection and intention to use*

Sathye [41] argued that internet access is one of the most important factors for the acceptance of electronic banking. Without a proper internet connection it is impossible to use electronic services [41]. Moreover, Pikkarainen, Pikkarainen, Karjaluoto and Pahnilaet [51] found out that a decent internet connection has significant influence on the intention to use e-Government services.

H6: Quality of the internet connection has a positive effect on the intention to use e-Government services.

### *G. Internet experience and intention to use*

Based on their empirical results, Corbitt, Thanasankit and Yi [52] argued that prior internet experience has a positive influence on the intention to shop online. Cho [53] reached the same conclusion, stating that experienced internet users are more likely to conduct an online transaction when they feel the urge to do so. It would be interesting to examine whether that is also the case with e-Government services.

H7: Internet experience has a positive effect on the intention to use e-Government services.

### *H. Computer self-efficacy and intention to use*

Wangpipatwong, Chutimaskul and Papasrator [1] empirically confirmed that the adoption of e-government depends upon the ability to use a computer.

H8: Computer self-efficacy has a positive effect on the intention to use e-Government services.

### *I. Self-image and intention to use*

Sang, Lee and Lee [4] have found out that individuals adopt an innovation in order to impress others that have not adopted it. On the other hand, Carter and Bélanger [5] failed to establish such a relationship, reporting that high levels of self-image do not directly affect the intention to use e-Government services. Therefore, it would be interesting to examine which is the case, adopting the hypothesis that:

H9: Self-image has a positive effect on the intention to use e-Government services.

### *J.Peer influence and intention to use*

According to Hung, Chang and Yu [26], peer influence is a key factor in using online tax applications. Moreover, Taylor and Todd [54] support that peer influence is an important determinant of Information System (IS) adoption. Since e-Government has the characteristics of these systems, it is only logical to hypothesize that:

H10: Peer Influence has a positive effect on the intention to use e-Government services.

Fig. 1 summarizes the above hypotheses, thus, presenting the proposed Conceptual Framework of the study.

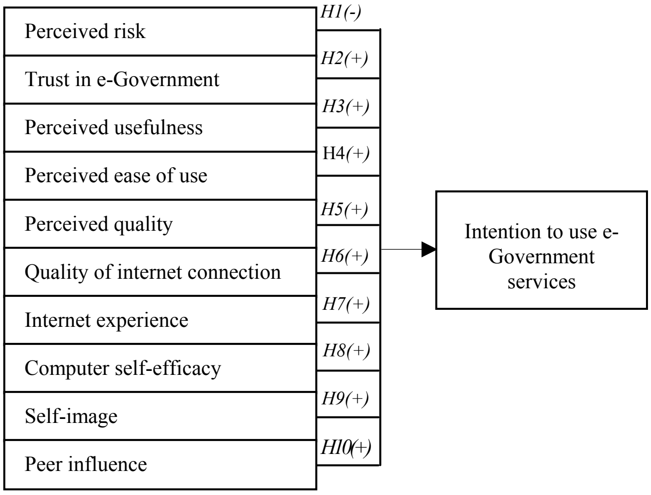


Fig. 1 The proposed conceptual framework of the study

## V. Research Methodology

### *A. The population of the study*

The proposed conceptual framework was tested with the use of a newly-developed structured questionnaire on a sample of Greek internet users. Hence, internet users living in Greece are the population of the present study. According to reliable data [58], internet users in Greece have reached 6.451.326 in December of 2013 (55,9% of the population).

Since the main dependent factor of the study is "intention to use e-Government services", respondents could be: (a) inexperienced e-Government users (non-users), (b) one-time e-Government users (not continual users), or (c) continual users. The sample of the study included individuals from the last two categories.

### *B. Measurement*

The measurement of each of the eleven factors of the proposed conceptual framework was conducted with the use of multiple questions (items) that were adopted from the international literature. The five point Likert scale was used for the measurement of all factors. Table I demonstrates the eleven factors of the study, the number of items used in each case and the studies from which they where adopted.

### *C. Data collection*

Primary data were collected from a random sample of internet users. The only criterion for citizens to participate in the survey was the use of internet. Personal interviews were conducted in order to collect the appropriate data. Citizens were selected in random (using the systematic sampling approach). The research period lasted two months (May to July 2013). Totally, 566 questionnaires were returned and 547 were used for data analysis.

TABLE I. FACTOR MEASUREMENT

| Factors | Items | Source |
|---|---|---|
| Perceived risk | 11 | 48, 49 |
| Trust in e-Government | 15 | 4, 23, 26 |
| Perceived usefulness | 4 | 10, 46, 55 |
| Perceived ease of use | 4 | 10, 39, 46 |
| Perceived quality | 4 | 24 |
| Quality of internet connection | 4 | 40, 50 |
| Internet experience | 3 | 51, 52 |
| Computer self-efficacy | 3 | 1 |
| Self-image | 4 | 3, 4 |
| Peer influence | 4 | 25 |
| Intention to use | 4 | 46, 54 |
| Total | 60 | |

### D. Reliability and validity

The instrument (questionnaire) that was used in the present study was tested for both its content and construct validity. The test for the content validity was conducted using a pilot study approach. Twenty (20) citizens were asked to fill in the final draft of the questionnaire and make comments concerning their level of understanding. Citizens' comments improved various aspects of the questionnaire (e.g. use of language).

For the control of the construct validity, each of the eleven research factors was evaluated: (a) for its unidimensionality and reliability (Table II), (b) for its goodness of fit to the proposed research framework (Table III).

The examination of the unidimensionality of each of the research factors was conducted using Explanatory Factor Analysis (EFA). Moreover, for the estimation of the reliability of these factors, Cronbach Alpha was used. All tests concluded that the scales used are valid and reliable (see Table II above for the main results).

The evaluation of the goodness of fit of each research factor to the proposed model was conducted using Confirmatory Factor Analysis (CFA). All tests produced satisfactory results (see Table III above for the main results).

It should be noted that Second Order CFA was conducted for the five dimensions of risk (performance risk, financial risk, social risk, time risk, security risk). All statistical measures extracted from this analysis where within satisfactory levels.

TABLE II. ESTIMATION OF UNIDIMENSIONALITY AND RELIABILITY

| Factors | KMO | Bartlett's Test | Eigen-value | TVE | Cronbach Alpha |
|---|---|---|---|---|---|
| Perceived risk | 0,82 | 1843,2[a] | 1,201 | 59,1% | 0,785 |
| Trust in e-Government | 0,70 | 595,9[a] | 2,195 | 73,1% | 0,817 |
| Perceived usefulness | 0,81 | 1624,9[a] | 3,174 | 79,3% | 0,913 |
| Perceived ease of use | 0,81 | 1680,9[a] | 3,181 | 79,5% | 0,914 |
| Perceived quality | 0,79 | 935,6[a] | 2,701 | 67,5% | 0,839 |
| Quality of internet connection | 0,80 | 1102,9[a] | 2,884 | 72,0% | 0,871 |
| Internet experience | 0,69 | 466,5[a] | 2,076 | 69,2% | 0,773 |
| Computer self-efficacy | 0,71 | 751,6[a] | 2,309 | 76,9% | 0,850 |
| Self-image | 0,73 | 1191,6[a] | 2,548 | 84,9% | 0,911 |
| Peer influence | 0,70 | 511,7[a] | 2,128 | 70,9% | 0,791 |
| Intention to use | 0,72 | 674,1[a] | 2,265 | 75,4% | 0,837 |

[a] p<0,01

TABLE III. ESTIMATION OF THE GOODNESS OF FIT

| Factors | Normed $X^2$ | C.R. | V.E. | RMSEA | CFI / GFI |
|---|---|---|---|---|---|
| Perceived risk | 2,49 | 0,84 | 78,8% | 0,091 | 0,99 / 0,97 |
| Trust in e-Government | 3,67 | 0,73 | 67,3% | 0,093 | 0,93 / 0,95 |
| Perceived usefulness | 3,34 | 0,61 | 55,6% | 0,089 | 0,94 / 0,96 |
| Perceived ease of use | 2,32 | 0,77 | 71,7% | 0,077 | 0,97 / 0,97 |
| Perceived quality | 2,53 | 0,64 | 75,1% | 0,094 | 0,99 / 0,99 |
| Quality of internet connection | 4,11 | 0,82 | 69,4% | 0,091 | 0,91 / 0,93 |
| Internet experience | 2,43 | 0,83 | 78,7% | 0,079 | 0,93 / 0,99 |
| Computer self-efficacy | 3,43 | 0,77 | 69,5% | 0,082 | 0,99 / 0,99 |
| Self-image | 2,93 | 0,68 | 73,7% | 0,099 | 0,91 / 0,94 |
| Peer influence | 1,97 | 0,84 | 69,3% | 0,095 | 0,97 / 0,99 |
| Intention to use | 3,43 | 0,88 | 89,2% | 0,084 | 0,93 / 0,97 |

## VI. RESULTS

### A. Basic measures

The examination of the proposed conceptual framework was conducted with the use of the "Structural Equation Modeling" (SEM) technique [59] [60].

To evaluate the fit of the (modified) overall model the chi-square value ($X^2$ = 99,9 with 26 degrees of freedom) and the p-value (p = 0,0356) were estimated. These values indicate a satisfactory fit of the data to the overall model. However, the sensitivity of the $X^2$ statistic to the sample size requires the use of other supplementary measures of evaluating the overall model, such as the “Normed-$X^{2}$” index (3,84), the RSMEA index (0,071) the CFI (0,94) and the GFI (0,96), that all indicate a very good fit. Additionally, the Construct Reliability (C.R.) and the Variance Extracted (V.E.) for all factors (constructs) are satisfactory.

### *B. Hypothesis testing*

As it can be seen on Table IV, six of the originally hypothesised paths were found significant (H1, H2, H3, H7, H8, H10), while thirteen new paths (presented on Figure II with dashed lines) were added to the model, based on the modification indexes function of AMOS. Moreover, two factors (quality of internet connection and self-image) were completely discarded from the model, while for four hypotheses (H4, H5, H6, H9) no statistical support was provided (although H4 is supported when indirect effects are considered).

The above modifications resulted in a modified structural model with improved fit (as already explained above). In more detail, the factors that are included in the final model can explain 52% of the variance in the dependent variable, i.e. intention to use e-Government services.

The modified conceptual framework presents an interesting view of the subject under consideration, arguing that every factor is significant in enhancing e-Government usage, either directly or indirectly. For example, perceived ease of use does not have a direct impact on intention, but it has an indirect effect, through perceived risk and perceived usefulness. On the other hand, perceived usefulness has both a direct and indirect effect on intention. Moreover, perceived quality does not seem to have any effect on intention whatsoever.

### *C. Further analysis*

The following observations can be made after reviewing the empirical results of the study (see Table IV and Fig. 2):

- Perceived usefulness directly affects intention. This is also confirmed by the Technology Acceptance Model (TAM) of Davis [11]. Moreover, this impact is not only statistical significant, but is also quite strong (r=0,51).

- Additionally, trust in e-Government has a significant direct effect on intention. The same conclusion has been drawn by Bélanger and Carter [24] and Colesca [25]. According to these authors [24] [25], when the user actually believes that the transaction and his personal data are secure, it is more likely to experience a higher level of intention to use the online services.

TABLE IV. RESULTS OF THE MODIFIED STRUCTURAL MODEL

| Causal Paths (hypotheses) | | Estimate | p | Result |
|---|---|---|---|---|
| H1 | Perceived risk → Intention to use | -0,10 | 0,000 | Accepted |
| H2 | Trust in e-Government → Intention to use | 0,25 | 0,000 | Accepted |
| H3 | Perceived usefulness → Intention to use | 0,51 | 0,000 | Accepted |
| H7 | Internet experience → Intention to use | 0,13 | 0,000 | Accepted |
| H8 | Computer self-efficacy → Intention to use | 0,10 | 0,000 | Accepted |
| H10 | Peer influence → Intention to use | 0,12 | 0,000 | Accepted |
| **New causal paths** | | | | |
| | Perceived ease of use → Perceived risk | -0,22 | 0,000 | New paths |
| | Perceived ease of use → Perceived usefulness | 0,42 | 0,000 | |
| | Perceived ease of use → Perceived quality | 0,13 | 0,000 | |
| | Computer self-efficacy → Perceived ease of use | 0,26 | 0,000 | |
| | Trust in e-Government → Perceived risk | -0,20 | 0,000 | |
| | Trust in e-Government → Perceived quality | 0,42 | 0,000 | |
| | Perceived usefulness → Trust in e-Government | 0,25 | 0,000 | |
| | Perceived usefulness → Perceived quality | 0,22 | 0,000 | |
| | Internet experience → Perceived ease of use | 0,61 | 0,000 | |
| | Internet experience → Computer self-efficacy | 0,76 | 0,000 | |
| | Internet experience → Perceived risk | -0,10 | 0,000 | |
| | Peer influence → Perceived usefulness | 0,16 | 0,000 | |

- Internet experience directly affects intention to use e-Government services. Cho [53] reached the same conclusion, arguing that repeated internet use for extended periods of time positively affects user behavior and system usage.
- Perceived risk has a negative effect on intention. This conclusion is in line with the studies of Lee [49] and Featherman and Pavlou [31]. For example, a loss of communication between user and server may have a negative impact on user trust and, therefore, in the intention to further use the online services.
- Peer influence positively affects intention. The same conclusion was reached by Hung, Chang and Yu [26]. A possible ‘reward’ from the immediate social environment predisposes the user towards using e-Government services.
- Moreover, computer self-efficacy has a direct positive effect on intention. Wangpipatwong, Chutimaskul and Papasrator [1] argue that people with increased computer

self-efficacy easily understand the functions of online services, thus increasing their confidence and their general intention to use such systems.

- In the same line as above, internet experience was found to have a quite strong impact on computer self-efficacy (r=0,76). It seems that experienced internet users believe that they are more capable of using their computer. Hence, internet experience affects intention both directly (r=0,13) and indirectly (through computer self-efficacy and perceived risk).
- Additionally, internet experience has both a direct and indirect effect on perceived ease of use. The same conclusion was reached by Cheong and Park [55]: the more time a user spends on the internet, the better his perception about the usability of the online platform.
- Perceived ease of use was found to have a direct positive effect on perceived usefulness, something that has been supported by, almost, every TAM study [9] [11] [49].
- The present study found that perceived usefulness has a positive impact on trust. In the same direction, Suh and Han [56] argued that when the user fully understands the usefulness of the online service, his confidence is being significantly increased.
- Furthermore, perceived ease of use was found to have a direct negative effect on perceived risk. This relationship was also identified by Hung, Chang and Yu [26], who concluded that users finding an online system easy to operate, believe that the overall risk of the transaction is quite small.
- On the same vein, perceived usefulness was found to have a negative effect on perceived risk. According to Bélanger and Carter [24], when a system is considered to be useful, it is also usually considered to be free of various risks.

## VII. Conclusions

### *A. General conclusions*

The present study developed a novel research framework in order to identify the main factors affecting the intention to use e-Government services. This framework (research model) was tested on a sample of 547 internet users.

The main conclusion of the present study does not concern a certain research factor. On the contrary, the empirical results support that both research and empirical focus should be drawn towards managing a bundle of factors (dimensions), which seem to be highly interrelated. The statistical analysis that was conducted highlighted various new causal relationships between the independent factors of the study. Therefore, one should not only pay attention to enhancing a limited number of factors, since the existence of various indirect effects underlined the need for a more integrated approach.

### *B. e-Government user profile*

The average user has the following characteristics: (a) is under the age of 40, (b) is quite educated, (c) his monthly income is satisfactory, (d) uses e-Government services for more than a year, (e) uses e-Government services in order to gather information and conduct transactions, (f) makes about five transactions each month.

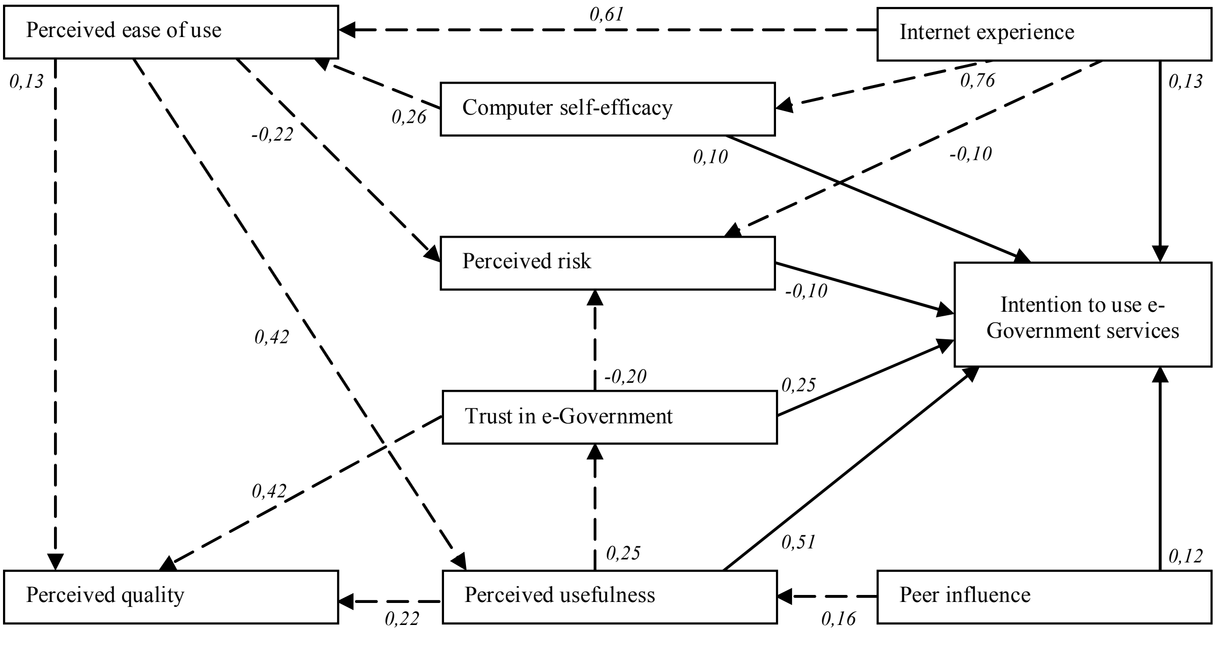


Fig 1. The modified conceptual framework of the study.

### C. *Managerial implications*

The empirical results showed that computer self-efficacy is positively related with the intention to use e-Government services. That finding is in line with Compeau and Higgins [45] and Wangpipatwong, Chutimaskul and Papasrator [1]. State officials should make efforts to fight computer illiteracy (seminars, incentives, etc), since only self efficient citizens will adopt e-Government.

Moreover, since trust is an essential element for the adoption of e-government, citizens should be convinced that state mechanisms are in place in order to protect their privacy and transactions in an impersonal medium, such as the internet. According to Dashti, Benbasat and Burton-Joneset [27] those responsible for public electronic services, should implement mechanisms for increasing public trust (receiving and answering to questions / comments, offer support via the phone, implement safety protocols, etc).

Additionally, according to Hung, Chang and Yu [26], external resources, such as television / news, significantly influence early-adopters and non-adopters of e-Government. By providing proper information to citizens (through the internet or in the form of leaflets), the state can bend the bias that exists towards online services [61] [62].

Finally, special focus should be given in enhancing the level of perceived usefulness [57]. Citizens are more interested in usefulness, as the effect of perceived ease of use is often reduced after a short period of time. In that direction, e-Government services should add value to citizens, offering access to useful information. There is no meaning in developing services that offer very little information and utility, while at the same time discourage citizens from future use. A useful application will attract the attention of the public and create more users in the near future.

### D. *Limitations and future research*

The present research used self-reported scales in order to measure the constructs of the proposed framework. This is a limitation inherent to all explanatory studies. Moreover, the sample of the study is national (Greece), while, on the other hand, an international sample would provide results that could be generalized to the international context. Finally, future studies could take the empirical results of the present study into consideration and further enhance its proposed conceptual framework.

## References

[1] S. Wangpipatwong, W. Chutimaskul, and B. Papasrator, “Understanding Citizen’s Continuance Intention to Use e-Government Website: a Composite View of Technology Acceptance Model and Computer Self-Efficacy”, *Electronic Journal of e-Government*, Vol. 6, No. 1, pp. 55-64, 2008.

[2] O.K. Lean, S. Zailani, T. Ramayah, and Y. Fernando, “Factors influencing intention to use e-government services among citizens in Malaysia”, *International Journal of Information Management*, Vol. 29, No. 6, pp. 458-475, 2009, http://dx.doi.org/10.1016/j.ijinfomgt. 2009.03.012.

[3] E Ziemba, T. Papaj, and M. Jadamus-Hacura, “Critical success factors for adopting state and local e-government polish insights”, in Proc. *13th International Conference e-Society 2015*, Portugal, 2015, pp. 95–102.

[4] S. Sang, J.-D. Lee, and J. Lee, “E-government adoption in ASEAN: the case of Cambodia”, *Internet Research*, Vol. 19, No. 5, pp. 517-534, 2009, http://dx.doi.org/10.1108/10662240910998869.

[5] L. Carter, and F. Bélanger, “The utilization of e-government services: citizen trust, innovation and acceptance factors”, *Information Systems Journal*, Vol. 15, No. 1, pp. 5-25, January 2005, http://dx.doi.org/10.1111/j.1365-2575.2005.00183.x.

[6] V. Bekkers, “E-government and the emergence of virtual organizations in the public sector”, *Information Policy*, Vol. 8, pp. 89-101, 2003.

[7] H.-S. Doong, H.-C. Wang, and G. Foxall, “Psychological traits and loyalty intentions towards e-Government services”, *International Journal of Information Management*, Vol. 30, No. 5, pp. 457-464, 2010, http://dx.doi.org/10.1016/j.ijinfomgt.2010.01.007.

[8] M. Rogers, *Diffusion of innovations*, 5th ed., New York: the Free Press, 2003.

[9] C.W. Phang, Y. Li, J. Sutanto, and A. Kankanhalli, “Senior citizens’ adoption of E-government: in quest of the antecedents of perceived usefulness”, HICSS 2005, *Proceedings of the 38th Annual Hawaii International Conference*, IEEE, pp. 130a-130a, 2005, http://dx.doi.org/10.1109/HICSS.2005.538.

[10] V. Venkatesh and F. Davis, “A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies”, *Management Science*, Vol. 46, No. 2, pp. 186-204, February 2000, http://dx.doi.org/10.1287/mnsc.46.2.186.11926.

[11] F.D. Davis, “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology”, *MIS Quarterly*, pp. 319-340, September 1989, http://dx.doi.org/10.2307/249008.

[12] M. Fishbein and I. Ajzen, *Belief, attitude, intention and behaviour: An introduction to theory and research*, Massachusetts: Wesley, 1975.

[13] E.L. Deci, *Intrinsic motivation*, New York: Plenum Press, 1975.

[14] I. Ajzen, “Nature and operation of attitudes”, *Annual Review of Psychology*, Vol. 52, No. 1, pp. 27-58, 2001, http://dx.doi.org/10.1146/annurev.psych.52.1.27.

[15] H.C. Triandis, *Attitude and Attitude Change*, New York: John Wiley, 1971.

[16] R.L. Thompson, C.A. Higgins, and J.M. Howell, “Personal computing: toward a conceptual model of utilization”, *MIS Quarterly*, pp. 124-143, 1991, http://dx.doi.org/10.2307/249443.

[17] A. Bandura, “Self-efficacy mechanism in human agency”, *American Psychologist*, Vol. 37, No. 2, pp. 122-147, 1982, http://dx.doi.org/10.1037/0003-066X.37.2.122.

[18] A. Bandura, *Social foundations of thought and action: A social cognitive theory*, Englewood Cliffs: Prentice Hall, 1986.

[19] J. Hiller, and F. Belanger, *Privacy Strategies for Electronic Government*, North America: Rowman and Littlefield Publishers, 2001.

[20] D.V. Dimitrova, and Y.-C. Chen, “Profiling the Adopters of E-Government Information and Services: The Influence of Psychological Characteristics”, *Social Science Computer Review*, Vol. 24, No. 2, pp. 172-188, 2006, http://dx.doi.org/10.1177/0894439305281517.

[21] E.M. Rogers, *Diffusion of innovations*, 4th ed., New York: The Free Press, 1995.

[22] J.C. Thomas, and G. Streib, “The new face of government: Citizen-initiated contacts in the era of e-Government”, *Journal of Public Administration Research and Theory*, Vol. 13, No. 1, pp. 83-102, 2003, http://dx.doi.org/10.1093/jpart/mug010.

[23] D.J. Atkin, L.W. Jeffres, and K. Neuendorf, “Understanding Internet Adoption as Telecommunications Behavior”, *Journal of Broadcasting & Electronic Media*, Vol. 42, No. 4, pp. 475-490, 1998, http://dx.doi.org/10.1080/08838159809364463.

[24] F. Bélanger, and L. Carter, “Trust and risk in e-government adoption”, *The Journal of Strategic Information Systems*, Vol. 17, No. 2, pp. 165-176, June 2008, http://dx.doi.org/10.1016/j.jsis.2007.12.002.

[25] S. Colesca, “Increasing e-trust: a solution to minimize risk in the e-government adoption”, *Journal of applied quantitative methods*, Vol. 4, No. 1, pp. 31-44, 2009.

[26] S.-Y. Hung, C.-M. Chang, and T.-J. Yu, “Determinants of user acceptance of the e-Government services: The case of online tax filing and payment system”, *Government Information Quarterly*, Vol. 23, No. 1, pp. 97-122, 2006, http://dx.doi.org/10.1016/j.giq.2005.11.005.

[27] A. Dashti, I. Benbasat, and A. Burton-Jones, “Developing trust reciprocity in electronic government: The role of felt trust”, *Proceedings of the European and Mediterranean Conference on Information Systems*, Izmir, Turkey , pp. 1-13, 2009.

[28] M.T. Frolich, and R. Dixon, “A taxonomy of manufacturing strategies revisited”, *Journal of Operations Management*, Vol. 19, No. 5, pp. 541-558, 2001, http://dx.doi.org/10.1016/S0272-6963(01)00063-8.

[29] R. Hubbard, D. Vetter, and E. Little, “Replication in strategic management: scientific testing for validity, generalizability, and usefulness”, *Strategic Management Journal*, Vol. 19, No. 3, pp. 243-254, 1998, http://dx.doi.org/10.1002/(SICI)1097-0266(199803)19:3<243::AID-SMJ951>3.0.CO;2-0.

[30] J.P. Peter, and M.J. Ryan, “An Investigation of Perceived Risk at the Brand Level”, *Journal of Marketing Research*, pp. 184-188, May 1976, http://dx.doi.org/10.2307/3150856.

[31] M. Featherman, and P. Pavlou, “Predicting e-services adoption: a perceived risk facets perspective”, *International Journal of Human-Computer Studies*, Vol. 59, No. 4, pp. 451-474, 2003, http://dx.doi.org/10.1016/S1071-5819(03)00111-3.

[32] J. Jacoby, and L.B. Kaplan, “The Components of Perceived Risk”, *Proceedings of the Third Annual Conference of the Association for Consumer Research*, pp. 382-393, 1972.

[33] T. Kuisma, T. Laukkanen, and M. Hiltunen, “Mapping the reasons for resistance to Internet banking: A means-end approach”, *International Journal of Information Management*, Vol. 27, No. 2, pp. 75-85, 2007, http://dx.doi.org/10.1016/j.ijinfomgt.2006.08.006.

[34] S. Forsythe, and B. Shi, “Consumer patronage and risk perceptions in Internet shopping”, *Journal of Business Research*, Vol. 56, No. 11, pp. 867-875, 2003, http://dx.doi.org/10.1016/S0148-2963(01)00273-9.

[35] N. Reavley, “Securing online banking”, *Card Technology Today*, pp. 12-13, October 2005, http://dx.doi.org/10.1016/S0965-2590(05)70389-3.

[36] D. Littler and D. Melanthiou, “Consumer perceptions of risk and uncertainty and the implications for behaviour towards innovative retail services: The case of Internet Banking”, *Journal of Retailing and Consumer Services*, Vol. 13, No. 6, pp. 431-443, 2006, http://dx.doi.org/10.1016/j.jretconser.2006.02.006.

[37] M.A. Eastlick, and S. Lotz, “Profiling potential adopters and non-adopters of an interactive electronic shopping medium”, *International Journal of Retail & Distribution Management*, Vol. 27, No. 6, pp. 209-223, 1999, http://dx.doi.org/10.1108/09590559910278560.

[38] G. Lohse, S. Bellman, and E. Johnson, “Consumer buying behavior on the Internet: findings from panel data”, *Journal of Interactive Marketing*, Vol. 14, No. 1, pp. 15-29, 2000, http://dx.doi.org/10.1002/(SICI)1520-6653(200024)14:1<15::AID-DIR2>3.0.CO;2-C.

[39] F. Heider, *The Psychology of Interpersonal Relations*, New York: Wiley, 1958.

[40] D. Gefen and D. Straub, “The Relative Importance of Perceived Ease of Use in IS Adoption: A Study of E-Commerce Adoption”, *Journal of the Association for Information Systems*, Vol. 1, No. 1, Article 8, 2000.

[41] M. Sathye, “Adoption of internet banking by Australian consumers: an empirical investigation”, *International Journal of Bank Marketing*, Vol. 17, No. 7, pp. 324-334, 1999, http://dx.doi.org/10.1108/02652329910305689.

[42] J. Rotter, “Generalized expectancies for interpersonal trust”, *American Psychologist*, Vol. 26, No. 5, pp. 443-452, May 1971, http://dx.doi.org/10.1037/h0031464.

[43] Y.-H. Tan, and W. Thoen, “Toward a Generic Model of Trust for Electronic Commerce”, *International Journal of Electronic Commerce*, Vol. 5, No. 2, pp. 61-74, Winter 2001, http://dx.doi.org/10.1080/10864415.2000.11044201.

[44] D.H. McKnight, V. Choudhury, and C. Kacmar, “Developing and Validating Trust Measures for e-Commerce: An Integrative Typology”, *Journal Information Systems Research*, Vol. 13, No. 3, pp. 334-359, September 2002.

[45] D.R. Compeau, and C.A. Higgins, “Computer self-efficacy: development of a measure and initial test”, *MIS Quarterly*, pp. 189-211, June 1995, http://dx.doi.org/10.2307/249688.

[46] D. Aaker, *Managing brand equity: Capitalizing on the value of a brand name*, New York, 1991.

[47] T.E. Cheng, D.Y. Lam, and A.C. Yeung, “Adoption of internet banking: an empirical study in Hong Kong”, *Decision Support Systems*, Vol. 42, No. 3, pp. 1558-1572, December 2006, http://dx.doi.org/10.1016/j.dss.2006.01.002.

[48] G. Gronier, and M. Lambert, “A model to measure the perceived quality of service in e-government”, *Public Research Centre Henri Tudor*, Luxembourg, 2010.

[49] M.-C. Lee, “Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit”, *Electronic Commerce Research and Applications*, Vol. 8, No. 3, pp. 130-141, June 2009, http://dx.doi.org/10.1016/j.elerap.2008.11.006.

[50] G. Pires, J. Stanton, and A. Eckford, “Influences on the perceived risk of purchasing online”, *Journal of Consumer Behaviour*, Vol. 4, No. 2, pp. 118131, December 2004, http://dx.doi.org/10.1002/cb.163.

[51] T. Pikkarainen, K. Pikkarainen, H. Karjaluoto, and S. Pahnila, “Consumer acceptance of online banking: an extension of the technology acceptance model”, *Internet Research*, Vol. 14, No. 3, pp. 224-235, 2004, http://dx.doi.org/10.1108/10662240410542652.

[52] B.J. Corbitt, T. Thanasankit, and H. Yi, “Trust and e-commerce: a study of consumer perceptions”, *Electronic Commerce Research and Applications*, Vol. 2, No. 3, pp. 203-215, Autumn 2003, http://dx.doi.org/10.1016/S1567-4223(03)00024-3.

[53] J. Cho, “Likelihood to abort an online transaction: influences from cognitive evaluations, attitudes, and behavioral variables”, *Information & Management*, Vol. 41, No. 7, pp. 827-838, September 2004, http://dx.doi.org/10.1016/j.im.2003.08.013.

[54] S. Taylor, and P.A. Todd, “Understanding Information Technology Usage: A Test of Competing Models”, *Information Systems Research*, Vol. 6, No. 2, pp. 144-176, 1995, http://dx.doi.org/10.1287/isre.6.2.144.

[55] J.H. Cheong, and M.-C. Park, “Mobile internet acceptance in Korea”, *Internet Research*, Vol. 15, No. 2, pp. 125-140, 2005, http://dx.doi.org/10.1108/10662240510590324.

[56] B. Suh, and I. Han, “Effect of trust on customer acceptance of Internet banking”, *Electronic Commerce Research and Applications*, Vol. 1, No. 3, pp. 247-263, 2002, http://dx.doi.org/10.1016/S1567-4223(02)00017-0.

[57] V. Venkatesh, M.G. Morris, G.B. Davis, and F.D. Davis, “User Acceptance of Information Technology: Toward a Unified View”, MIS Quarterly, pp. 425-478, September 2003.

[58] Internetworldstats.com, European Union Statistics of Internet Usage, Accessed at 12-2-2015 from: http://www.internetworldstats.com/ europa.htm#gr.

[59] E.K. Kelloway, *Using LISREL for Structural Equation Modeling: A Researcher's Guide*, Thousand Oaks, CA: Sage, 1998.

[60] F. Hair, R. Anderson, R. Tatham, and W. Black, *Multivariate Data Analysis with Readings*, Prentice-Hall International: London, 1995.

[61] S. Alghamdi, and N. Beloff, “Towards a comprehensive model for e-Government adoption and utilisation analysis: The case of Saudi Arabia”, in Proc. *2014 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Warsaw, 2014, pp. 1217-1225, http://dx.doi.org/10.15439/2014F146.

[62] E. Ziemba, T. Papaj, and D. Descours, “Assessing the quality of e-government portals-the Polish experience”, in Proc. *2014 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Warsaw, 2014, pp. 1259-1267, http://dx.doi.org/10.15439/ 2014F121.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1499–1504

DOI: 10.15439/2015F43
ACSIS, Vol. 5

# Comparative Analysis of Electronic Banking Websites in Selected Banks in Poland in 2014

Witold Chmielarz
University of Warsaw, Faculty of Management,
ul. Szturmowa 1/3, 02-678 Warsaw, Poland
Email: witold@chmielarz.eu

Marek Zborowski
University of Warsaw, Faculty of Management,
ul. Szturmowa 1/3, 02-678 Warsaw, Poland
Email: mzborowski@wz.uw.edu.pl

***Abstract*—The main objective of this article is to identify the best e-banking websites in Poland from the point of view of an individual customer. Using modern IT tools for communications with customers of banking services, banks create competitive advantages, as well as, opportunities for providing banking services in a convenient way for consumers. After a short introduction the authors defines the assumptions for the study. The methodological approach—based on theoretical and empirical study in the field of e-banking, allows them to build the evaluation model for the construction of high quality e-banking website. Subsequently, the authors carried out multilateral analyses and presented the conclusions of the study. The identified categories are classified into three groups: economic, technological and anti-crisis. The originality of the work comes down to knowledge of the determinants of customer's quality perception of websites and a starting point for an effective quality management of their e-services system.**

## I. Introduction

IT APPEARS that the consequences of the worldwide crisis in electronic banking in Poland strengthen the tendencies which show that the crisis, which started in the second half of 2008, does not concern this area. Compared to 2013, the number of individual clients with potential access to account increased almost by 15% (9% more than in 2012) reaching over 25 million users; the number of active individual clients went up by over 5% reaching 13.060 million [12]. Undoubtedly it is the fastest growing banking sector and—as indicated in earlier articles—nothing points to the fact that something may undermine these positive trends. The increase in absolute numbers of clients is shown in Fig. 1. The increase in the number of clients with potential access to account via the Internet is accompanied by a continuous increase of active customers (at least one transaction a month). Since the end of 2008 till the end of 2014 the number increased by over 13 million users, which is an increase by 123%. Every year the population of new customers using the possibilities offered by the Internet to handle banking transactions is growing. In 2008 more than 890,000 people started to use e-banking services, and in 2014 the number was more than 3 million. There are nearly 53% of active users, out of all clients having electronic access to account.

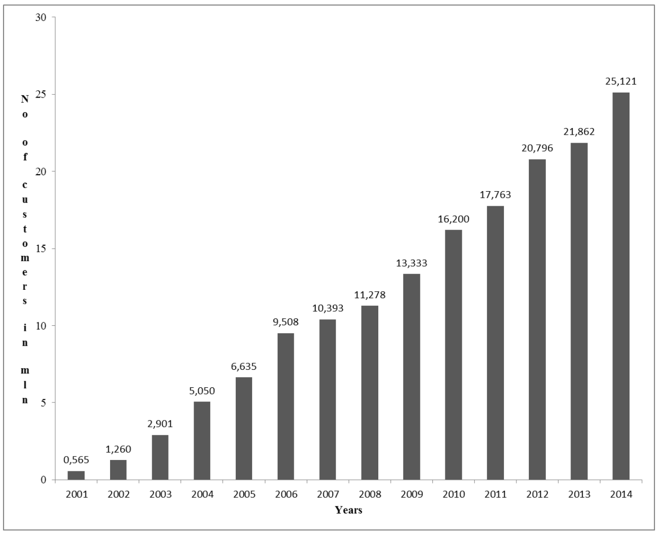


Fig. 1. The evolution of the number of clients with electronic access to account in 2001-2014 in millions [12]

Poland in the European statistics—with regard to penetration—compares quite well—according to ComScore [7] report—it takes the sixth place (52.3%), while the European average is 40%. The largest e-banking penetration is in the Netherlands—66%, the lowest in Switzerland (18.8%). France (60%), Finland (56.4%), Sweden (54.2%) are ahead of us, behind are among others: Germany, Spain, Denmark and Norway. The dynamics of the increase of the number of e-banking clients in Poland is still one of the highest on our continent—in recent years we note the increase of over one million every year. So, this is very important area of e-services, worth to conduct research.

There are many publications concerning the issue of evaluation of websites [3], [9], [13], [14], [15] and access to e-banking services [1], [10], [11] but there is no easy solution to the problems encountered [12], [6], [16]. The review of the literature shows that e-banking websites (as well as e-commerce websites) may be analysed from the point of view of:

- usability (site map, directory),
- interactivity (availability and responsiveness),
- functionality (search, navigation, relevance of content),
- visualisation (colour scheme, background, graphics, text),
- efficiency (cost of purchase, transport, the difference in prices offered by traditional and online shops),

978-83-60810-66-8/$25.00©2015, IEEE

- reliability and availability.

Most of evaluation methods are traditional scoring methods based on specific criteria sets, evaluated by means of an applied scale. Technical and functional criteria are the most commonly applied. Most of them contain factors which may be evaluated in a very subjective way: text clarity, attractive colours, images and pictures, the speed and intuitiveness of navigation, etc. Moreover, some users do not treat particular criteria sets in an equivalent way. However, on the other hand, there occur frequent problems with determining preferences for particular criteria and the evaluation of relations between them. This part of the work concerns the application of the authors own, though based on the literature, set of criteria for a scoring evaluation and a selection of electronic services of selected banks.

We are making an comparative analysis mainly in three cases, enabling:

- specification and accurate research into the area in which the software works,
- creating a ranking of IT solutions existing on the market,
- identification of the features which make particular solutions better than others.

Here we are concentrating on the third case.

## II. Assumptions of the Study

At the first half of 2014 (March-June), the authors carried out research on the quality of websites offering electronic access to services of the most popular banks among Polish individual clients on a sample of 361 people, where 311 respondents completed surveys correctly. The participants of the survey were students, aged 19-45, Faculty of Management University of Warsaw and Vistula University in Warsaw. Among the respondents, 69% were women and 31% men, mainly from Warsaw and surrounding areas. Each of the respondents declared to have at least one electronic access account with one of the banks operating in Poland (fifteen—used e-banking services provided by two banks, two people—of three banks), thus, in total, the authors examined access to 339 active electronic accounts (see Fig. 2).

In the surveyed population majority of people held accounts in the banks which are considered to be internet banks (mBank, Inteligo PKO BP), or regarded as modern (AliorBank, Millenium), or the largest ones (PKO BP, BZ WBK, CityBank). This does not correspond to the numbers of electronic access accounts declared by particular banks, however, considering the facts that only the active accounts were described and the fact that the surveyed population is relatively young and specific, the structure of the use of accounts is probably closer to reality than the one presented on the basis of official statistics.

This study belongs to a series of cyclical, yearly analyses concerning the factors influencing the usability of websites with online access to individual accounts in the banks [4], [5]. The same set of criteria has been applied in the evaluation of e-banking condition at 2013, 2011 and earlier, before the crisis began in 2008. They were created on the base of internet discussion among leading researchers in this field from some universities in Poland (Wrocław University of Economics, University of Economics in Katowice, Poznań University of Economics, University of Szczecin, Faculty of Economic Sciences and Faculty of Management University of Warsaw).

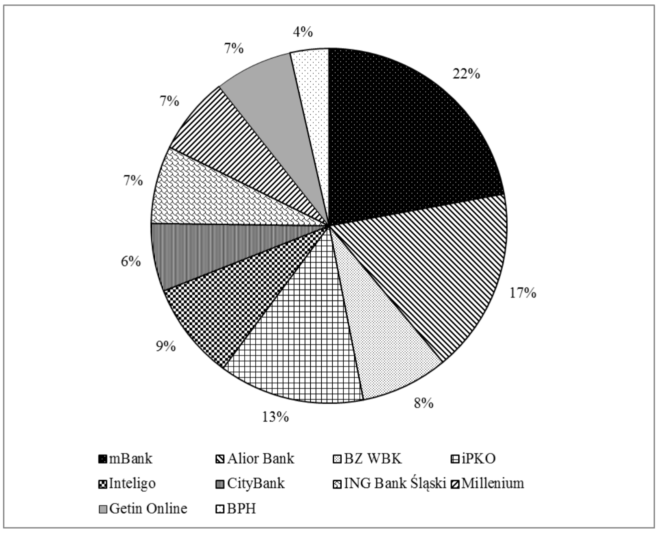


Fig. 2. The percentages of holders of accounts with electronic access in ten the most popular banks

The respondents filled in the tables evaluating e-banking websites of the banks where they had their accounts, performing the analysis and assessment of the obtained results. The tables which they completed were sent by electronic mail. In the second stage, they imposed preference coefficients on particular criteria and performed further calculations. The obtained findings were supplemented with comments. All calculations in the present study are carried out with the application of the authors own, though based on the literature and consultations with experts, set of criteria for a scoring evaluation and a selection of electronic access to services of selected banks.

Criteria applied in this study can be divided into two main groups:

- economic—annual nominal interest rate of personal accounts, account maintenance PLN/month, fee for a transfer to a parent bank, fee for a transfer to another bank, payment order, fee for issuing a debit card, fee for a card PLN/month, interest on savings accounts, interest rate on deposits of 10,000, interest rate on loans 10,000;
- technological—additional services (such as: insurance, investment funds, cross-border transfer or foreign currency account), functionality (set of function available for user), access channels to an account (branches, the Internet, Call Centre, mobile phone), security (ID and password, token, SSL protocol, a list of single-use passwords, a list of single-use codes), visualization (colours,

fonts, background, photos etc.), navigation, clarity and ease of use.

Considering the situation of the signs of economic crisis spreading, the authors applied a set of psychological criteria in addition to the criteria used previously in the evaluation of e-banking websites which were discussed above. The psychological criteria included the so-called anti-crisis criteria related to—according to the experts cooperating with the authors—all those activities, which were to counteract potential impact of the crisis on the banking sphere [6]. This additional group of criteria was also included in the previous evaluation of e-banking websites. The group of the considered anti-crisis measures includes:

- dynamics of interest rates on deposits (reduction, increase, differences in rates, tendencies),
- dynamics of interest rates on credits (reduction, increase, differences in rates, tendencies),
- stability of the policy related to basic fees (the number and the nature of changes),
- degree of customer confidence (the number of individual clients, its dynamics, how long the bank has been operating in the Polish market),
- the average places occupied in the rankings in the Internet and trade magazines last year.

In the scoring method the authors collected information on selected criteria; they were assigned values according to the assumed scoring scale and the results were analysed in a combined table. For the purposes of the evaluation the authors applied—as in previous studies—a typical R. Likert scale [8]. A scoring method was used in two variations: simple—where criteria were treated equivalently; and one with a preference scale—where sets of criteria were assigned indicator values differentiating their treatment by clients (the total of coefficients = 1).

In a simple scoring method you need to measure the distance from the maximum value to be obtained (according to the assumed scoring scale). It concerns the value of criterion measure and in the sense of a distance it is the same when we measure the distance from one criterion to another as the other way round. However, we do not define the relations between particular criteria. Assigning a preference scale to particular criteria (or sets of criteria) can be regarded as such a measure. A linear preference scale in a normalized form defines in turn the participation of particular criteria in the final score. It establishes a one-time relation between criteria in relation to the final score, it is also a specific "averaged" measure of criteria in particular cases, without the individualization of the evaluation for any of them. However, it does not specify to what degree one criterion is better/ worse than the other. It is merely a derivative of the normalized distance.

## III. Comparative Analysis of Internet Access to E-Banking Accounts Using a Scoring Method

To evaluate cost, functional, technological and anti-crisis criteria the authors used a preliminary table presenting bank offers related to internet banking services used by respondents and fees connected with using bank accounts operated via the Internet. This table has been generated on the basis of data obtained from websites of selected banks. On the basis of the surveys the authors created an averaged combined table for the criteria generated by users.

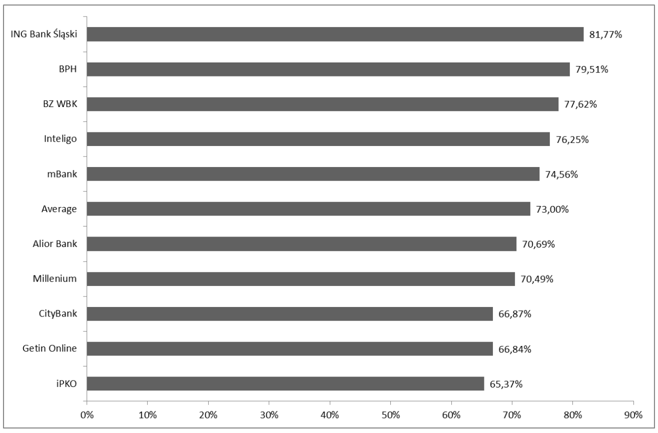


Fig. 3. Ranking of banks for assessing electronic access to individual accounts in selected banks in Poland

The spread in the respondents' evaluations of the analysed banks amounts to 16.4 percentage points (compared to 5.1 percentage points in 2011, and 2.3 points in 2008), which reflects the growing diversity of evaluations; which confirms the thesis that the period of crisis increased the radicalism of evaluations and heightened expectations concerning the tools used to access an account.

This time the best in the ranking were: ING Bank Şląski (81.77%) and BPH (79.51%). Directly behind are: BZ WBK and Inteligo. The low position of mBank (fifth position in the reverse order), came as a surprise because this bank so far occupied leading positions and it was very popular with the analysed group of people (22% of respondents). It is worth mentioning that in the rankings [17] till May 2011 it held the first position. There occurred a reversal of the situation from three years ago—the banks which two years ago fell in the rankings, at present, are trying to make up for the previous losses. Another issue which seems to be characteristic of this study—general scores for the quality of the websites increased (see Fig. 3.).

In the majority of analysed cases there are no obligatory payments for issuing a debit card; transfers to the parent bank are usually free of charge. The level of security can be regarded as satisfactory for clients. Actually, it has not changed from 2008. Based on the compilation, we can conclude that a fee for issuing a card (usually there is no fee for such a service) reached a level which, at present, may satisfy clients' needs almost in 100% (96%). Undoubtedly, the worst indicator is interest rate on savings accounts functionality (evaluated by the majority of users as too low—37.92% of the maximum scores). The interest rates on deposits reached over 57% of maximum score (Fig. 4). From the factors not listed within

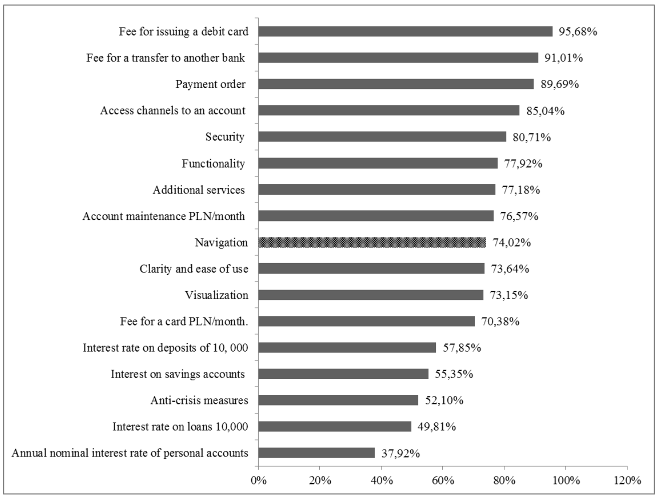


Fig. 4. Ranking of criteria for assessing electronic access to individual accounts in selected banks in Poland in the beginning of 2014

the criteria clients paid attention to the lack of possibility to make a cross-border transfer (e.g. SWIFT in Inteligo) or no possibility of fully automatic obtaining a credit—via the Internet. In 2008 there were no anti-crisis measures among the criteria; however, if we compare this study with the research carried out in 2011, we have to admit that during the crisis e-banking clients neither noticed any signs of the crisis nor were able to define anti-crisis measures undertaken by the banks, and at present they sometimes suggest criteria to be applied for their evaluation.

## IV. Comparative Analysis of Internet Access to E-Banking Accounts by Means of a Scoring Method With Preferences

One of the methods limiting a specific subjectivity in the experts' or users' evaluations (apart from the previously used averaging of scores) is applying unitary preferences with regard to particular criteria or sets of criteria. For each group the authors applied one dominant variant:

- economic (70% for economic criteria and 15% for the remaining ones),
- technological (70% for technological criteria and 15% for the remaining ones),
- anti-crisis (70% for anti-crisis criteria and 15% for the remaining ones).

In the first case the three leading positions are taken by ING Bank Śląski, Inteligo and BZ WBK. Next, BPH and mBank are among the best banks with regard to economic factors. iPKO, CitiBank, Alior Bank occupy the last positions of the ten analysed banks. In the leading positions (in the relation to research without preferences) Inteligo moves by two places, Millenium by one place. iPKO BP records the greatest fall.

In the second case the leaders are as follows: BPH, ING Bank Śląski and BZ WBK—further positions are taken by: Inteligo and mBank. The last positions were occupied by: Getin Online, iPKO and Citi Bank.

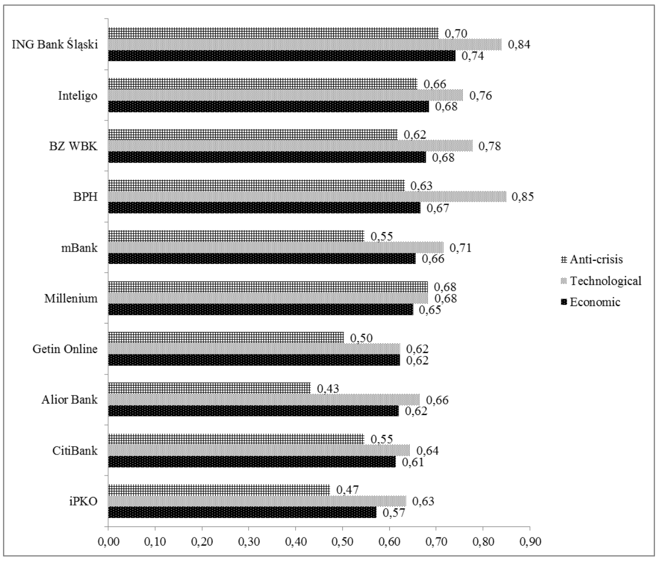


Fig. 5. Ranking of the scores according to various types of preferences for ten selected banks in Poland in the beginning of 2014, according to the order of economic criteria

In the third case the order is very similar to the other evaluations. The first is ING Bank Śląski, next—Millenium and Inteligo. The last positions are occupied by Alior Bank, iPKO and Getin Online.

All in all, multiplaying the values by preference coefficients did not bring about any significant changes in the analysed cases, because the order of the examined websites offering access to e-banking services remained basically the same. The results for the groups are presented in Fig. 5.

## V. Summary and Some Conclusions

The present analysis has shown that the crisis situation, whose signs are visible in various industries, does not apply to electronic banking. While in 2011 it could be a one-off phenomenon, after five years we may conclude that it starts to be seen as a clear trend. Also, it confirms the changes concerning the banking customers' awareness. The choice of the access to an account starts to be a matter of an informed choice, not a chance or habit. The decision is determined both economic and technical conditions. The result is the choices made by clients reflected in the presented study and commented on in the surveys. It is true that some of the opinions indicate—despite the awareness of certain shortcoming of the bank where they hold an account—resistance to changes, but it is the first step to move their account to another bank.

The selection of the research sample and its limitations affected the obtained findings. Students are a group which have relatively the greatest number of access to e-banking and use it most frequently. And know e-banking websites very well. Taking the above factors into consideration, it appears to be the best sample for carrying out the research. This group in the population also has a wide, and perhaps the greatest, knowledge concerning the newest technologies and their use. The selection of the student groups was random, the dominance of women and students of early years of study are accidental. However, it is a group which does not have considerable financial resources, and this may be the reason why the representatives of this group have relatively the cheapest access to the internet banking websites. Simultaneously, the awareness of the high usability of information technology means that this particular group is able to appreciate the practical (or entertainment) value of the offered e-banking services. In addition, the representatives of this group, after a few years of using e-banking websites designed for them, are less likely to be satisfied with a product or service—in the form of e-banking website for example—of lower quality.

Taking into consideration the conducted analyses, we may draw the following conclusions:

- in the minds of users of electronic banking a clear distinction between the virtual banks (electronic access only) and electronic access services of traditional banks lost its importance, and it appears to be a continued trend. It is caused by the following phenomena (quotation from a student's survey: ...as we can see the results are comparable. This is due to the large similarity of services offered by banks with regard to visualisation and "starting package", that is all for 0zł ...): virtual and traditional banks try to maximally increase the number of communication channels, it is difficult to separate a virtual bank from a traditional one, e-banking websites of traditional banks are just as technologically advanced and modern as those of virtual banks, we observe lowering of that prices of basic e-banking services in traditional banks, sometimes below the prices of virtual banks, we have a more possibility of access—phenomenon of mobile banking via mobile applications (smartphones, tablets),
- users have higher expectations with regard to the quality of e-services. The averages from the rankings—previously relatively constant—have become dynamic and fluctuate,
- entering the market (for example—the case of Alior Bank) and allocating significant resources to a clever advertising campaign does not guarantee an automatic promotion to the top position in the rankings (for example—Credit Agricole),
- having two or more accounts in two or more banks to perform various financial transactions is still a rare phenomenon,
- too few clients dynamically respond to changes in banking services market,
- vast majority of active bank customers consider economic criteria to be the most important criteria in the evaluation of electronic access to banking services—usually the prices of the most frequently used services. More and more people admit, however, that when selecting a website, to a certain degree, they tend to focus on user-friendliness and intuitiveness as well as the visual attractiveness of the website (quotation from a student's survey (original phrasing): ... personally, when selecting banking services, I am more influenced by economic factors than the visualisation or the simplicity or complexity of navigation. Visual qualities depend on a personal taste, and I think they should not be a decisive factor in the selection of a financial institution, to which we entrust our own money. With regard to the technical and functional level of the service, it does not determine my choice—especially since the current market seems to have developed a certain standard, below which banks can no longer operate because customers are bound to verify it quickly and leave ...),
- users of electronic banking services more frequently notice anti-crisis measures of banks and even though they do not influence their choices in any considerable degree, they can note and identify them.

This confirms the authors thesis about the inadequacy and a specific superficiality of standard, unified, quantitative methodologies used for evaluation and selection of e-banking websites. It also points to the need of further studies into constructing multi-dimensional, multi-criteria, hierarchical and multi-faceted system for websites' evaluation, with the consideration of additional, more specific information, e.g. customer profile [3].

Nevertheless, despite the problems related to using e-banking services, which the article presents, from year to year we observe that the Internet tends to assume the role of the main (also for an individual client) channel of communication with the bank. Undoubtedly, this development irrevocably changes the expectations, perceptions and habits related to using banking services which users have had so far, and also, simultaneously, it urges the banks to introduce quick changes of the medium which would take into account holders' requirements.

## REFERENCES

[1] H. Bauer, M. Hammerschmidt, T. Falk, "Measuring the quality of e-banking portals", International Journal of Bank Marketing, vol. 23 no. 2, 2005, pp. 153–175.

[2] W. C. Chiou, C. C. Lin, C. Perng, "A strategic framework for website evaluation based on a review of the literature from 1995–2006", Information & Management, vol. 47, no. 5-6, 2010, pp. 282–290.

[3] W. Chmielarz, "Koncepcja ekspertowego systemu oceny i selekcji witryn internetowych", chapter 4, Koncepcje zastosowań Systemów ekspertowych [in:] Wiedza i komunikacja w innowacyjnych organizacjach. Systemy ekspertowe—wczoraj, dziÅŕ, jutro, edited by J.Gołuchowski, B. Filipczyk, Prace Naukowe UE w Katowicach, Wydawnictwo UE w Katowicach, Katowice. pp. 183–190.

[4] W. Chmielarz, "Methodological Aspects of the Evaluation of Individual E-Banking Services for Selected Banks in Poland", chapter 11 [in:] Infonomics for Distributed Business and Decision-Making Environments. Creating Information System Ecology, edited by M. Pańkowska, IGI Global, Business Science Reference, Hershey-New York, 2010, pp. 201–216.

[5] W. Chmielarz, "Comparative analysis of electronic banking services in selected banks in Poland in 2013", in: Current Problems of Banking Sector Functioning in Poland and East European Countries, red. Naukowa A. Gospodarowicz, D. Wawrzyniak, nr 316, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, 2013, pp. 16–29.

[6] W. Chmielarz, "Metody oceny elektronicznych usług bankowych dla klientów indywidualnych w Polsce", [in:] edited by A. Gospodarowicz: Bankowość detaliczna — idee, modele, procesy, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 54, Wydawnictwo UE we Wrocławiu, Wrocław, chapter 1, pp. 9–26.

[7] http://www.egov.vic.gov.au/focus-on-countries/europe/trends-and-issues-europe/statistics-europe/internet-statistics-europe/comscore-releases-2014, Apr. 2014.

[8] R. Likert, "A Technique for the Measurement of Attitudes", [in:] Archives of Psychology, no. 140, 1932, pp. 1–55.

[9] M. B. Mateos, A. C. Mera, F. J. Gonzales, O. R. Lopez, "A new Web assessment index: Spanish universities analysis", [in:] Internet Research: Electronic Application and Policy, no. 11(3), 2001, pp. 226–234.

[10] Y.K. Migdadi, "Quantitative Evaluation of the Internet Banking Service Encounter's Quality: Comparative Study between Jordan and the UK Retail Banks", [in:] Journal of Internet Banking and Commerce, no. 2 (13), 2008, pp. 70–79.

[11] F. J. Miranda, R. Cortes, C. Barriuso, "Quantitative Evaluation of e-Banking Web Sites: an Empirical Study of Spanish Banks", [in:] The Electronic Journal Information Systems Evaluation, no. 2(9), 2004, pp. 73–82, at http://www.eiise.com.

[12] NETB@nk—Raport Bankowość internetowa i płatnokści bezgotówkowe. Podsumowanie IV kwartału 2014 r. at: http://www.zbp.pl/Netbank_Q4_2014v3.pdf, April 2015.

[13] M. Sikorski, Usługi on-line. Jakość, interakcje, satysfakcja klienta, Wydawnictwo PJWSTK, 2013, Warszawa.

[14] H. W. Webb, L. A. Webb, "SiteQual: an integrated measure of Web site quality" [in:] Journal of Enterprise Information Management, vol. 17, no. 6, 2004, pp. 430–440.

[15] J. Wielki, "Modele wpływu przestrzeni elektronicznej na organizacje gospodarcze", Wydawnictwo UE we Wrocławiu, Wrocław, 2012.

[16] Z. Yang, S. Cai, Z. Zhou, N. Zhou, "Development and validation of an instrument to measure user perceived service quality of information presenting Web Portals". [in:] Information & Management, vol. 42, no. 4, 2005, pp. 575–589.

[17] "Znajdź swój bank", Neewsweek, ranking June/July 2010. SMG/KRC auditors' group, Newsweek, October 3-th 2010, p. 78.

DOI: 10.15439/2015F162
ACSIS, Vol. 5

# Conceptual design of financial ontology

Helena Dudycz
Wrocław University of Economics
Komandorska Str. 118/120,
PL 53-345 Wroclaw, Poland
Email: helena.dudycz@ue.wroc.pl

Jerzy Korczak
Wrocław University of Economics
Komandorska Str. 118/120,
PL 53-345 Wroclaw, Poland
Email:jerzy.korczak@ue.wroc.pl

*Abstract*—**The article presents the approach to the conceptualization of the financial knowledge used for an Intelligent Dashboard for Managers. The content of the knowledge is focused on essential financial concepts and relationships related to the management of small and medium enterprises (SME). That includes the illustration of the conceptualization of the ontology. The designed ontology was split into six ontologies describing areas of Cash Flow at Risk, Comprehensive Risk Measurement, Early Warning Models, Credit Scoring, Financial Market, and General Financial Knowledge. The examples of the topic map of key financial indicators and their interpretation are given. The results of this research have been implemented in the Business Intelligence system.**

## I. Introduction

TO make optimal decisions, managers need very useful, adequate and easy to interpret information. They must analyse various economic indicators assessing the financial situation of an enterprise. Data for analyses are usually extracted from different information systems. To interpret a financial indicator, a manager should analyze relations between indicators and economic data which have influence on its value. However, available information systems concentrate mainly on providing information reflecting hierarchic relationships between examined indicators. Decision-makers evaluate semantic associations existing between them. Such an analysis of indicators can potentially ease and shorten the time needed, inter alia, to identify chances of advancement and threats of breakdown related to carrying out an activity. In order to facilitate the process of data analysis, the usage of the ontology is proposed as a model of financial knowledge about the analysis of indicators.

The decision-makers of small and medium enterprises (SMEs), in comparison to managers of big companies, may not have access to all essential strategic information. Usually financial expertise is either not available or too expensive. Big companies have at their disposal strategic consultation and possess standard procedures to solve problems in the case of essential changes in business environment. For financial and personnel reasons most SMEs cannot afford these types of facilities. It should be noted that SMEs operate in a definitely more uncertain and risky environment than big enterprises, because of a complex and dynamic market that has much more important impact on SMEs' financial situation than on big companies [1].

In general, most existing Business Intelligence (BI) and Executive Information Systems (EIS) provide the functionality of data aggregation and visualization. Many reports and papers in this domain underline that decision makers expect new ICT solutions to interactively provide not only relevant and up-to-date information on the financial situation of their companies, but also explanations taking into account the contextual relationships.

Our research concentrates on two essential issues: supporting decision makers in the area of analysis of economic and financial information using solutions for representing the ontology of economic and financial data (for example: topic map)[1], and using tools for visualization of the semantic network, which is based on an ontology model of the economic knowledge and data from all relevant information systems[2].

The aim of this article is to present the conceptual design of financial ontology. The structure of the paper is as follows. In the next section, the functional schema of the system is discussed. The main domain areas of financial knowledge are presented and detailed by the topic map of the main financial indicators. Section 3 describes the process of ontology development, in particular the actual design of the ontology. A case study in section 4 illustrates an example of financial ontology conceptualisation. To show the reasoning a case for explanation of financial data is specified. In the conclusion, the future research directions are indicated.

[1] This research was supported by the National Research and Development Centre within the Innotech program (track In-Tech), grant agreement no. INNOTECH-K1/IN1/34/153437/NCBR/12. The name of this project was the Intelligent Dashboard for Managers, which was conducted by a consortium led by the Wrocław University of Economics, Poland (the leader: Jerzy Korczak), and the other principal member is the company UNIT4 TETA BI Center. Credit Agricole Polska also participated in the project. The project was realized during the period 2012-2014.

[2] This research was supported by the Polish Ministry of Science and Higher Education (grant N N111 284038 ). This grant was realised by Helena Dudycz during the period 2010-2012.

978-83-60810-66-8/$25.00©2015, IEEE

## II. Proposal widen Business Intelligence system functionalities

The Business Intelligence (BI) system is used for the analysis of all basic areas of an enterprise's activities, such as, e.g., finance and accounting, manufacturing, logistics, marketing, sales, and customer relationships. These applications provide many reports containing valuable information in each statement. Retrieval information from these reports is eased by the use of appropriate forms of its presentation, and of a friendly and easy user interface. Nowadays decision-makers want not only to look at static reports or even ad hoc reports, but also easy-to-use tools to assess goals and key performance indicators to identify any chances of advancement and threats of breakdown. The usefulness of the BI system is not related to the amount of generated information, but to the information which is required at the right moment. These were basic motives for developing and applying a new technology and knowledge representation in the BI system. In the literature, the development of BI systems towards BI 2.0 (using semantic search) is described (see [2]-[4]). This system is focused on the semantic analysis of data, using data and information from multiple sources (including external sources). One of the main artifacts to create a semantic network is the ontology, because the architecture of BI 2.0 has new components, such as ontologies and service ontologies (see [2]). The ontologies are used to create the necessary knowledge models for defining and explaining functionalities in analytical tools. Using ontologies and semantic networks for a visual interface supporting an information search in the BI system may help to reduce the following weaknesses of management information systems (see [4], [5]):

- lack of support in defining business rules for getting proactive information and support in consulting in the process of decision making;
- lack of a semantic layer describing relations between different economic topics;
- lack of support in presenting the information of different users (employees) and their individual needs;
- difficulty in rapidly modifying existing databases and data warehouses in the case of new analytic requirements.

In Figure 1 a functional architecture of the information system is presented, with ontology applications. Various mechanisms can be seen for extracting source data from transactional systems (ETL), its data warehouse, and external sources. However, the available solutions – in particular the standard analyses, reports and analytical statements generated by the system – are complemented by economic and financial knowledge (most importantly ontologies). This enables a dynamic, interactive analysis of key economic and financial indicators. Such architecture concept was used in the project InKoM (a wide review of the issue is presented in: [6], [7]). This solution will significantly extend existing BI and EIS functionalities.

To support the analysis, SMEs decision makers need economic and financial knowledge. The scope of required knowledge was divided arbitrarily by experts into six selected areas, namely: Cash Flow at Risk, Comprehensive Risk Measurement, Early Warning Models, Credit Scoring, Financial Market, and General Financial Knowledge (Fig. 2). They are described in [6]. Between these fields there are intersections, and some topics belong to two or three areas.

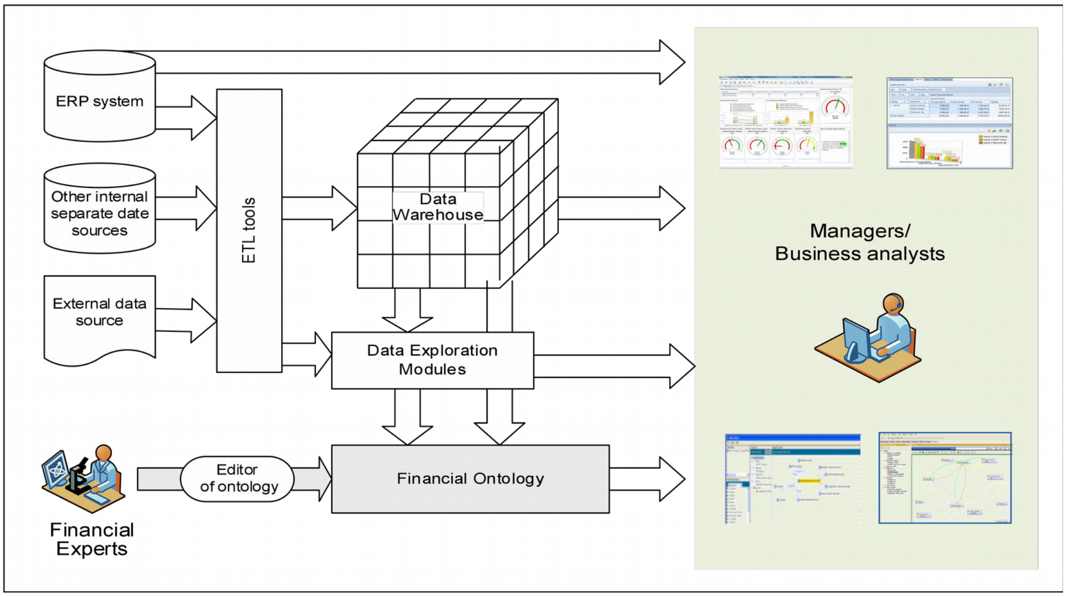


Fig. 1. Functional architecture of information system with ontology applications

Source: based on [1, p. 57].

The system that enables semantic information retrieval should be intuitive to use or easy to understand. For managers, the presentation layer is the most critical aspect of a BI system, since it broadly shapes their core understanding of the data displayed [8]. The basic assumption of navigation is that managers should be able to view focus and context areas at the same time to present an overview of the whole knowledge structure [9].

Ontology of financial knowledge is the foundation of creating a semantic network. In our project, special attention was paid to the role of the visualization of a semantic network, which is not only a tool for presenting data, but also provides an interface allowing interactive visual information retrieval (see inter alia [10], [11]). Working from the displayed semantic structure of a built-in ontology of financial and economic knowledge, it is possible to interactively choose analyzed topics or relations, to change the area of presented details, and to obtain relevant source data.

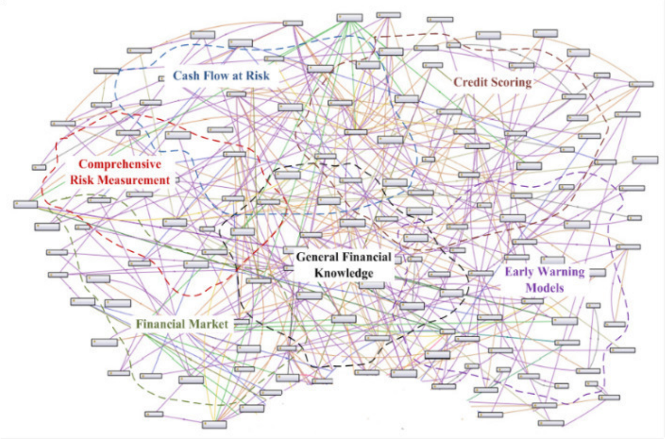


Fig 2. Six selected areas of ontology in the Intelligent Dashboard for Managers

Source: [1, p. 61].

In the years 2011-2013 we carried out four experiments with a created prototype and users participation. The results of these experiments are optimistic (discussed inter alia [12]). However, the usefulness of these solutions depends mainly on substantive content, that is, correctly building the ontology of financial or economic knowledge.

## III. Design process of financial ontology

In the literature many different approaches to design of an ontology can be found (a wide review of the issue is presented in: [13]). There are many methods describing the methods of creating ontology for information systems. These are inter alia: Cyc, KBSI, TOVE, EMA, HOLSAPPLE, HCONE, System KACTUS, SENSUS, UPON, METHAONTOLOGIA, On-To-Knowledge method (a wide review of the issue is presented in: [14], [15]). But so far there is no single approach accepted by all.

Based on the analysis of existing methodologies and our research, a method of creating an ontology of financial indicators has been proposed. In this method, the following stages are distinguished (see also: [1], [6], [16], [17]):

*1.* Definition of the goals, scope, and constraints of the created ontology. While creating an ontology, assumptions about the created model of knowledge that will apply during its building have to be provided. That requires an answer to the question: *what will the created ontology be used for?*

*2.* Conceptualization of the ontology. Independently of the field that is to be modeled by using an ontological approach, it is the most important stage in creating a model based on ontology (see inter alia [18, p. 2036]). It includes the identification of all concepts, definition of classes and their hierarchic structures, modeling relations, identification of instances, specification of axioms, and rules of reasoning.

*3.* Verification of the ontology's correctness by experts. In this stage, the constructed ontology is verified by experts who did not participate in the process of conceptualization. Verification is carried out in two steps. The first concerns a formal verification of the specified ontology (e.g. incorrect relations are indicated) with the use of a given editor. The second step is carried out by experts from the given field and concerns content verification which includes verification of the correctness of topics' definitions, correctness of taxonomic topics, and correctness of relational dependences between topics. In the proposed method of building an ontology of financial knowledge verification and validation were isolated in accordance with approach used in software engineering (see [19]).

*4.* Encoding the ontology is described in the formal language or editor of ontology. The result of this stage is the encoded ontology. Two basic stages of encoding of ontology are: (1) entering all topics and creating a taxonomy of these topics, and (2) entering all other types of relations between topics.

*5.* Validation and evaluation of the built ontology. In this stage, the encoded ontology is checked to ensure it meets the needs of the managers. Validation is carried out in three areas. Firstly, validation of usefulness and correctness of the created ontology by experts (managers) who will potentially use it. Secondly, evaluation of the application with a created ontology is carried out by managers. Finally, the validation of predefined use cases is carried out. That requires an answer to the questions: *will the created ontology be useful for the managers who will use it?*

Figure 3 shows the design process of an ontology of financial knowledge.

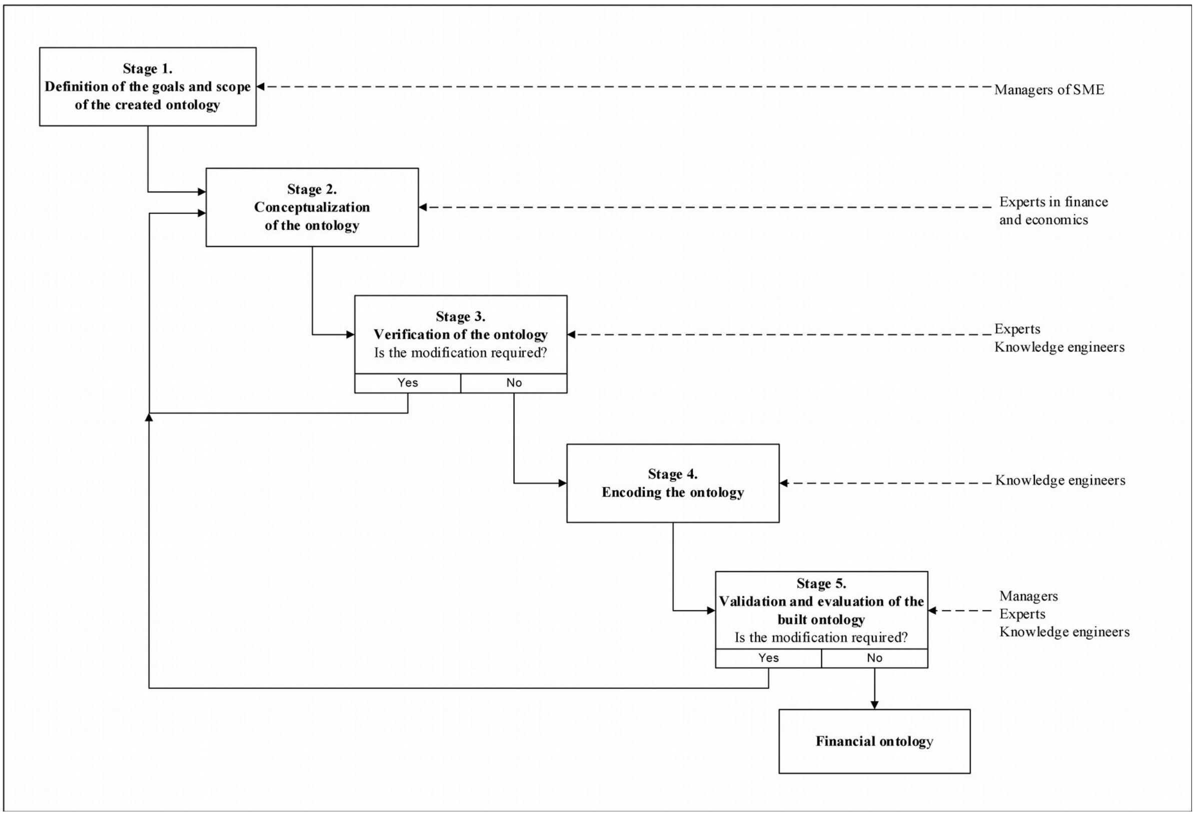

Stage 1.
Definition of the goals and scope of the created ontology
Managers of SME
Stage 2.
Conceptualization of the ontology
Experts in finance and economics
Stage 3.
Verification of the ontology
Is the modification required?
Yes
No
Experts
Knowledge engineers
Stage 4.
Encoding the ontology
Knowledge engineers
Stage 5.
Validation and evaluation of the built ontology
Is the modification required?
Yes
No
Managers
Experts
Knowledge engineers
Financial ontology

Fig. 3. Design process for an ontology of financial knowledge

Source: own elaboration.

The important stage in the described method is the conceptualization of financial indicators. The process of conceptualization of an ontology is an intellectual activity of organizing knowledge acquired from a given domain knowledge. This is carried out by the person, either an expert or in collaboration with an expert, responsible for creating the model of knowledge without the support of automated tools (see inter alia [18, p. 2036]). In the literature [14], [15]) the following phases in the conceptualization of the ontology of financial knowledge are shown (see also: [6]):

(a) Identification and definition of all topics. A topic, representing any concept, is "a syntactic construct that corresponds to the expression of a real-world in a computer system" [10, p. 60]. A topics' list is determined by experts in a given domain of economic knowledge. These topics include, beside their names, also their synonyms and descriptions.

(b) Creating a taxonomy of topics. Specification of taxonomic relations between distinguished topics and defining classes and subclasses. In general, these relationships describe the topics generalization. The description of a taxonomy can be presented in graphic or tabular form. An interesting approach to creating a taxonomy is proposed in METHONTOLOGIA (see i.e. [14]).

(c) Definition of all other types of relations between topics, notably the basic relationships aggregate of (*Aggregate – Member*) was defined. Moreover, within each ontology, additional relations were defined.

(d) The list of all the individual relationships existing in the ontology. The list includes: the name of the relationship, source topic, and target topic.

(e) Description of functions and rules. This description contains: name, input, output, initial and final conditions, and definition of operations.

(f) Description of usage scenarios. Usage scenarios, also called use case views, describe demonstration analyses of economic topics occurring in this ontology.

Building an ontology always denotes analysis and organizing of knowledge. That work has required multi-domain expertise, both theoretical and practical, in economics, finance, and informatics.

## IV. CASE STUDY - CONCEPTUALIZATION OF FINANCIAL ONTOLOGY

The important stage in the described method of creating an ontology is the conceptualization of ontology. We will illustrate it in the example of the analysis of Return on Sales (ROS) indicator. ROS is one of indicators that managerial staff analyses to evaluate a company's efficiency. This measure is helpful to management by providing insight into the profit structure of sales. An increasing ROS indicates that the company is growing more efficiently, while a decreasing ROS signals financial troubles. Managers often use the ROS indicator and sales analysis reports to identify market opportunities and areas where they could increase the volume of sales.

The conceptualization of the ontology of the ROS indicator was as follows:

*1.* Identification and definition of all topics. Table I presents the example of the topics list.

TABLE I.
THE EXAMPLE OF TOPICS LIST

| Name | Synonym | Description |
|---|---|---|
| Return on sales | ROS | A ratio widely used to evaluate a company's operational efficiency. ROS is also known as a firm's "operating profit margin". It is calculated using this formula:<br>Net_profit / Revenues_from_sales<br>Recommendation: Compare a company's ROS over time to look for trends, and compare it to other companies in the industry. |
| Net profit | NP | Net profit is calculated by subtracting a company's total cost from total income, thus showing what the company has earned (or lost) in a given period of time (usually one year). Also called net income. |

Source: own elaboration.

*2.* Creating a taxonomy of topics. Figure 4 shows the taxonomy for topic *Indicators* and topic *Profitability evaluation indicators.*

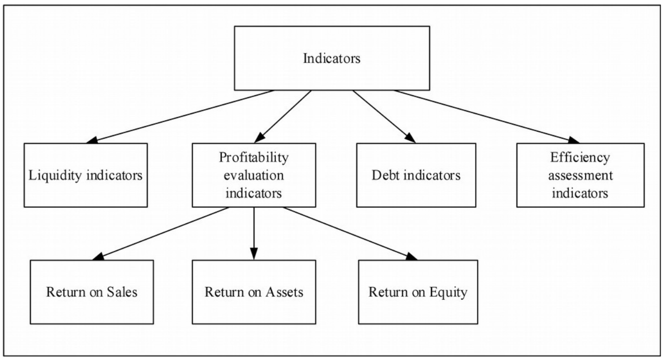


Fig. 4. Example of the taxonomy for topic *Indicators* and topic *Profitability evaluation indicator*
Source: [1, p. 64].

*3.* Definition of all other types of relations between topics. In this ontology, the basic relationship *aggregate of (Aggregate – Member*) is defined. Moreover, additional relations are defined, for example: *potential growth, proportional positive/negative change, is the sum, is the quotient.*

*4.* The list of all the user defined relationships existing in the ontology. The description of a taxonomy can be presented in graphic or tabular form. Figure 5 shows the relations existing between topic *Return on Sales* (in the class *Profitability evaluation indicators*) and topics: Net profit, and Revenues from sales ((in the class Total income). Solid lines denote taxonomic relations (relation *Subclass – of*), whereas broken lines denote domain-specific relations (e.g. relation *is the quotient*).

5. Description of functions and rules. The definition describes how to compute and interpret their values. This description can contain: name, input, output, initial and final pre-conditions, and definition of formula (see also: [9]). The following description specifies the example of the indicator *Return on Sales*:

***Name****:*
  *Indicator Return on Sales (ROS)*

***Input****:*
 *Result of Net profit (NP)*
 *type: value extracted from Balance Sheet*
*Revenues from sales (RS)*
*type: number, value extracted from Balance Sheet*

***Output****:*
*Return on Sales*
*Description/formula:*
    *ROS = NP/ RS*

```
Final conditions:
if (ROS < value_1)
  Interpretation_1
else if (value_1  > ROS  < value_2)
  Interpretation_2

else if  ....
  ....

else if ( ROS > value_n)
  Interpretation_n
```

6. Description of usage scenarios. Usage scenarios, also called use case views, describe demonstration analyses of economic topics occurring in this ontology. For example, a manager analyzes the ROS indicator and would like to identify causes of a decreasing value ROS:

*a.* The manager analyzes the semantic network, from which it follows that the *Return on Sales* indicator depends on two values: *Net profit* and *Revenues from sales*.

*b.* From the BI system the manager receives the values of the *Net profit* and *Revenues from sales*. It notices that the company's value of net profit is worse than for the previous period (Fig. 6).

*c.* The manager analyzes the semantic network of the *Net profit* to identify two parameters: *Total income* and *General costs*.

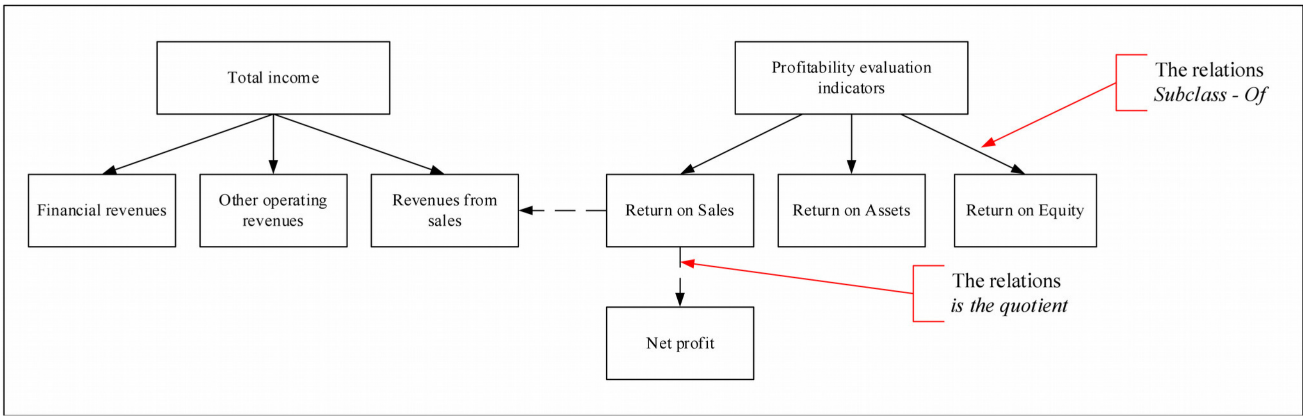


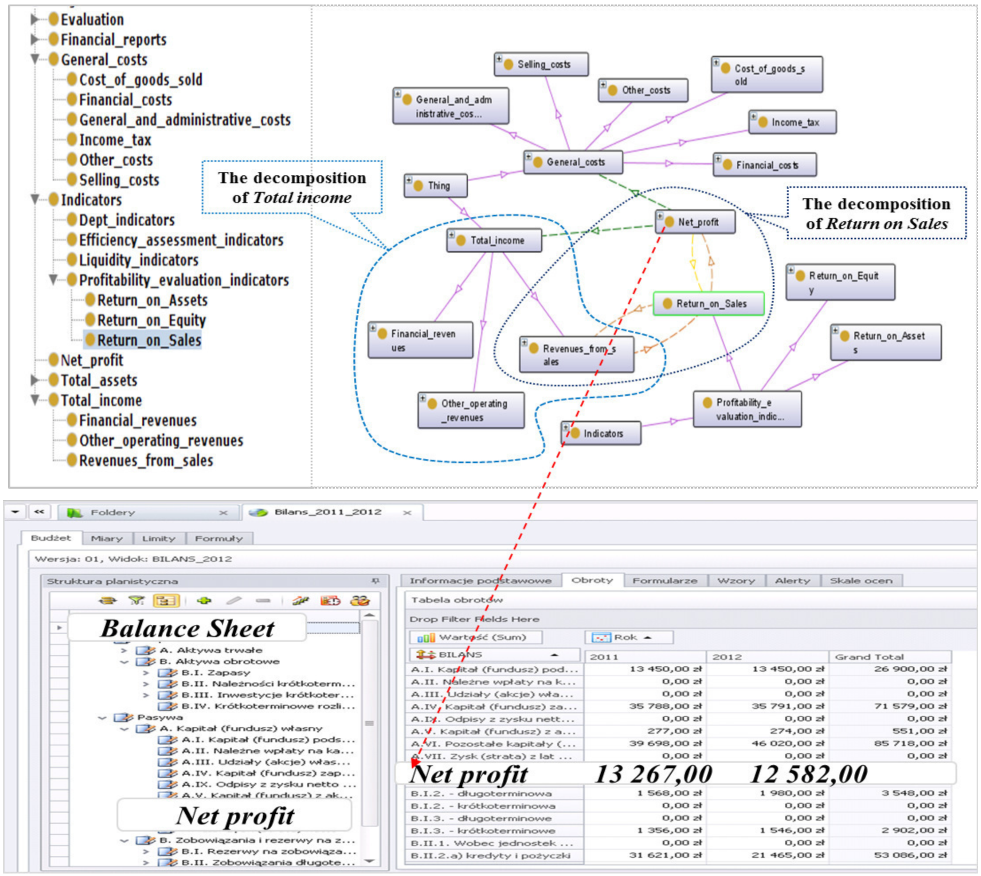
Fig. 5. The example of illustrated relationships *is the quotient*

Source: own elaboration.

*d.* From the BI system the manager receives the values of the *Total income* and *General costs*. It notices that the company's value for General costs is worse than the previous period.

*e.* The manager analyzes the semantic network of the General costs to identify causes of unfavorable values from the ROS indicator.

*f.* Based on the analysis conducted of economic indicators, the manager can undertake corrective actions which may potentially result in improving the company's *Return on Sales*.



Wersja: 01, Widok: BILANS_2012

Tabela obrotów

| BILANS | 2011 | 2012 | Grand Total |
|---|---|---|---|
| A.I. Kapitał (fundusz) pod... | 13 450,00 zł | 13 450,00 zł | 26 900,00 zł |
| A.II. Należne wpłaty na k... | 0,00 zł | 0,00 zł | 0,00 zł |
| A.III. Udziały (akcje) wła... | 0,00 zł | 0,00 zł | 0,00 zł |
| A.IV. Kapitał (fundusz) za... | 35 788,00 zł | 35 791,00 zł | 71 579,00 zł |
| A.IX. Odpisy z zysku nett... | 0,00 zł | 0,00 zł | 0,00 zł |
| A.V. Kapitał (fundusz) z a... | 277,00 zł | 274,00 zł | 551,00 zł |
| A.VI. Pozostałe kapitały (... | 39 698,00 zł | 46 020,00 zł | 85 718,00 zł |
| A.VII. Zysk (strata) z lat ... | 0,00 zł | 0,00 zł | 0,00 zł |
| **Net profit** | **13 267,00** | **12 582,00** | |
| B.I.2. - długoterminowa | 1 568,00 zł | 1 980,00 zł | 3 548,00 zł |
| B.I.2. - krótkoterminowa | 0,00 zł | 0,00 zł | 0,00 zł |
| B.I.3. - długoterminowe | 0,00 zł | 0,00 zł | 0,00 zł |
| B.I.3. - krótkoterminowe | 1 356,00 zł | 1 546,00 zł | 2 902,00 zł |
| B.II.1. Wobec jednostek ... | 0,00 zł | 0,00 zł | 0,00 zł |
| B.II.2.a) kredyty i pożyczki | 31 621,00 zł | 21 465,00 zł | 53 086,00 zł |

Fig. 6. . Example of visualization of entered ontology of ROS indicator and of Balance Sheet extracted from the TETA BI system

Source: own elaboration

.

The scenario is presented in Figure 6. The screenshot shows the expansion of the selected topic: *General costs, Total income* and *Profitability evaluation indicators*. On the diagram, it is the area encircled by a dashed line, with new topics being a subclass of the topic *Total income*. A semantic search is provided to avoid difficulties related to decision makers' interpretation of financial information. This gives the user the opportunity to search data sources taking into account not only structural dependences, but also the semantic context. In this figure there are two types of lines between topics: (1) the solid line represents a relation *Subclass – of* and (2) the dashed line represents the experts' defined relations. Business data contains a lot of hidden relationships and dependencies that make their usage difficult. To interpret the values of financial indicators correctly, many measures and ratios need to be examined that either directly or indirectly influence the final result. Explicit visualization not only makes the interpretation of indicators easier, but it also contributes to finding explanations of current values of indicators.

## V. Conclusion

The use of a financial ontology seems to be a promising extension for Business Intelligence systems. It not only improves the efficiency of analysis, but also increases the capacity of understanding of financial data.

In this paper, the approach to the ontology of the financial knowledge design process was presented. The stages of ontology design were described and illustrated using the Business Intelligence system. In the case study, the topic map of key performance indicators is presented, with their structures and relationships. The financial ontology was implemented in the extended Business Intelligence system, which will be soon commercialized by TETA BI Center.

Many extensions and applications of this work are possible. Current work is directed toward the development of smart navigation throughout the very large field of ontological concepts, and the method of financial ontology updating by adding new concepts either through a SME manager or data mining modules. Suggested approach could be used in bigger enterprises too.

## References

[1] J. Korczak, H. Dudycz and M. Dyczkowski, "Specification of Financial Knowledge – Case of Intelligent Dashboard for Managers", *Business Informatics*, Wrocław University of Economics Research Papers, no. 2 (28), 2013, pp. 56-76.

[2] G. S. Nelson, Business Intelligence 2.0: Are we there yet?, SAS Global Forum 2010, http://support.sas.com/resources/papers/proceedings10/040-2010.pdf.

[3] N. Raden, Business Intelligence 2.0: Simpler, More Accessible, Inevitable, February 01, 2007, http://www.informationweek.com/news/software/bi/197002610.

[4] D. Sell, L. Cabral, E. Motta, J. Domingue and R. Pacheco, Adding Semantics to Business Intelligence, 2008, http://dip.semanticweb.org/documents/WebSpaperOUV2.pdf.

[5] J. Korczak and H. Dudycz, "Approach to visualisation of financial information using topic maps", in: *Information Management*, B.F. Kubiak, A. Korowicki, Eds., Gdańsk: Gdańsk University Press, 2009, pp. 86-97.

[6] J. Korczak, H. Dudycz and M. Dyczkowski, "Design of Financial Knowledge in Dashboard for SME Managers", in: *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems. Annals of Computer Science and Information Systems*, vol. 1, M. Ganzha, L. Maciaszek, M. Paprzycki, Eds. Polskie Towarzystwo Informatyczne, IEEE Computer Society Press, Warsaw, Los Alamitos, CA, 2013, pp. 1111–1118.

[7] J. Korczak, H. Dudycz and M. Dyczkowski, "Intelligent Dashboard for SME Managers. Architecture and Functions", in: *Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS 2012*, M. Ganzha, L. Maciaszek, M. Paprzycki, Eds., Polskie Towarzystwo Informatyczne, IEEE Computer Society Press, Warsaw, Los Alamitos, CA, 2012, pp. 1003–1007.

[8] L. Wise, The Emerging Importance of Data Visualization, part 1, October 29, 2008, http://www.dashboardinsight.com/articles/business-performance-management/the-emerging-importance-of-data-visualization-part-1.aspx

[9] S. Smolnik and I. Erdmann, "Visual Navigation of Distributed Knowledge Structures in Groupware – Base Organizational Memories", *Business Process Management Journal*, vol. 9, no. 3, 2003, pp. 261-280.

[10] B. L. Grant and M. Soto, "Topic maps, RDF Graphs, and ontologies visualization", in: *Visualizing the Semantic Web. XML-based Internet and information visualization*, second edition, V. Geroimenko, C. Chen Eds., London: Springer-Verlag, 2010, pp. 59-79.

[11] L. W. M Wienhofen, "Using Graphically Represented Ontologies for searching Content on the Semantic Web", in: *Visualizing the Semantic Web. XML-Based Internet and Information Visualization*, V. Geroimenko, C. Chen, Eds., London: Springer-Verlag, 2010, pp. 137-153.

[12] H. Dudycz, "Heuristic evaluation of visualization in the semantic searching economic information. The Comparative Analysis of Four Experiments", *Information Systems in Management*, vol. 2, no. 3, 2013, pp. 194-206.

[13] B. Smith, Ontology and Information Systems, 2010, http://ontology.buffalo.edu/ontology%28PIC%29.pdf

[14] A. Gomez-Perez, O. Corcho and M. Fernandez-Lopez, *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*, London: Springer-Verlag, 2004.

[15] F. N. Noy and D. L. McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology, 2005 http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html.

[16] H. Dudycz, *The topic map as a visual representation of economic knowledge* (in polish), Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, 2013.

[17] H. Dudycz, "Approach to the conceptualization of an ontology of an early warning system", in: "*Information Systems in Management XI. Data Bases, Distant Learning, and Web Solutions Technologies*", P. Jałowiecki, P. Łukasiewicz, A. Orłowski, Eds., Warsaw: Warsaw University of Life Sciences, Department of Informatics, 2011, pp. 29-39.

[18] M. B. Almeida and R. R. Barbarosa, "Ontologies in Knowledge Management Support: A Case Study", *Journal of the American Society for Information Science and Technology*, no. 10 (60), 2009, pp. 2032-2047.

[19] I. Sommerville, *Software engineering*, 9th ed., Harlow: Addison-Wesley, 2010.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1513–1519

DOI: 10.15439/2015F221
ACSIS, Vol. 5

# Organizational Impacts of Enhancing a BI-Supported Performance Measurement System on the Israeli Police

**Marina Vugalter**
Ben-Gurion University of the Negev,
Department of Industrial Engineering and Management.
POB 653, 8410501, Beer-Sheva, Israel
Email: vugalter@post.bgu.ac.il

**Adir Even**
Ben-Gurion University of the Negev,
Department of Industrial Engineering and Management.
POB 653, 8410501, Beer-Sheva, Israel
Email: adireven@bgu.ac.il

*Abstract* - **Performance Measurement Systems (PMS) have long captured the attention of organizational behavior and information systems (IS) research. The PMS in the study was implemented by public police forces, using advanced Business Intelligence (BI) technologies. The study examines the impact of enhancing that PMS, through analysis of the metric results over an 8-year time period that covered a transition between two major system versions. The analysis results indeed show a significant impact of transitioning to the new PMS in most (75%) performance metrics. A noticeable impact of the transition is the temporary performance decline, followed by some improvement that can be attributed in part to the redefinition of some metrics. Further, the results confirmed the preliminary assumptions that the improvement in the measured performance is positively and significantly associated with human-resource allocation; however, with some mediation effects of the crime category and the organization unit**.

## I. Introduction and Background

Today, PMS are broadly adopted across all industries and business domains, and their implementation is often supported by Business Intelligence (BI) technologies [1], [2]. PMS implementation is driven by the notion that consistent and well-organized measurement, reflecting past behavior and current state of organizational units, can serve as an important tool for supporting decision-support and driving continuous performance improvement [3]. PMS may back these goals by providing solutions for gathering, analyzing, and distributing relevant information [4]. Much of the PMS research so far is non-empirical, and not supported by rigorous evidence [5]. Further, PMS studies mostly focus on for-profit businesses [6], but less on organizations that are not necessarily driven by profit goals, such as governmental agencies or community-based groups [7].

This study explores the impact of a PMS in a real-world setting, within a governmental agency – the Israeli Police Forces. The research scope is examining the transition between two PMS: the “MENAHEL” that was used between 2006 and 2010, and the “MIFNE” that replaced it in 2010. Both PMS were implemented with advanced Business Intelligence (BI) technologies and embedded a number of performance metrics that reflect the police strategy and activities. Beyond enhancement of BI technologies, a key difference between the systems was the re-design of the metrics structure. While the "MENAHEL" offered a large number of metrics (over 150), some with low relative weight, the "MIFNE" offered a reduced and more focused set that was developed based on the knowledge and insight that were gained during the operation of the "MENAHEL". The focus of “MIFNE” is not the quantity of measures, as the quality and essence. The newer system exposes police officers not only to performance measures in their own stations, but in other stations as well for the sake of comparison. Further, the newer performance measurement system serves not only police stations, but rather all organizational police units. One new main module that is part of  “MIFNE” is “internal services”- as part of the comprehension of the importance of police officer’s and his family’s social satisfaction. Moreover, “MIFNE” is targeting every organizational unit specifically, compared to “MENAHEL” that defined a uniform target for all units. The new PMS enables to drilldown and investigation from high-level information to more detailed. It enables identification of exceptional trends or events that required an attention. According to research and statistics unit of the police, "MIFNE" brought a significant change in perception as part of a 3-years change program named “MIFNE” that has been started in 2011. Main guideline of this program is that public trust of the police, has an important and critical rule. In order to focus all organizational units on achieving “MIFNE” strategy, “MIFNE” program is based on a unique PMS named “MIFNE”.

A question that motivates this study is whether or not the transition to a new PMS promoted the desired improvement in police-units performance. Was the influence positive, or rather negative in some cases?  Is the influence moderated by certain characteristics of the organizational unit? This study suggests that these questions may have important implications for performance measurement in the Israeli police as for the adoption of PMS in other organizations.

This work was not supported by any organization

978-83-60810-66-8/$25.00©2015, IEEE

## II. Model Development

The transition between the two PMS systems raised questions that this study aims to explore: did the transition indeed gained the desired effect on performance measurement? What factors may possibly affect the actual contribution of replacing a PMS? One could expect that the actual contribution will be affected by factors such as the size of police station, socio-economic characteristics of its region, or the activity or crime category that the performance metrics reflect. These questions are reflected in the model that underlies this study (Figure 1), which is presented next.

The model observes the following constructs:

**Performance (P):** The dependent variable observed is the organizational performance as reflected in the values of the PMS metrics. Positive and/or negative trends in metric values reflect the influence transitioning to a new PMS and possibly other factors.

**PMS Replacement (R):** Transition to a new PMS is expected to have a significant impact on performance. Accordingly, the first set of hypotheses reflects possible manifestations of that impact:

- *H1a:* PMS Replacement will significantly affect performance.
- *H1b:* PMS Replacement significantly improves performance over time.
- *H1c:* PMS Replacement may result in a temporary performance decline during the transition period

***Resources Allocation (A):*** An alternative, or possibly complementary, explanation to the effect of PMS replacement on performance is the allocation of resources. Performance could have been enhanced not only by PMS replacement, but also by extending human-resource allocation, where the effect is expected to be positive:

- *H2:* Increasing human resource allocation will improve the measured performance

***Socio-Economic Ranking (S):*** The common assumption in literature is that socio-economic ranking may influence on the dependent variable in some researches as in politics and government [8] and entrepreneurship [9]. Certain demographic characteristics, such as the socio-economic ranking of the region in which the police station acts, are expected to moderate the effect of PMS replacement on the measured performance. We therefore assume that:

- *H3a:* Stations that act in regions with different socio-economic ranking will show difference in the influence of replacing PMS on the performance

***Performance Category (PC):*** PMS commonly arrange the performance metrics in clusters. For example, in PMS that are based on the Balanced Score-Card (BSC) methodology [10], the measurements are categorized under four key clusters. Such categorization could be detected in the police PMS as well, and the model assumes that this categorization will have some effect:

- *H3b:* Different performance categories will show different influence of PMS replacement on performance.

***Crime Category (CC):*** The police typically characterize different types of crime along different categories. The model assumes that different crimes category will show different effect of replacing PMS on performance:

- *H3c:* Different crime categories will influence of PMS replacement on performance differently.

***Police Station (PS):*** There is a different effect on measures among different police stations so stations differ from each other in PMS enhancing. Considering common qualities of police stations assist to divide them to clusters. Each Cluster is differ from other in internal and external characters. This examination helps assessing whether or not performance measures are effected from internal (HR allocation, socio-economic ranking) and external (crime records) police stations characteristics. The assumption is that the influence of replacement PMS on performance is expected to be different between different police stations. The model assumes that different police stations will show different effect of replacing PMS on performance:

- *H3d:* Influence of PMS replacement on performance is different between different police stations.

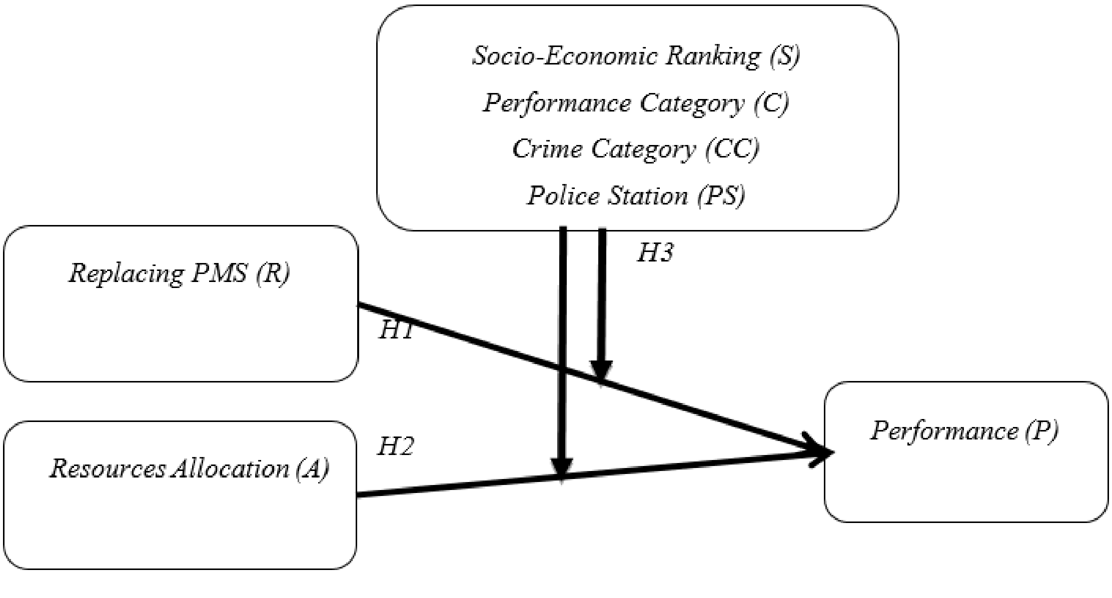


Fig. 1 Research Model

## III. Data Preparation and Preliminary Analysis

The study was based on performance measurements that were retrieved from the police database, with some additional pieces of data that were retrieved from other sources. The measurements covered an 8-year period (January 2006 to December 2013), 146 organizational units (districts, regions, special units, etc.), and 33 performance metrics - 8 annual measures, per metric and per unit. The 33 metrics can be classified to 3 main categories: a) crime records, b) charges, and c) arrests. Each category includes 11 crimes that could be divided to exposure offenses versus the rest. Exposure offenses are crimes with no previous complaint or report; hence, exposed only by initiated activity. Exposure-based crimes are considered to be more challenging to handle than the others; hence, get extended attention by the police. One of main goals the police has put as a focus in front of her is increasing the probability of the criminal to be caught in exposure offences. 3 crimes out of 11 are considered as exposure offenses: weapons, illegal stay and drugs.

Another data resource that was used includes human resource allocation for each organizational unit per year. The datasets also contain socio-economic ranking, which was calculated as an average for all cities and regional councils' socio-economic ranking that under the responsibility of a certain police station. After removing records with missing or inconsistent values, and keeping only police stations that have full information of HR allocation and performance measures over 8-year period- 2006-2013, 60 valid stations remained, and were included in the analysis. The final dataset include one annual record per police station.

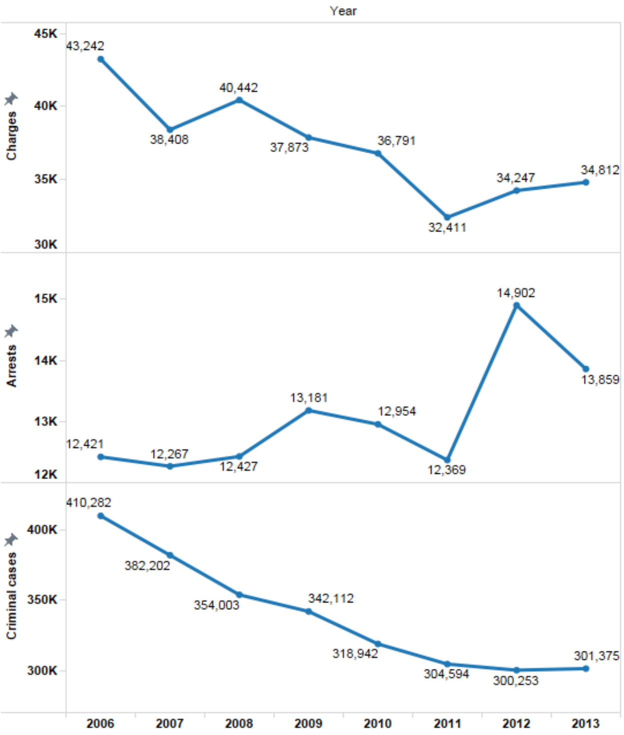


Fig. 2 Preliminary Performance Analysis

A preliminary analysis of 3 main categories: criminal cases, charges and arrests shows an improvement in metrics after the switch to the new system (2012); however, with some period of instability during the transition in 2011 (Figure 2). In addition, the ratio between arrests and crime records has a positive trend, and the ratio between arrests and charges is increasing as well over time. Among 15 performance measures out of 33 (45.45%) there is an increase in performance following the switch and among the 33 measures, only for 2 (6%) transition period (2011) is followed by improvement in performance, the rest of measures (94%) are followed by decline in performance. Furthermore preliminary analysis shows a positive relationship between HR (number of police officers in a station) and performance measures: crime records, arrests and charges. A preliminary examination shows a strong correlation between crime records and arrests especially (Figure 3); however the correlation with charges measure seems to be weaker.

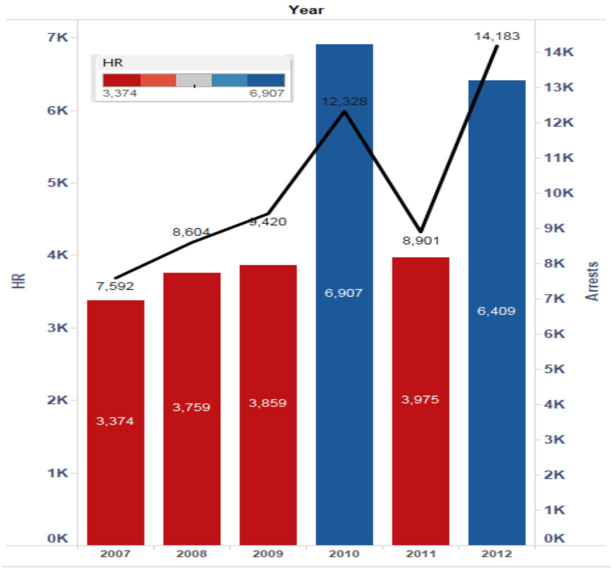


Fig.3 Arrests vs. Investment in Human Resources

## IV. Evaluation Results

This chapter reviews the analysis results of the model that is describes in chapter 3.

*Testing Hypothesis H1: The Impact of Enhancing Performance Measurement Systems*

The first hypothesis (H1) suggests that replacing the PMS will result in significant decline in performance during the transition period and also significant performance improvement following the replacement; the results support it partially. The datasets that were examined for this hypothesis is 3 year period between 2010 and 2012. In 2010 the former PMS was active. The transition to the new PMS system began in 2011. Then, the first version was active, followed by a year of the actual transition. By 2012, the transition was over and the new system became live in the organization.

The three derived hypotheses - H1a, H1b, and H1c - were tested by examining the rate of measures that had a significant effect (P-Value < 0.05) (Table 1). The H1a assumption that measures show significant change over time is largely supported, as for 25 out of 33 measures (75.7%). The H1b assumption that measures will significantly improve following replacement wasn't supported- 24 out of 33 measures have no significant improvement (72.7%). However most of the measures that were improved significantly are included in the arrests category. The H1c assumption that measures will significantly decrease during transition is also not fully supported. 31 measures among 33 had a decline in performance during the transition period (93%), however, 10 only (30%) had a statistical significant decrease.

TABLE 1.
CHANGE IN PERFORMANCE MEASURES (H1)

| Hypothesis | Significant measures | Ratio |
|---|---|---|
| H1a | 25 | 75.75% |
| H1b | 9 | 27.27% |
| H1c | 10 | 30% |

*Testing Hypothesis H2: The improvement in performance will be positively affected by human-resource allocation*

H2 suggests that performance will be improved with higher human-resource allocation. The results support this assumption significantly. A preliminary analysis shows consistently-high correlation between HR allocation and performance improvement in three key metrics (Tables 2).

TABLE 2.
CORRELATION BETWEEN HR ALLOCATION AND PERFORMANCE (H2)

| Crime Metric | Correlation with Crime Records Performance | Correlation with Charge Performance | Correlation with Arrests Performance |
|---|---|---|---|
| Regular violence | 83.58% | 57.98% | 61.22% |
| Aggravated violence | 73.7% | 62.1% | 61.70% |
| Total violence | 83.61% | 59.89% | 60.86% |
| Property | 63.81% | 51.86% | 53.86% |
| Sex | 54.67% | 43.60% | 39.73% |
| Regular attack | 78.74% | 54.45% | 56.56% |
| Aggravated attack | 69.53% | 61.20% | 50.67% |
| Weapons | 1% | 22.1% | 0% |
| Illegal stay | 11.2% | 0% | 21.0% |
| Drugs | 48.52% | 57.2% | 30.30% |

A more detailed analysis shows that the extent of improvement varies between performance types. For some measures (e.g., violence, aggravated attack, regular attack, etc.), the correlation is significant (P-Value < 0.05), while for others – "exposure offenses" (e.g., illegal stay, weapons, drugs trade) it is less significant (P-Value > 0.05). The most significant difference is between exposure offenses and the others. As seen in table 2, among the 30 indicators that demonstrated in 3 categories (a table is presented for each category), 7 (70% of all crime measures) have a significant correlation with HR allocation in all categories. All 3 exposure measures (30% of all crime measures) have a low and not significant correlation.

*Testing Hypothesis H3a: Station will show difference in the influence of replacing PMS, based on the socio-economic of the region under their control.*

Hypothesis H3a, suggesting that the impact of replacing the PMS on performance is moderated by the socio-economic ranking, is supported in part. A repeated-measures ANOVA test shows influence of replacing the PMS on performance measures for the 3 key measures (H1a). The test confirmed the assumption H1a with high significance (P-Value = ~0). However significant influence with respect to the interaction with Socio-Economic Ranking (H3a) could be found among 17 measures out of 33 (~50%). In crime records category, high significance (P-Value = ~0) interaction was found in 7 out of 10 crimes categories (~70%), however for arrests category only among 3 crimes out of 10 (30%) the interaction is significant and for charges category among 5 out of 10 can be seen a high significant interaction (50%).This interaction is found as significant in all categories for crimes like violence and attack more than other crimes. The main conclusion that derived is that different stations will show difference in the influence of replacing PMS, based on the socio-economic characteristics of the region under their control and crime category.

TABLE 3.
RATIOS BETWEEN SIGNIFICANT PERFORMANCE MEASURES (H3B)

| Impact | Category | Significant Metrics | Ratio |
|---|---|---|---|
| Total Change | Crime records | 7 | 70% |
| | Charges | 7 | 70% |
| | Arrests | 9 | 90% |
| Increase | Crime records | 1 | 10% |
| | Charges | 1 | 10% |
| | Arrests | 7 | 70% |
| Decrease | Crime records | 5 | 50% |
| | Charges | 4 | 40% |
| | Arrests | 1 | 10% |

*Testing Hypothesis H3b: The influence of PMS replacement on performance is will differ between metric categories*

The third hypothesis (H3b), is suggesting that the impact of enhancing PMS on performance will be moderated by performance category. This sub-hypothesis is derive from H1, which assumes that replacing PMS has an impact on the performance. This hypothesis (H3b) suggests that replacing the PMS will result in significant decline in performance during transition period and also significant performance improvement following the replacement is dependent on performance category; Results of examination the impact on performance by considering 3 performance categories are presented in Table 3.

For most crimes in all 3 categories some significant change in performance measures could be detected (P-Value < 0.05). For arrests, the ratio of crimes that had a significant change, is the highest- 90%, and for the other 2 categories the ratio is 70% which is pretty high as well. Regarding the increase in performance following the replacement, for both crime records and charges categories, the ratio is 10%, however for arrests category- ratio of crimes with significant improvement is 70%. It seems that there was no significant decrease in performance in transition period for none of the category in all crimes. For crime records category, 50% of crimes had a significant decline in performance, for charges category 40% and for arrests only 10% from crimes.

Table 4.
P-Value results (H3c)

| Category | Arrests | Charges | Crime Records |
|---|---|---|---|
| Regular violence | 0.00 | 0.00 | 0.00 |
| Aggravated violence | 0.01 | 0.08 | 0.02 |
| Total violence | 0.00 | 0.00 | 0.00 |
| Property | 0.02 | 0.316 | 0.05 |
| Sex | 0.588 | 0.55 | 0.115 |
| Regular attack | 0.05 | 0.01 | 0.00 |
| Aggravated attack | 0.11 | 0.14 | 0.187 |
| *Weapons | 0.06 | 0.06 | 0.01 |
| *Illegal stay | 0.00 | 0.00 | 0.638 |
| *Drugs | 0.00 | 0.14 | 0.00 |
| total | 0.00 | 0.00 | 0.00 |

*Testing Hypothesis H3c: The influence of PMS replacement on performance will be differ between crimes categories.*

The third hypothesis (H3c), is suggesting that the impact of enhancing PMS on performance will be moderated by the crime category. This sub-hypothesis is derive from H1 as well. This hypothesis (H3c) suggests that replacing the PMS will result in significant decline in performance during transition period and also significant performance improvement following the replacement, dependent on crime category; examination results of the impact on performance by considering 11 crime categories show that for sex and aggravated attack crimes there is no significant change in performance in none of the categories (Table 4). Also, for illegal stay – crime records category and for drugs charges, the change in performance is no significant; however for rest of crimes (80%) the change is significant in all categories for 6 crimes and for the rest 2, is significant in 2 categories out of 3. None of the above crimes has a significant improvement or a significant decline in all categories (Table 4).

*Testing Hypothesis H3d: Influence of PMS replacement on performance is different between different police stations.*

The third hypothesis (H3d), is suggesting that the impact of enhancing PMS on performance will be moderated by the police station. This sub-hypothesis is suggesting that the influence of replacing PMS is differ from one station to other. In order to test this assumption, all 57 police stations were divided to 6 clusters. One cluster included all “non-jewish” police stations and the rest of 45 police stations were classified to 5 clusters by k- medoids algorithm. This classification shows a partial support to the assumption; for sex crimes and illegal stay- charges and arrests categories there is a significant interaction with clustering (Table 5). Also, for aggravated attack crimes –charges, the interaction is significant. Cluster analysis was made according to measures: HR allocation, socio-economic ranking, and criminal charges in crimes: regular violence, aggravated violence, aggravated attack and property. After a cluster analysis was done, results of repeated measures ANOVA show that the only crime category with significant interaction is criminal charges. However this result is obvious since the cluster analysis is based on criminal charges measures, as mentioned above. Moreover it is clear (Table 5) that for some measures as sex and illegal stay in arrests and charges categories and also in aggravated attack charges, the influence on performance measures is dependent on the cluster attribute.

Table 5.
P_Value results- K=5 (H3d)

| Crime Metric | Crime Records | Charges | Arrests |
|---|---|---|---|
| Regular violence | 0.00 | 0.14 | 0.76 |
| Aggravated violence | 0.00 | 0.14 | 0.98 |
| Total violence | 0.00 | 0.09 | 0.90 |
| Property | 0.00 | 0.09 | 0.12 |
| Sex | 0.43 | 0.00 | 0.00 |
| Regular attack | 0.09 | 0.09 | 0.47 |
| Aggravated attack | 0.00 | 0.03 | 0.71 |
| Weapons | 0.69 | 0.985 | 0.92 |
| Illegal stay | 0.13 | 0.02 | 0.02 |
| Drugs | 0.08 | 0.498 | 0.11 |

Following repeated measures ANOVA, determining directionality of measures is defined as next step. Tukey HSD test was done in order to determine what police stations clusters achieved better performance in each year and who improved over time. Results of Tukey HSD test for *illegal stay arrests* measure show that cluster 2 is in total better than clusters 3 and 4; police stations that relate to cluster 2 perform better than those that relate to clusters 3 or 4. In addition, the performance is higher in 2012 than in 2010. Results of Tukey HSD test for *illegal stay charges* measure show that cluster 1 is in total better than clusters 4 and 5, so the number of illegal stay charges is higher in stations from cluster 1 than clusters 4 or 5. For conclude illegal stay crimes, one can say that there is significant difference between police stations that relate to different clusters in performance. Regarding *sex arrests* measure, cluster 1 is seen to be better than 4 and 5 in 2010-2012. Moreover, cluster 2 is better than 3,4 and 5 and 3 is better than 4 and 5. Police stations that related to cluster 3 improved in sex arrests measure significantly over time. Regarding *sex charges measure*, the results show that clusters 1 and 2 have higher performance than 3, 4 and 5. Moreover cluster 3 is better than 5 and also than 4 in 2011-2012. In sex charges, as in sex arrests, cluster 3 has improved in 2012. These results support the assumption that influence of PMS replacement on performance is different between different police stations however it's significant with dependent on crime category.

TABLE 6.
SUMMARY OF RESULTS

| Hypothesis | | | Supported? |
|---|---|---|---|
| H1 | H1a | PMS Replacement will significantly affect performance | Strongly |
| | H1b | PMS Replacement significantly improves performance over time. | Partially |
| | H1c | PMS Replacement may result in a temporary performance decline during the transition period | Partially |
| H2 | | Increasing human resource allocation will improve the measured performance | Partially |
| H3 | H3a | Stations that act in regions with different socio-economic ranking will show difference in the influence of replacing PMS on the performance | Strongly |
| | H3b | Different performance categories will show different influence of PMS replacement on performance. | Strongly |
| | H3c | Different crime categories will influence of PMS replacement on performance differently. | Partially |
| | H3d | Influence of PMS replacement on performance is different between different police stations. | Partially |

## V. DISCUSSION

This paper had discussed the implementation and enhancing of a PMS in the Israeli public sector organization.

The literature and case data presented clearly shows first, the importance of updating and enhancing performance measurement systems in organizations so that they change over time, and second the factors that affect the performance measures as a result of enhancing process. There is not a significant evidence in the literature to enhancing PMS process in organizations, and in particular not in the public sector. This paper discusses many of these issues in the context of case study data in Israeli police relating to performance measurement systems evolution. A considerable amount has been written about design and implementation of performance measurement systems in organizations. Moreover one can find many researches about performance measures methodologies, success and failure factors and PMS development over time; however, there is a little discussion in the literature of what are the influences of enhancing PMS, and in the public sector even less. Israeli police data show clearly that enhancing PMS may have a strong impact on performance- some decline during transition period and improvement after it.

It is also clear that for performance measurement systems to enhance effectively there are factors that an organization must consider as internal characteristics of organizational units and measures categories. Reviewing these factors is an important stage in management enhancing PMS process in organizations. This research provides a view of the factors that may influence a process of enhancing a PMS and the impact of that.

## VI. CONCLUSIONS

This research examined the impact of enhancing PMS and factors that are possibly involved in that impact. The study shows that, generally, there is a significant change in performance in most of crimes' categories following enhancing PMS. The only crimes' categories that weren't affected from PMS enhancing are sex, aggravated violence, drugs trade charges and illegal stay records. The study shows that the improvement in performance following the change is mostly expressed in crimes from arrests category (an output measure): violence, attack, and illegal stay. Furthermore, for most of crime categories a decrease in performance during transition period was seen, however it's statistical significant only for 30% of measures. For violence and aggravated attack crimes there is a significant decrease in performance for arrests and charges, for drugs trade there is a significant decrease in crime records and arrests. In addition, it confirmed that performance will be improved with higher human-resource allocation in most crimes categories, except from exposure crimes; main conclusion for the police is that no matter how resource allocation is high, the impact on performance wasn't significant. We can conclude that without any consideration in size of human-resource in a police station, its performance in dealing with crimes like illegal stay, drugs trade and weapons remains the same. Police officers can learn from this study that

The quantity of the police force staff in a police station will not dramatically improve the quality of those measures, so the focus should be on the quality of HR and less the quantity. Moreover, as suggested previously, the socio-economic ranking is shown to have a strong moderating effect on performance in crime records category, however for output measures as charges and arrests, it seems that changes in performance are not affected by socio-economic ranking.

The baseline assumption that a PMS system and, in particular, a BI-supported one has an influence on organizational change and improvement in performance has been confirmed. However, it's forbidden to neglect the negative incline in performance during the transition period that preceded the improvement.

## VII. Future Research

The literature claims that there is a need to a right management of performance measurement systems in both public and non-public sectors. Many researches are focusing on developing methodologies and reviewing the literature on PMS, however little empirical research has been carried out to assess their influence on performance measures, especially in the public sector. In order to better understand performance measurement in general and in the public sector in particular, this study can be extended in a few possible directions.

Some limitations are ought to be further addresses and explored since the study explores only a limited number of moderator variables that can influence the effect of replacing a PMS. Studies focusing on the following research questions would be of interest.

- Are the moderate variables that were analysed are the only main ones that can influence performance or other variables need to be considered?
- Is enhancing PMS improves performance over long period? How performance is influenced in the long term?
- Are other public organizations show same behaviour as the Israeli police? Is the influence of enhancing PMS is specific to this case study or common in other organizations as well?
- Is the influence on performance measures is the same on other measures that weren't included in this research but are part of the PMS?

To answer most of these research questions, additional academic and empirical research are required.

## References

[1] M. Bourne, 'Performance Management:learning from the past and projecting the future.' *Measuring Business Excellence,* 2008, Vol. 12, pp. 67-72. 10.1108/01443570010330739

[2] N. Yadav, M. Sagar 'Performance Measurement and Management Frameworks- Research trends to the last two decades.' *Business Process Management Journal,* Vol. 19, No.6, 2013, PP.947-970. 10.1108/BPMJ-01-2013-0003

[3] V. Vuksic, M. Bach, A. Popvic, Supporting PM with business performance and Business Intelligence: A case analysis of integration and orchestration. *International Journal Of Information Management,* Vol. 33, 2013, PP. 613-619. 10.1016/j.ijinfomgt.2013.03.008

[4] U. Bititci, P. Garengo, V. Dorfler, Performance measurement: Challenges for tomorrow.' *International Journal of Management Reviews,* 2012, PP. 327-305. 10.1111/j.1468-2370.2011.00318.x

[5] P. Micheli, M. Kennerley .'Performance measurement frameworks in public and non-profit sectors. *Production Planning & Control,* 2005, 16(2)134-125 . 10.1080/09537280512331333039

[6] H. K .Rantanen, 'Performance measurement systems in the Finnish public sector,' *International Journal of Public Sector Management,* 2007.20(5). 10.1108/09513550710772521

[7] T. Boland, A. Fowler, 'A System Prespective of Performance Management in Public Sector Organizations'. *The International Journal of Public Sector Management,* Vol 13, No.5,2000, PP. 417-446. 10.1108/09513550010350832

[8] A. Lijphart 'Corporatism and consensus democracy in eighteen countries: Conceptual and empirical linkages' *British Journal of political science,* 1991,246- 235 . 10.1017/S0007123400006128

[9] B. Batjargal 'Social capital and entrepreneurial performance in Russia: A longitudinal study.' *Organization Studies* 2003, 535-556. 10.1177/0170840603024004002

[10] R. S. Kaplan, D. P. Norton.'The balanced Scorecard-measures that drive performance.' *Harvard Busisness Review,* 1992, 71-79. 10. 2308/acch.2001.15.2.147

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1521–1531

DOI: 10.15439/2015F279
ACSIS, Vol. 5

# TRADITIONAL and AGILE PROJECT MANAGEMENT in PUBLIC SECTOR and ICT

Anna Kaczorowska
University of Lodz
Faculty of Management, Department of Computer Science
ul. Matejki 22/26, 90-237 Lodz, Poland
Email: annak@wzmail.uni.lodz.pl

*Abstract*—**The article comprises the characteristics of traditional project management (TPM) and agile project management (APM) and indicates that when using a specific concept we should take into account the conditions of the sector in which the organization implementing or participating in IT projects is functioning. For it is not in all organizations that APM is more effective than TPM. Agility at the project level is one of the possibilities of which we should remember when seeking a tool for achievement of the organization's strategic objectives. However, such a tool becomes less effective if its use is not preceded by analysis of specific attainable benefits and conditions which have to be met to achieve such benefits. These conditions comprise, among other, the organizational and decision-making culture, projects financing method, as well as approach to change servicing, risk management or standardization of project management practices.**

## I. Introduction

AGILITY at the executive – project level was defined in the principles of Agile Manifesto [4], and in practice implemented in numerous agile methods [1]-[2]-[3], the most popular of which is SCRUM [5]-[6]-[27].

The main project benefits which when reached facilitate an agile approach involve: higher easiness to cope with variable priorities, abbreviated time to market, decreased project risk, better adjustment of the objectives of IT and business [10]-[17].

However, the use of one of the agile methods does not guarantee the appearance of the mentioned benefits in each project, or their contribution to a higher efficacy of the whole organization.

Both the public sector organizations and the ICT (Information and Communication Technology) organizations undertake the IT (Information Technology) projects. Yet, they are functioning in different conditions and implementing such projects differently [37].

IT projects exhibit many identical characteristics, as is the case with other measures. They also bear their own specificity. The information project management consists of many activities related to planning, management and control.

According to B. Lent, the information project (IT project) is the „temporary form of organization aimed at designing and performing of applications, data banks, organizational solutions, computer accessories, system platforms and other solutions within the computer science" [9].

The public sector[1] is connected with the provision of a number of public services. Thus, the public administration bears the nature of services and is functioning within the legal regulations system [12]-[15].

*The act on informatization of entities performing public tasks* [20], of 17 February 2005 (hereinafter referred to as the UINF), constitutes the e-activity of all offices from this sector.

Owing to defining in the UINF [20] the term of information project of public use (art. 3 §6) legal use of such projects was enabled and entered the road of the formal setting up projects in entities of governmental and self-governmental administration.

"Information projects should be established in this sector mostly to make available further e-services and teleinformation systems owing to which they may be provided" [38].

The ICT sector is defined in many ways by various organizations. In the simplest way it is understood as a combination of IT sector and telecommunication sector.

The core of the definition of ICT sector is International Standard Industrial Classification – ISIC [14] which applies product distinguishing that consists in specification of conditions which have to be met by the products of a given business activity in order to qualify it as an element of the sector.

ISIC defines ICT sector as a set of enterprises conducting production and services activities consisting in seizing, transferring and displaying the data and information electronically [11].

[1] According to the definition submitted by the Ministry of Finances the „Public sector is a part of the national economy which consists of:
1. State and self-governmental organizational units which are not corporate entities.
2. State and self-governmental special funds (State higher schools and State and municipal culture institutions as well as State enterprises are not included into the public sector)" (Art. 9 of the Act on public finances).

978-83-60810-66-8/$25.00©2015, IEEE

Organizations from ICT sector are usually suppliers of information solutions for the public sector. These organizations are much more experienced in using the Agile approach in IT projects implementation and may convince the public administration decision-makers to a more comprehensive use of APM in its future projects. First, however, suppliers from the ICT sector should try to gain trust through reliable indication that they had found a method to bend the rules without breaking them and that they can abide by them through co-implementation of IT projects in public administration entities.

Agile methods require such skills from project participants as: self-organization of teams, undertaking of group commitments and decisions, self-reliance, creativity and courage. These are elements of organizational culture, and not just project culture. The use of agile approach to project management in organizations from different sectors should also involve this aspect.

This paper is written from the perspective of the public and ICT sectors in Poland.

The comparison of traditional and agile project management in these sectors in Poland is based on the following criteria:

- legal consideration related to conducting of projects,
- approach to change and risk servicing,
- organizational and decision-making culture,
- project financing method,
- maturity in project management.

The public sector implemented from the private sector a form of management through the development of projects. Evaluation of the use of TPM and APM approach in Polish public administration entities was preceded by analysis of the most important legal acts [12]-[15]-[20] in this respect and positioning of our country in acknowledged European research within *eGovernment Benchmark Measurement* from the years 2004-2014 [21]-[22]-[23]-[24]-[25]-[26]-[31]-[32]-[33].

Opinions on the use of agile and traditional approach in IT project management was based on analysis of CHAOS reports [13]-[18]-[19] and conclusions from the research carried out by K. Jasińska and T. Szapiro [28]-[29] on the factors of success in project implementation processes management among Polish enterprises from the ICT sector.

## II. TRADITIONAL VERSUS AGILE PROJECT MANAGEMENT

R. Wysocki [16] singles out the following methods of project management:

- Traditional Project Management (TPM),
- Adaptive Project Framework (APF),
- Extreme project management (XPM).

The method allocation criterion is the basis for project implementation. And so in TPM such basis is a strictly defined plan. In APF the implementation is based on an earlier analysis and defining of the project structure. XPM – otherwise referred to as management in extreme conditions – is based on the principle of a fast response to ongoing changes and appropriate facing of complex, unplanned situations.

The project approach to changes appeared in public administration in Poland quite recently, so its modifications may seem unnecessary to some decision-makers. However, new proposals have already appeared on how to increase the efficiency and effectiveness of project management, therefore the earlier approach is referred to as the traditional project management, whereas the newer one – the agile project management (APM).

Both the TPM and the APM are focused on the golden triangle (also referred to as the main triad [8]) of the project management. This is an equilateral triangle whose sides are the following parameters:

- the project operations range (resulting from the project objective),
- cost (budget which is the project financial restriction),
- time (the time framework of the project implementation).

These parameters, which at the same time are the project’s main determinants, are the most important factors decisive of the success of the measure under implementation.

Assumed as a resultant of all the three parameters is the quality of the implemented project for which the project manager is responsible. It is his task to constantly monitor and improve the activities related to the project’s key parameters. Each appearance of deviations from assumed arrangements or even the risk of such deviations is reported to the project sponsor who may decide that it is legitimate to start introducing remedial plans.

The critical factors of the project’s success are strictly inter-related, they condition each other and determine the project implementation path and decide about the risk in a given project. Modification of the factors rests only with the project sponsor, and the relations between the parameters in case of such modifications should be as follows:

- lowering of the costs of the project – reducing the range of the project,
- abbreviation of the project implementation time – raising the costs of the project or/and decreasing the range of the project,
- increasing the range of the project – increasing the costs of the project and lengthening the project implementation time.

As early as at the stage of determining the project concept, the first requirements as to the project key parameters should be agreed between the sponsor and manager of the project. Furthermore, their priorities should be determined as early as possible. Such priorities indicate which of the factors may be subjected to more changes and which should rather remain unchanged. Therefore, the factors should be balanced together, because all unbalanced proposals of the project limitations cause a risk that the measure will not be performed at the set time, cost or range.

At the TPM the project manager does his best to determine, describe and „freeze" the project range so as to base on it the determination of the project time and budget. He carries out laborious analyses at the beginning in order to prepare a detailed plan and avoid as many changes in the future as possible.

When the change is unavoidable, it is subject to analysis and once approved the changed elements are added to the project range. This usually leads to extension of the implementation time and increase in the project budget.

At the APM the time and costs are considered to be constant parameters, whereas management is aimed at an appropriate adjustment of the range to the current situation. Simultaneous creation of the best possible conditions for the work performance yields very good effects and increases the teams' creativity on both the business and technical side.

The project implementing team's expectations differ from those of the future user. Contractors aim at the compliance with agreed parameters of the project, to the extent consistent with the specification. The user represents a different point of view. He expects the actual use of the created and implemented SI which will effectively support the organization's basic activity, pursuant to both identified and unspecified needs.

The project management involves also activities connected with the best use of available resources. The primary resource are the people, other resources are various tools, equipment or premises. Competent organization of the people's work, methods of motivating to everyday productivity is an important element of the project implementation and constitutes an indispensable condition of maintained consistency with the basic parameters of the project.

APM as compared to TPM is more similar to the not quite perfect human nature, because it follows the principle: „ Do not change people. Change systems". This is one of the greatest „soft" differences between the traditional project management and agile project management. In APM the so called time boxes were used, i.e. constant segments of time, in the form of sprints covering the periods from 1 to 4 weeks. The determined length of the sprint must not change even by one hour. This was meant to avoid the syndrome of the student who performs his tasks at the very end in situations when the objective is considerably remote. On the contrary: in APM the objective is close and precisely determined in time, while as the sprint's effect the ready-to-use value arises.

Truly enough, there were attempts to use similar mechanisms in TPM, such as for instance the division of the project life cycle into phases. But completion of a phase at a strictly determined time is not perceived restrictively [36].

In the traditional approach the push concept is used, whereas in the agile concept – the pull concept is used. In push the tasks are allocated to the contractors well in advance. In pull – to the contrary – the sprint tasks are selected by a self-organized team whose individual members may use their preferable method of implementation. This is an element which significantly contributes to the contractor's higher commitment. The team does what is most important in a given moment of the project for the client's best possible results. Every short sprint is a consequence of the following steps: planning, implementation, survey of effects and retrospection consisting in collection of the so far acquired experiences. This is motivated by the idea of faster learning and adaptation to new situations in result of better working conditions and increased involvement and creativity of the team.

The TPM and APM approaches perceive the risks to the project differently. For the traditional management the highest risk is exceeding the planned date of the project completion, while for agile management – the lack of involvement of the business party. Every sprint is planned with the Product Owner (PO) who is a business representative. PO must know exactly what is most important during planning of a concrete sprint at a given moment of the project and share such information with the team. An adverse aspect of the fact that the PO may make plans in short segments of time is an unexpected lack of PO, because then the team cannot continue its work. What is positive is that he may introduce changes appropriate to the current situation in the project environment [27].

There is a great difference between the two approaches in the project objective. In APM the project objective is not its termination on time and within the planned budget but what matters most is the product and its development. The point of view in this approach is oriented to execution of the product, and not the project itself. This is particularly important when a change characterized by a high innovation is planned [7].

The project team in APM consists of the same roles, apart from the project manager whose competencies were assigned to the PO. All roles have one objective, which is the project product that will „enrapture" the client. PO draws up a list – arranged according to the business value - of functionalities necessary for implementation of the product or functioning of the service, which is referred to as the backlog. The lower the position of a given functionality on the list, the lower its value and accuracy of description.

A complete backlog is delivered to the project's team's meeting at which the sprint is planned. The team is servicing only that segment of functionality from the backlog which it is capable to perform within one sprint.

On the other hand there is a stepwise implementation of the sprint which ends in acquisition of active single functionality (or functionalities) (it is not yet the whole product – e.g. SI because of the lack of other functionalities contained in the backlog) whose activity may be checked by the PO. At the same time PO is capable to recognize and prepare the next required functionalities. In the last step of the sprint – retrospection the team discusses its activities and the process and selects and can estimate the most important elements which might be improved in the next sprint[2]. With each consecutive retrospection the team empirically knows better and better the pace of its work and is able to estimate with a higher accuracy the time needed for implementation of the functionalities remaining still in the backlog.

The product owner has the right to introduce changes in the backlog according to the rule of choosing the required and most needed functionalities for the client, which may be accomplished within the fixed time and budget. The PO may also take a decision to withhold implementation of the effects of several sprints till the moment of creating a set of functionalities which are important for the client.

In each consecutive sprint the consecutive functionalities of the product are performed or those performed earlier are changed according to the feedback information from stakeholders. Such a mode of work requires a lot of commitment from the PO during continuous updating of the backlog and when work is allocated to the team. The backlog updating moment is very important because it is exactly at that time that the product owner may consider the product to be good enough so that there is no need to develop it any further. The PO may also give up the least important functionalities of the product to replace them with other functionalities, even such which initially were omitted in the backlog.

## III. Traditional and Agile Project Management in Public Sector

All public projects in Poland may be established owing to the *The act on informatization of entities performing public tasks* (UINF) [20], but they are also subject to the *Public Procurement Law* [12] hereinafter referred to as the UPZP. Most of the limitations and difficulties in project implementation result from the entries of that particular act.

Project implementation according to one of agile methodologies assumes the lack of precise specification of the final product parameters. Public administration institutions are obliged by the act of the Public Procurement Law [12] to specify in detail the subject of procurement (at SIWZ) prior to selection of the contractor.

It is impossible to determine complete functionality of the solution at the beginning of the project. However, no matter which approach to the project management is applied, we should absolutely specify precisely the key assumptions and parameters. It is of utmost importance to specify the project purpose in a measurable way which allows to check if it is achievable. To put it generally, we should specify as precisely as possible what has to be done, how it will be checked, if it is implemented, and then to offer relatively much freedom as to how it should be accomplished. Public projects at the very start demand a very detailed description of the solution, thereby making it difficult to introduce subsequent changes.

For the choice of the contractors of the largest public projects in Poland the same criterion is used, i.e. the criterion of experience and knowledge, as is the case with less complicated works. Success in every project depends not so much on technical skills, but rather on the offerer' skills within the project management. Meanwhile, the national public procurement law demands specification of the procurement subject according to European, Polish or international standards, but does not require any description of offerers' qualifications within the project management. This creates quite a high probability of choosing a company which will not cope with implementation of a huge project or will cause selecting a foreign contractor who does not know Polish conditions.

The more and more prevalent (irrespective of the UPZP entries) requirement that the project managers have certificates confirming their management qualifications meeting the PMI or PRINCE2® standards is insufficient. Legal sanctioning of such a requirement would not quite solve the problem either, because the project manager may undertake activities exclusively within the procedures of the public administration entity for which he works.

A severe problem connected with the UPZP is that the Terms of Reference (SIWZ) do not include the risks which may occur during implementation of a given project. The lack of if only preliminary estimation of the risks by the public sector contracting authority deprives the potential contractors from the ICT sector the important knowledge about the realities of implementation of a future project.

SIWZ should be supplemented not only with a list of risks to which the project is exposed, but also with indication for which of them the responsibility rests with the contracting authority, for which – with the contractor, and for which both participants of the project process are responsible. Exclusively the contractor should be held responsible for the risks connected with direct project management, including those resulting from cooperation

[2] Improvement - in a consecutive sprint – of the most important elements from the previous one is referred to as implementation of the Lessons Learned.

with sub-providers. The list of risks included in the SIWZ by the contracting authority would be specified by the contractor within the submitted bid. The contractor would be also obliged to estimate the consequences of all identified risks for the budget and schedule of works. Leaving the risk servicing almost exclusively on the contractor's side will probably make the offerers try and include their cost in the price they suggest. The cost of risk servicing should constitute a separated part of the bid's budget, to be launched at the moment a given risk occurs [7].

The use of agile approach to management of IT projects implemented in public sector is also exposed to difficulties caused by the principles of the provider selection and cooperation with him and also evaluation of the extent to which he managed to implement the project.

Presently, due to legal limitations, the complete departure from the cascade model of public IT projects implementation is impossible.

In the ICT sector we deal with clients, while in the public sector – with applicants or tax-payers. Contrary to the clients, the tax-payers have no choice and may not go to another provider when the product does not meet their expectations [3]. Consequently, the market pressure on improvement of public services is insignificant.

Introducing of project management into the public sector in Poland enables informatization of this sector, but first of all causes the development of eGgovernment, i.e. administration providing services electronically.

Analysis of the Polish eGovernment position in the research within the *eGovernment Measurement* [21]-[22]-[23]-[24]-[25]-[26]-[33] covering the years 2004-2013 (Table I and Table II) points to extensive backwardness of the public sector in Poland, as compared to other countries. This problem may be solved by effective management of IT projects financed with the EU support.

Successful IT project management in public sector is currently very important because more and more EU funds are obtained. It should be underlined, that not only the amount of EU funding is a measure of the beneficiaries success but the issues of effective project management ought to be the area of greatest engagement of executive management.

Before 2013 Poland participated six times in the *eGovernment Benchmark Measurement* [33]-[38]. For the first three years our country had one of the last positions on ranking lists, both in respect of the index of complete availability on-line of 20 basic public services and index of their maturity (Table I).

The last but one report on public services online has subtitle *Đigital by Default or by Detour* [26]. It states that public services must be designed and delivered not in administration-centric but in a customer-centric manner. The new benchmark framework was used in order to aligned it with the policy priorities of *the Digital Agenda for Europe* [31] and the current *eGgovernment Action Plan* (AP). One of four priorities of AP is „results driven government". „The results are based on a survey sample of more than 28 000 internet-using respondents in 32 countries who were questioned for this study" [26] and were named EU-27+.

TABLE I.
20 COMMON PUBLIC SERVICES IN POLAND IN RANKINGS
OF E-GOVERNMENT AUTHORIZED BY THE EC

| **The year when the report was prepared** | **Poland's position in view of full online availability of services** | **Poland's position with regard to services maturity** | **Number of states participating in the study** |
|---|---|---|---|
| 2004 | 26 | 27 | 28 |
| 2006 | 25 | 26 | 28 |
| 2007 | 30 | 30 | 31 |
| 2009 | 25 | 24 | 31 |
| 2010 | 19 | 20 | 32 |

„Results driven government" evaluates the efficiency and effectiveness of government on the basics of synthetic indicate *Effective Government* (which building is presented in Table II; counted according to formula: *Effective Government* = average of *(eGovernment efficiency* and *eGovernment impact)* * percent of eGovernment users scaled on 100). Value of this indicator shows the extent to which governments succeed in satisfying their online users.

The synthetic indicator *eGovernment efficiency* is an average of e-government users satisfaction and fulfillment of expectation. While *eGovernment impact* is average of *Likelihood of re-use* and agreement with *Perceived benefits.*

TABLE II.
INDICATORS BUILDING THE EFFECTIVE GOVERNMENT BENCHMARK
AND VALUES OF ITS COMPONENTS
FOR POLAND VERSUS EU-27+

| **EFFECTIVE GOVERNMENT – Poland / EU-27+**<br>**EFFECTIVE GOVERNMENT – 18% / 26%** | | | |
|---|---|---|---|
| **eGovernment efficiency** – Poland / EU 27+<br>**eGovernment efficiency** – 39% / 40% | | **eGovernment impact** – Poland / EU 27+<br>**eGovernment impact** – 64% / 71% | |
| **User Satisfaction – Poland / EU-27+** | **Fulfillment of expectations Poland / EU-27+** | **Likelihood of re-use Poland / EU-27+** | **Perceived benefits Poland / EU-27+** |
| Top level satisfaction scores (8-9-10) across 19 life situations | % „better" and „much better than expected" | % „likely" and „very likely" to re-use | % „agree" and "strongly agree" with 8 perceived benefits" |
| 37% / 38% | 42% / 41% | 83% / 86% | 45% / 56% |

The latest report from this series (subtitled *Future-proofing eGovernment for a Digital Single Market*) [31]-[32] accesses the state-of–play of the implementation of digital services in 33 European countries (including all countries of the EU, Iceland, Norway, Serbia, Switzerland and Turkey) who were named EU-28+.

A two-step clustering study has been carried out to place the performance of individual countries in the national context of exogenous factors such as eGoverment demand or the environment.

The first step of this study makes it possible to determine eGovernment maturity within countries and to identify different clusters of countries with similar eGovernment maturity performance. Five clusters have been identified: neophytes, high potential, progressive, builders and mature. Poland together with Austria, Germany, Bulgaria, Czech Republic, Italy, Latvia and Slovenia were qualified for the progressive cluster. Countries in this cluster have been working on a digital approach, but there are some factors that constrain full distribution of satisfying eGovernment services and the progressive cluster should focus on removing those barriers. Policies and innovation plans in countries from this cluster should specifically address and support deployment of a citizen - centric approach to further increase use of eGovernment services.

In the second step is taken into account that eGovernment maturity is affected by different variables. At the same time, undertaking an eGovernment project could have different meanings in different countries. Therefore, it is important to understand the impact of the national context on performance. Five groups of countries with a similar context are identified, based on the values of the context variables which were defined per country (eGovernment supply, eGovernment demand and environment).

Having categorised countries in terms of both absolute performance and their relative context, it is possible to analyze peers. The cross-analysis puts the individual performance of a country in its context. The purpose of mapping absolute performance clusters with clusters of countries with a similar context is to compare peers and to identify specific policy recommendations for each country that could support policy makers in moving forward. Although the background (e.g. economic, demographic and institutional factors) of European countries varies, all countries can find good example among their own peers. All contextual groups have at least one country in the well performing high potential, mature, progressive or builders cluster.

Poland was classified as a country of Group 2. Germany, Italy, France, the United Kingdom and Spain were also included in this group. Group 2 is composed of high income countries with the largest populations (and those populations are relatively older), level of education and the take-up of the internet are in line with the EU average. The ICT infrastructure is highly developed but the level of centralisation is low.

Clustering countries by contextual group and performance Poland found at the same cluster as Denmark and Italy (our peers). Policy recommendations received by our country are following: „Compared to the benchmark, in Poland context factors that limit digitisation may be the availability of digital skills and the difficulty to coordinate the efforts of the public bodies, although these factors are not likely to jeopardize the effectiveness of an appropriate eGovernment strategy. Similar considerations can be extended to Italy, although Poland may count on higher digital skills level" [31]-[32].

For the public sector entities as clients of the ICT sector enterprises the problems in project implementation are often tantamount to multimillion losses, disturbances in their functioning, and deteriorating their image in the society.

In the EU the public sector bureaucratization is increasing. It seems, however, that in Polish offices the projects will be managed pursuant to the APM principles. This, however, will demand bending the APM rules to one's needs and becoming creative. For example: if a public administration entity is required to conduct documentation related to the process or architecture of the implemented solution, such entity may prepare a video recording of the people creating on a board some graphic models of the system process or architecture. Such video recording is much easier to implement and understand. Actually, it may be also acknowledged as some type of documentation.

In public administration the projects are mostly managed in a traditional way. Effective implementation of projects requires continuous improvements. Even the division of a complex large project into smaller sub-projects while leaving the same budget, scope and general project imposed in advance, may yield the following effects: reduced risk and feedback information loop and encouraging people to extend their competences.

The on-going improvement of the project approach in public sector in Poland directly contributes to a change in public organizations management.

## IV. Traditional and Agile Project Management in ICT Sector

According to ISIC [14], the ICT sector enterprises may conduct production (e.g. production of office machines and computers, wiring and components, production of instruments and devices used for measuring, control, testing, navigation, and for other purposes, except for industrial use) or services (e.g.: telecommunication, services connected with hardware, wholesale of computers and their components and software, electronic and telecommunication devices).

The following types of enterprises may be singled out among the ICT sector enterprises:

- providers of IT devices and (or) software
- providers of IT services (from simple repairs and servicing to solutions in the cloud and professional services),
- software developers,
- distributors – companies specialized in logistics, having preferential price conditions for some goods, reached through framework agreements with selected providers,
- integrators – companies constructing complex solutions and specialized in specific ICT technologies,
- telecommunication operators.

The companies of this sector implement projects without such legal limitations as those identified in the public sector.

Representing the contractor's party in the public project with EU cofinancing, such companies are subject to evaluation carried out by the project financing institution. The criteria used by the European Union during evaluation are: relevance, drawing the project up and its plan, efficiency, effectiveness, impact and sustainability.

The relationships between the ICT sector enterprises are characterized by a high level of cooperation, temporariness and insignificant formalization.

The ICT market is the client's market where the supply of ICT solutions exceeds the demand. The decisive party in the provider-purchaser relation is the client with whom we should continually and mutually modify the project results and adapt to the changing conditions of the project implementation.

The conditions of companies functioning in the ICT sector, where the enterprises are not so significantly limited by legal regulations and often together on the client-provider line search for the best model of cooperation largely predestine this sector organizations to conduct projects with agile approach.

In many enterprises participating in IT projects a view prevails that project implementation is an individual and unique activity, therefore the project management processes are not subject of the process management concern. This view should be verified, especially among the ICT sector enterprises, because the project implementation process is not functioning in the organization alone but is a part of a large system of all its processes.

Although the project itself is an individual and unique undertaking, yet we can distinguish various types of projects in the ICT sector, which constitute groups of undertakings characterized by similar features. In the ICT sector these may be the following groups of projects:

- - information projects – their final products mostly consist in development of software or services related to its development, service and maintenance,
- - IT projects – their final products mostly refer to computer hardware or related services,
- - telecommunication projects – their final products are mostly related to the network infrastructure,
- - electronic projects – aimed at the production of electronic devices, therefore they are colloquially called hardware devices,
- - electrotechnical projects – their final products mostly refer to electrotechnical solutions; they are also called low-signal projects; the monitoring network project may serve as such a project,
- - ICT projects – combining all the mentioned categories of projects; they are interdisciplinary but we could hardly indicate a dominant domain in them; corporate network projects may serve as an example here.

Implementation of specific groups of processes bears common features and involves resources with specific competences. Therefore, aiming at standardization of project activities should be a natural aim of every enterprise from the ICT sector oriented towards project implementation. Those enterprises in this sector which would each time start project implementation individually would be ineffective, because in each newly undertaken project they would have to build its implementation structures and acquire new resources. Understanding a project management in ICT sector enterprises as a process, conforms with the concept of B. Lent [9] and additionally analyses project management in the context of operational project activities (e.g. drawing up documentation), supporting (e.g. legal or book-keeping services) and mostly managerial and administrative activities (all activities oriented to harmonization of operational and supporting activities; they include project management). B. Lent differentiates between the project conducting process and the notion of project management itself. He means as the project conducting process the process during which the project is implemented starting from its concept, through planning processes and implementation till its successful completion. Meanwhile, the notion of project management itself refers to single processes of the project of the type of activities following a cyclical pattern of measures.

The project implementation process involves additional input and output processes connected with new projects and their course after formal closure. These activities are strongly connected with the specificity of the project implementing enterprise.

Management of project implementation processes with reference to processes going on in the ICT sector enterprise requires a favorable internal environment in which the project and process activities integration is of key importance.

Perception of project management in ICT sector enterprises as a process increases the possibilities of enterprises of arranging repeatable elements of project implementation into systematized standards which form schemata ready for multiple use. Consequently, a project implementation process arises, which on the one hand enables carrying out unique projects, on the other hand it may be improved as any other repeatable process in the enterprise [29].

The studies carried out so far on the key factors of ICT project success are important, because they allowed not only for specification of the factors of success, but they were also an attempt of identification of implementation problems and searching for methods to limit or eliminate them.

As statistics of ICT projects failures, in Poland and elsewhere in the world, usually quoted are the research results of the Standish Group (tSG). The data presented by tSG should be interpreted with reference to established criteria of the project success. The Standish Group defines project success as on time, on budget, and with all features and functions as initially specified.

Observation of changes occurring on lists of key factors of ICT projects success allows to state that systematically growing is the importance of such factors as agile management processes and activities aimed at optimization of project operations.

Agile projects are successful three times more often than non-agile projects, according to the 2011 CHAOS report from the Standish Group. Exactly the report states that: “The agile process is the universal remedy for software development project failure. Software applications developed through the agile process have three times the success rate of the traditional waterfall method and a much lower percentage of time and cost overruns” [18]. Moreover, they do not report how many projects are in their database but say that the results are from projects conducted from 2002 through 2010.

When plans are established in the initial phase of the project, there is a need to increase the control processes which verify the consistency of the plan with actual requirements of the project at a given moment of implementation. This increases formalization which in turn absorbs time and budget. The complete departure from traditional planning is impossible and encumbered with a high risk of implementing the project ad hoc – only through responding to the ensuing events.

A higher flexibility in implementation of ICT projects expresses increased interest in the agile project management approach. Agile project management is presently treated as one of the most important ways of reducing the formalization and eliminating the difficulties with preliminary planning in projects.

In the latest tSG report of 2014 [19] the IT projects were classified into three types:

- successful projects,
- challenged projects (such projects are completed and operational but exceeded the funding package and time and offer fewer functions than planned),
- cancelled projects (at a certain moment during the development cycle).

The Standish Group further segmented successful, challenged and cancelled projects by large (any company with greater than 500 million dollars in revenue per year), medium (defined as having 200 to 500 million in yearly revenue) and small companies (from 100 to 200 million dollars).

Generally, on the part of the successfully completed projects the average reaches only 16.2% for IT projects completed on time and within the funding package (in smaller companies in total in 78.4% software projects at least 74.2% of their planned features and functions will be implemented, while in bigger companies only 9% of their projects are provided on time and within the funding package) [19]. The detailed data of such cross-analysis are presented in Table III.

TABLE III.
SUCCESSFUL, CHALLENGED AND CANCELLED PROJECTS BY LARGE, MEDIUM AND SMALL COMPANIES

| TYPES OF PROJECTS / TYPES OF COMPANIES | Large | Medium | Small |
|---|---|---|---|
| **Successful** – 16.2% | 9.0% | 16.2% | 28.0% |
| **Challenged** – 52.7% | 61.5% | 46.7% | 50.4% |
| **Cancelled** – 31.1% | 29.5% | 37.1% | 21.6% |

The three major reasons that a project will succeed are: user involvement (15.9% of responses), executive management support (13.9%), and clear statement of requirements (13.0%).

Software professionals from ICT sector are getting more and more knowledgeable about agile development and are now scaling it more broadly within their organizations, as compared to what officials could do [17].

The results of the studies carried out by K. Jasińska [28]-[29] among the Polish ICT sector enterprises indicate that the most significant limitation of ICT project implementation was the company’s internal organization maladjusted to project implementation. To identify the organization’s elements which generate the highest implementation difficulties, subjected to analysis was the area of processes, organizational structures and project management methods used in enterprises.

The highest percentage (92%) of respondents among those using the organization processes or their elements for ICT project implementation, used processes of planning the solution and sale and marketing. The fewest respondents (14%) used the technology development process, which may be associated with a limited need to develop technologies during the ICT project implementation. The processes with inherent most significant limitations in the decreasing order: processes of designing the solution, sale and marketing, implementation of the solution at the client and company management.

In the planning process most prevalent were the problems associated with the lack of knowledge about the client's needs, purposeful reduction of the scope, costs and time limits, and insufficient integration.

In the process of production and sale most errors were associated with the lack of support for selling activities, improper project qualification, and taking the selling decisions by managers without suitable consulting the project team.

On the other hand, in the company management process most errors resulted from conflicts between the project and formal structures, maladjustment of the project management methods to the specificity of implemented projects and communication problems.

In the organizational structures area of research the highest appraisal was granted to the linear-staff structure as such which to the highest extent supports the ICT project implementation, whereas the lowest appraisal was that of the linear implementation of projects with functional isolation.

Within the project management methods use in the ICT sector enterprises in Poland it was found out that the most often used methods were: the Project Management Institute, and next - PRINCE2®.

The enterprises of the highest project maturity [34] are the integrators (level 4 according to H. Kerzner model [35]), and the lowest – telecommunication operators (level 1).

The results of the studies [28]-[29] connected with the critical factors of successful project and difficulties in project implementation in ICT enterprises in Poland are similar to those connected with failures of IT projects obtained by tSG and do not differ from the average results obtained by other organizations (the OASIG Survey, The Robbins-Gioia Survey or The Conference Bard Survey). This enables to use the results of the studies on IT project management, carried out by large organizations specialized in this area, in Polish ICT sector enterprises, to increase the efficiency of project implementation using the APM or TPM approach, depending on their effectiveness in implementation conditions of a specific project and conditions of the sector from which the client and provider come.

## V. CONCLUSION

The development of project management seems to reflect evolution of management as a whole. Such development started from hard aspects in conducting of projects and aims at inclusion of more and more soft psychosocial aspects at the end of the 20th and the beginning of the 21st century. The development of an appropriate system of elements of hard and soft projecting largely depends on the project manager and the team he created.

The project management is a new approach in management, whose pillar has always been information but now it is also knowledge [8]. The most effective approach to project management is the project-resources approach which involves both the resources of information and the resources of the knowledge of the project-launching organization, furthermore – the intellectual capital is worked out owing to the people who use appropriate tools to accomplish specific project processes.

The greatest benefit of agile project management, as compared to the traditional management, is an increase in the risk management quality within the project. The agile – iterative approach allows to follow the project progress with a very high accuracy and gives a possibility to react in advance in case of risks of exceeding the budget, delays in implementation or supplying a solution the functionality of which differs from that expected by the client. Advance active response to risks finally translates into benefits in various areas of the project.

Many failures related to implementation of projects (especially large projects) in public sector point to the need for a profound reconstruction of the Public Procurement Office (UZP). For the largest projects, e.g. nationwide information projects, separate terms of their commissioning and implementation should be introduced. The criterion which qualifies a given measure to the group of large projects could be e.g. the preliminary estimation of the budget of works. It should not be important whether the budget of such a project amounts to 350 or 700 million PLN; what should matter is determination of specific valuating thresholds which would be the decisive factors as to whether a given project may be included into the cluster of large projects.

The Public Procurement Office (UZP) as a central institution most directly connected with implementation of public projects should first of all determine the projects management standards[3] as well as the ways of evaluating the contractors competencies (at best according to the project maturity models), and also supervise their subsequent use, at least in the largest projects. Actually, it is not important which of the internationally recognized standards (PMBOK® or PRINCE2®) and maturity models - Capability Maturity Model Integration® (CMMI) or Organizational Project Management Maturity Model® (OPM3) [34]-[35] will be selected as obligatory on the

[3] In Great Britain even the Highways Agency which is a counterpart of the Polish General Directorate of National Roads and Highways, and not of a higher level such as the UZP, cares about the level of project management and issues recommendations for the projects of highways construction.

Polish market, because they all performed well in the project implementation practice. Perhaps the UZP will be assisted by the International Organization for Standardization (ISO) which in September 2012 issued a new standard of project management, marked as ISO 21500[4].

The Public Procurement Law should also allow to specify the scope of works during the project implementation. The cooperation between the employer and contractor at a change of the scope of works would then express a partnership approach to project implementation. Such form of cooperation between the employer from the public sector and the contractor from the IT branch is more and more popular in Hong Kong, Great Britain and the USA.

There are several differences between TPM and APM. The greatest methodological difference between the traditional ad agile project management is the approach to the project management golden triangle. Information about other differences may be obtained directly from Manifest Agile [4]. These are, for example, the following values put above other in agile project management:

- „Response to changes above following the plan",
- „People and interactions above processes and tools",
- „Cooperation with the client above formal decisions",
- „Software over extensive documentation".

Statistically [13]-[17] the project implementation in the APM model in most cases lasts from 4 to 5 times shorter, as compared to the TPM model. This mainly results from the fact the APM is used to form the most important functionalities, whereas other functionalities are skipped, which decreases the range of the project. Efficiency of work at the team's level is increased by approx. 200-300%, similarly to the probability of success. Efficiency at the team's level grows, because each member of the team may have an individual impact on the product development and soon sees the effects of his work, but at the same time the whole team works for one common objective.

The *8$^{th}$ Annual State of Agile$^{TM}$ Survey* [17] emphasizes that agile project management gives the client who orders the IT solution the following benefits:

- he has a high freedom within introducing changes and adjusting his requirements to the changing business environment,
- the investment may bring profits as early as during the project implementation,
- comprehensive outlook on the project – not only during its implementation but also during its subsequent maintenance,
- relatively low costs of the project maintenance due to a respectively high quality,
- the work performance concept which is flexible and prone to changes (the provider continually aims at creating a product ready for the fastest possible use),
- better concept on the client-provider line, which enables the provider a possibility of actual involvement through searching for more advantageous solutions, reporting one's own suggestions, sharing one's specialized knowledge and using it for the client's activities optimization.

Owing to agile project management, both the client and the provider are less trustful and want to achieve success together. The final result of the project depends largely on how much the parties trust each other, and therefore – cooperate with each other.

The traditional project management may in turn lead to the situation in which the client and provider are less trustful towards each other, while every change and not quite specified requirements are perceived as an obstacle.

Educating the clients within social cooperation on the contracting authority line is inscribed into the Agile principles. In practice, the ICT sector contractors help the public sector clients in the process of acquiring experience and understanding their role in the project. The exemplary support forms may be as follows:

- training courses organized by the provider personally or by external consultants,
- close cooperation of leaders in the role of the Scrum Master on the provider's side with the Product Owner on the client's side,
- supporting the Product Owner by the Proxy's role on the provider's side.

## REFERENCES

[1] J. Apello, *Management 3.0. Leading Agile Developers, Developing Agile Leaders*. Boston, MA: Addison-Wesley Signature Series (Cohn), 2010.

[2] J. Apello, *How to Change the World: Change Management 3.0.* Kindle Edition, 2012.

[3] J. Apello, "Agile w sektorze publicznym", *Zarządzanie projektami. Magazyn o projektach i kropka*, vol. 1(8), Gdańsk, Poland: Fundacja Instytut Rozwoju Projektów, 2015, p. 62.

[4] K. Beck, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, … D. Thomas, *Manifest Zwinnego Tworzenia Oprogramowania*, 2001, Retrived May 3, 2015, from http://agilemanifesto.org/iso/pl/

[5] M. Chrapko, *SCRUM. O zwinnym zarządzaniu projektami*, Wydanie II rozszerzone. Gliwice, Poland: Helion, 2015.

[6] M. Cohn, *Succeeding with Agile: Software Development Using Scrum*. Boston, MA: Addison-Wesley Professional, 2010.

[7] A. Kaczorowska, *E-usługi administracji publicznej w warunkach zarządzania projektami*, Łódź, Poland: Wydawnictwo Uniwersytetu Łódzkiego, 2013, pp. 67-100.

[8] J. Kisielnicki, *Zarządzanie projektami. Ludzie – procedury – wyniki*, Warszawa, Poland: Wolters Kluwer Polska, 2011, p. 15.

[9] B. Lent, *Zarządzanie procesami prowadzenia projektów. Informatyka i Telekomunikacja*, Warszawa, Poland: Difin, 2005, p. W-3.

[10] M. Łubiarz, "Zwinne organizacje", *Zarządzanie projektami. Magazyn o projektach i kropka*, vol. 1(8). Gdańsk, Poland: Fundacja Instytut Rozwoju Projektów, 2015, pp. 87-91.

[4] ISO 21500 document was based on standard PMBOK®.

[11] OECD, *Measuring the information economy, Annex 1. The OECD definition of ICT sector,* 2002, p. 81, Retrieved May 3, 2015, from http://www.oecd.org/sti/ieconomy/2771153.pdf

[12] Sejm, *Prawo zamówień publicznych,* Dz. U. 2014, poz. 1232, Retrieved May 3, 2015, from http://isap.sejm.gov.pl/DetailsServlet?id=WDU20140001232

[13] The Standish Group, *Big Bang Boom,* 2014, Retrieved May 3, 2015, from http://www.standishgroup.com/sample_research_files/BigBangBoom.pdf

[14] United Nations Statistical Commission, *ISIC Rev 3.1, International Standard Industrial Classification of All Economic Activities,* 2002.

[15] Ustawa z dnia 27 sierpnia 2009 r. o finansach publicznych, *Dz. U.* 2009, nr 157, poz. 1240 z późn. zm.

[16] R. Wysocki, *Efektywne zarządzanie projektami. Tradycyjne, zwinne, ekstremalne,* 6th ed., Gliwice, Poland: Helion, 2013, p. 29.

[17] VersionOne, *8th Annual State of AgileTM Development Survey,* 2014, Retrieved June, 24, 2015, from http://www.versionone.com/pdf/2013-state-of-agile-survey.pdf

[18] CHAOS MANIFESTO. *The Laws of CHAOS and the CHAOS 100 Best PM Practices,* p. 25, Retrieved June, 25, 2015, from http://www.versionone.com/assets/img/files/ChaosManifest_2011.pdf

[19] The Standish Group, CHAOS Report: *Twentieth First Anniversary Edition,* Retrieved June, 26, 2015, from http://www.projectsmart.co.uk/docs/chaos-report.pdf

[20] Ustawa o informatyzacji działalności podmiotów realizujących zadania publiczne, *Dz. U.* 2005, nr 64, poz. 565, z późn. zm.

[21] Capgemini, "Online Availability of Public Services: How is Europe Progressing?", in *Web Based Survey on Electronic Public Services.* Report of the Fifth Measurement, Retrieved June, 27, 2015, from http://ec.europa.eu/information_society/eeurope/i2010/docs/online_pub_serv_5th_meas_fv4.pdf

[22] Capgemini, "Online Availability of Public Services: How Is Europe Progressing?", in *Web Based Survey on Electronic Public Services. Report of the 6th Measurement,* June 2006, Retrieved June, 28, 2015, from http://ec.europa.eu/information_society/eeurope/i2010/docs/benchmarking/online_availability_2006.pdf

[23] Capgemini, The User Challenge. *Benchmarking The Supply of Online Public Services. 7th Measurement,* 2007, Retrieved June, 29, 2015, from http://ec.europa.eu/information_society/eeurope/i2010/docs/benchmarking/egov_benchmark_2007.pdf

[24] Capgemini, IDC, Rand Europe, Sogeti and DTi, "Smarter, Faster, Better eGovernment. I2010 Information Space Innovation and Investment in R&D Inclusion", in *eGovernment Benchmark Survey 2009*, 2009. Retrieved June, 29, 2015, from http://ec.europa.eu/information_society/eeurope/i2010/docs/benchmarking/egov_benchmark_2009.pdf

[25] Capgemini, IDC, Rand Europe, Sogeti and DTi (2011) Digitizing Public Services in Europe: Putting Ambition into Action. *9$^{th}$ Benchmark Measurement, in: eGovernment Benchmark Survey 2010,* 2011, Retrieved June, 29, 2015, from http://www.epractice.eu/en/library/5283331

[26] Capgemini, Rand Europe, IDC, Sogeti, Danish Technological Institute (2013) "Public Services Online. Digital by Default or by Detour. Assesing User Centric eGovernment performance in Europe – eGovernment Benchmark 2012", in *Final Background Report,* 2013, DOI: 10.2759/14318.

[27] K. S. Rubin, *Essential Scrum: A Practical Guide to the Most Popular Agile Process,* Addison-Wesley Signature Series (Cohn), Kindle Edition, 2012.

[28] K. Jasińska, „Zarządzanie procesami realizacji projektów na przykładzie sektora ICT", *Przegląd Organizacji,* no. 9, pp. 50-56, 2013.

[29] K. Jasińska, T. Szapiro, *Zarządzanie procesami realizacji projektów w sektorze ICT,* Warsaw, Poland: Wydawnictwo Naukowe PWN, 2014, chs. 5-8.

[30] European Commission, *Europe 2020: A European Strategy for smart, sustainable and inclusive growth,* Aug. 2010, Retrieved June, 26, 2015 from http://Europe-2020-A-European-Strategy-for-Smart-Sustainable-and-Inclusive-Growth.pdf

[31] Capgemini, IDC, Sogeti, Politecnico di Milano, Future-proofing eGovernment for a Digital Single Market. *INSIGHT REPORT,* June 2015, DOI: 10.2759/32843.

[32] Capgemini, IDC, Sogeti, Politecnico di Milano, Future-proofing eGovernment for a Digital Single Market, *BACKGROUND REPORT,* June 2015, DOI: 10.2759/687905.

[33] A. Kaczorowska, K. Ciach, "The effectiveness of e-government development in Poland in 2004-2013", in *Information Systems in Management,* vol. 2, no. 4, pp. 274-288, 2013.

[34] R. Wendler, "Development of the Organizational Agility Maturity Model", in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, ACSIS,* vol. 2, pp. 1197–1206, 2014, DOI: 10.15439/2014F79.

[35] H. Kerzner, *Project Management a Systems Approach to Planning, Scheduling, and Controlling,* Tenth Edition, John Wiley & Sons, 2009.

[36] R. Wendler, "The Structure of Agility from Different Perspectives" in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems,* pp. 1165–1172, 2013.

[37] D. Kisperska-Moron and A. Swierczek, "The agile capabilities of Polish companies in the supply chain: An empirical study," *International Journal of Production Economics,* vol. 118, no. 1, pp. 217–224, Mar. 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0925527308002673

[38] A. Kaczorowska, "IT projects for the development of e-government in Poland", in *Information Systems in Management XVII. The role and Applications of ICT in Digital Economy and e-Business,* P. Jałowiecki, P. Łukasiewicz, A. Orłowski, Ed. Warsaw: WULS Press, 2012, pp. 5-19.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1533–1538

DOI: 10.15439/2015F170
ACSIS, Vol. 5

# Smart Technologies for Improved Software Maintenance

Zane Bicevska
DIVI Grupa Ltd
Riga, Latvia
Email: Zane.Bicevska@di.lv

Janis Bicevskis
University of Latvia
Riga, Latvia
Email: Janis.Bicevskis@lu.lv

Ivo Oditis
DIVI Grupa Ltd
Riga, Latvia
Email: Ivo.Oditis@di.lv

**Abstract**–***Steadily increasing complexity of software systems makes them difficult to configure and use without special IT knowledge. One of the solutions is to improve software systems making them "smarter", i.e. to supplement software systems with features of self-management, at least partially. This paper describes several software components known as smart technologies, which facilitate software use and maintenance. As to date smart technologies incorporate version updating, execution environment testing, self-testing, runtime verification and business process execution. The proposed approach has been successfully applied in several software projects.***

***Keywords-Autonomic computing, smart technologies, self-managing systems, software maintenance***

## I. Introduction

Rapid development of information technologies has created systems of unprecedented complexity; some authors [1] refer to as „computing systems with complexity approaching boundaries of human ability". They indicate that the ultimate dream of a pervasive computing – billions of computing systems simultaneously connected to the internet – can quickly become unmanageable and may soon turn into evil "nightmare". The authors predict even further increase of information systems' complexity that would almost eliminate human ability to perform software installation, configuration, optimization and maintenance.

Solution to this problem certainly lies within prospects of information technologies. In complex systems operations that are too sophisticated to be managed by a regular user should be entrusted to the system itself. This can be executed by implementing respective components into software and setting environment, in which the system is used.

IBM has proposed a solution described in its autonomic computing manifesto in 2001. The main statement implies targeted development of information systems that were able to self-management thus overcoming gap between users and increasingly complex world of information technologies.

The manifesto listed four aspects of autonomic computing:

- Self-configuration - automated configuration of components and systems follows high-level policies, rest of system adjusts automatically and seamlessly;
- Self-optimization - components and systems continually seek opportunities to improve their own performance and efficiency;
- Self-healing - system automatically detects, diagnoses, and repairs localized software and hardware problems;
- Self-protection - system automatically defends against malicious attacks or cascading failures.

Achievements of autonomic computing movement during its first decade after publication of the manifesto have been explicitly demonstrated in [2], as well as in [3]. As of now, manifesto's targets have been met only to some extent.

The concept of smart technologies was created by authors [4], and its main objectives are similar to those of autonomic computing. The approach contains a set of practically applicable improvements of non-functional features to simplify the maintenance and daily use of information systems. Below are described five types of smart technologies, which need was identified in real software development projects. The proposed smart technologies cover only part of requirements outlined in the autonomic computing manifesto. Nevertheless they are suitable for practical implementations and can serve as valuable improvement of new and existing software systems.

The second chapter of this paper deals with related research and solutions. The third chapter describes the proposed architecture of smart technologies.

## II. Related works

The autonomic computing manifesto declares a vision of fully independent computer systems (not just software) that are able to self-management. It also defines evaluation criteria to check the maturity of autonomic systems [5] - from basic level (manually maintainable information systems) to completely autonomic systems that are able to function operate accordingly to guidelines set by humans.

The manifesto does not include any instructions about implementation issues, but some authors discuss ideas about essential components of autonomic systems. For instance R. Sterritt [6] describes an autonomic environment consisting of autonomic elements, which are mutually connected via

The research leading to these results has received funding from the research project "Information and Communication Technology Competence Center" of EU Structural funds, contract nr. L-KC-11-0003 signed between ICT Competence Centre and Investment and Development Agency of Latvia, Research No. 1.5 "Platform for business process description and modelling in event-oriented systems".

978-83-60810-66-8/$25.00©2015, IEEE

autonomous channels. Every autonomic element has a kernel, so called manageable component (the component implementing the business logic), and it is controlled by an "autonomous supervisor". The supervising component uses sensors and effectors, and its main functions are monitoring of internal and external states, accumulation of knowledge base and communication with other autonomic components using autonomous communication channels. A separate component in this system is so-called "heartbeat monitor" which communicates with any existing system components through autonomous communication channels and supervises the system as a whole.

The autonomic computing approach has also been criticized [7], and the main reasons are as follows:

- the lack of precise definitions;
- avoidance of the real complexity of the problem;
- ignoring of inter-componential links.

Despite these criticisms, autonomic system objectives are so attractive that there seemed to be no reason to abandon the ideas. In 2003 IBM extended the list of autonomic aspects to eight characteristic aspects [1]. The initial autonomic characteristics were enhanced by system's ability to "know itself" and manage its resources in a proper way. An autonomic system should know its environment as well as the context surrounding its activity and act accordingly – to adjust and operate in heterogeneous environment accordingly open standards - , as well as anticipate the optimized resources needed while keeping its complexity hidden. Some years later the, so called, self-management features were supplemented with new self-properties reaching a total of 24 features [3]. Continuing efforts on autonomic systems include both, theoretical research and practical implementation [2].

The concept of smart technologies created by authors [4] is consistent with the primary objective of autonomic computing. Unlike the traditional implementation of autonomic computing where universal autonomous software components are built, the smart technologies approach deals with embedding of specific system features into information systems directly but in a uniform way.

Although the smart technologies approach and the autonomic computing approach seemingly share some similarities, it should be emphasized that the smart technologies approach was developed independently. The practical results gained in IT projects provide evidence of the usefulness of the approach.

## III. Components of Smart Technologies

There are five fields of smart technologies where practical results were gained: embedded software versioning and data syncing, embedded dynamic business model, testing of external environment, self-testing, and runtime verification.

### A. Software Versioning and Data Syncing

Every successful software solution is being used and improved significantly longer than the development of its first version has taken. Information systems are in use for many years, and the software is gradually modified, updated with new features, improved to approximate to user needs.

To ensure reliability of software in long-term, the system should already in its initial development time include not only the required (customer specified) functionality, but also supporting mechanism – "updater" (see Fig. 1) for software, data structures and templates upgrading.

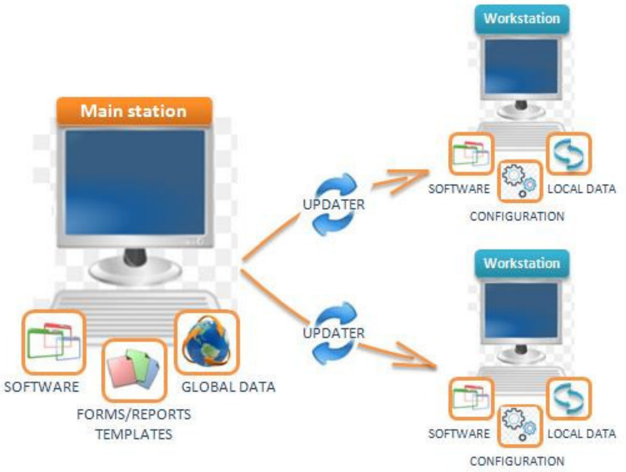


Fig. 1 Software versioning

The supporting mechanisms should be built into systems, and they should include features for deploying of new versions without any user intervention. The following should be ensured automatically during deployment process:

- check the compliance of the new software version with the external environment
- download and install a new software version
- update configuration and information about data structures, screen forms, report templates etc.
- migrate stored data into the new data structures of the database as well as the personalization and configuration data
- perform self-testing of the new system's version to check correctness of the essential system's functionality
- create backups to be able to recover the system in case of incidents

The majority of information systems today support some of the characteristics listed above, but in most cases - to a limited extent only. Authors of this paper have prototyped the characteristics in some projects, the research results are described in [8].

### B. Execution environment testing

One of the most spectacular smart technology solutions is described by the authors in [9]. It is quite common that programs have specific requirements for their successful operation at a given environment – the computer, network, operating system, etc. The proposed solution implies gathering these requirements in a "software profile" to be able to validate the execution environment before starting the information system. Such validation should be performed on demand, for instance, before each session; however, some authors propose validation during installation. Validation of execution environment allows avoiding failure in business processes in case the information system relies on properties of the environment.

Quite often, software is developed based on assumptions about other component's work, not on their specification [10], [11]. Similarly, developers sometimes assume that software, which works in development environment, will keep working after it is deployed elsewhere, hence encoding some assumptions about the environment into the program. As a result, when the software is installed in other environment, which is different from the development environment, the software may fail or work only partially correct.

The authors [12] propose a technology, which allows independent environment checks, performed by the software in order to validate if the execution environment is suitable for normal execution (see Fig. 2). Unlike the built-in test method, which validates the ability of software itself to fulfill its "contracts"; this technology measures livability of the external conditions. Only if the results of all checks are satisfactory, the program can be considered prepared for work at a given environment, otherwise the session is stopped, giving the user an explanation, why it is not possible to perform work.

A program execution profile is a document achieved when all the requirement descriptions of software are combined together. The profile can be formalized as a separate document and supplemented to typical software deliverables such as code and documentation. The main, but not the only use of the profile is validation of execution environment during program use.

The practical environment testing task is carried out by environment validation modules. Each module is an atomic unit, which enforces validation of a single type of requirement; this is done by reading information from the environment and comparing it to reference values. In a simple scenario, each requirement describes required value of some resource's attribute (for instance, data base server must be reachable). When the testing functionality of the module is invoked, it uses the information available in execution environment to do the "inspection".

To be able to modify the set of checks to be performed without modifying the program code, information about the checks (both the algorithms and reference values) must be stored outside the code. This concept is different from other approaches used in practice – both from the ones, which validate the environment straightaway after installation or updating, and from the others, which try to "hide" the checks in source code.

To be able to describe requirements regarding execution environment, a formal language is required to encode the requirements, moreover, the language must be extendable, when new kinds of requirements are defined. Such aspect complicates the construction of test coordinator, since it has to be compatible with a language, which is not fully defined during development of coordinator. The problem is solved by assigning the coordinator only the role of language syntax analysis, but the semantic analysis of requirements is performed in environment validation modules.

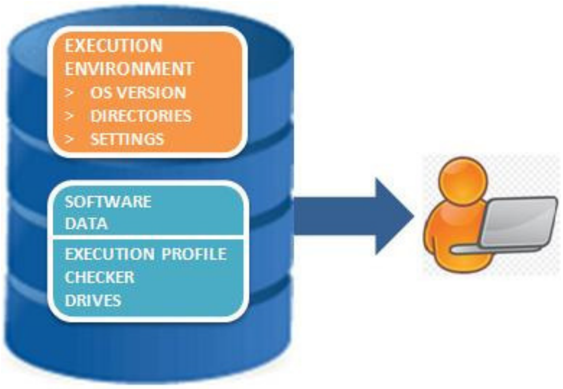


Fig. 2 Execution environment testing

The practical implementation showed that development of the proposed approach requires relatively little programming resources.

## C. Self-testing

The research of the authors [13] offers an original approach to software testing, named as self-testing. Self-testing is a software's ability to test itself automatically prior to operation, and it can be performed even in a productive environment. The self-testing feature in software is similar to hardware self-tests that are executed every time after the device is turned on. Instead of traditional testing that verifies correctness of software in testing environments using testing tools, the self-testing property is built-in software component that executes accumulated test cases using means of the information system. It helps to perform tests not only in testing environment, but also to verify software correctness in action with real data in production environment.

Self-testing contains two main components:

- test cases that are designed for checking of critical functions of the software
- built-in automated testing mechanism providing automatic execution of tests and result comparison with benchmark values.

Designing of test cases covering the critical functionality (lack of these essential functions causes inoperability of the whole system) is a part of requirement analysis.

Implementation mechanism of self-testing approach uses software instrumentation, and it has been offered quite a while ago [14], [15]. The idea is to supplement the source code with extra routines for self-testing purposes that are executed if the software is run in the testing mode. The points in source code where the routines are included are named as test points. Testing routines allow to monitor values of variables and to compare them with benchmark values therefore checking the correctness of the information system. Unfortunately, this solution is usable only for those information systems whose development is in the testers' influence sphere.

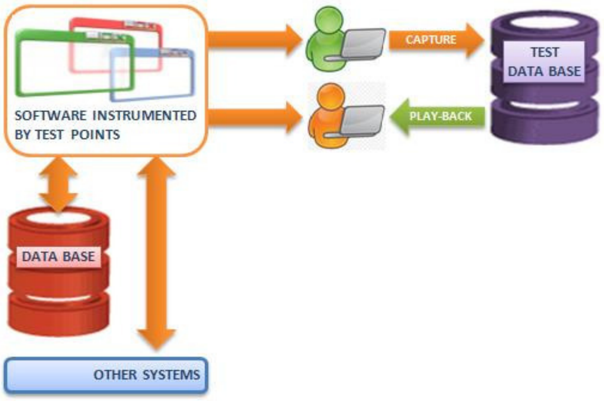


Fig. 3 Self-testing

Self-testing can be used in four modes (see Fig. 3):

1. test-capturing – running of software instrumented by test points and capturing of new test cases into test data base or editing of the existing ones;
2. self-testing – automated self-testing of software by automated execution of the captured test cases;
3. normal usage – running of information system without any testing activities;
4. demo mode – running of information system using pre-captured demo scenarios.

Comparison of self-testing implementations with automated testing tools leads to the following conclusions:

- Unlike the majority of globally recognized testing support tools, the self-testing approach offers some additional options: testing of external interfaces to other information systems and database management systems, testing in production environment, testing with the white-box method, possibility for users without IT knowledge to capture tests.
- The self-testing technology makes possible to test software throughout the whole life cycle of an information system – from early stages of software development till maintenance activities, because it is suitable for testing in all development, testing and production environments.
- The self-testing functionality should be integrated into software already during development of the software.
- The self-testing requires additional work to include the self-testing functionality in the software and to design test cases; on the other hand, self-testing saves time as repeated (regression) testing of the existing functionality is available
- Implementation of the self-testing functionality is useful in incremental life cycle models, in particular if information systems are improved gradually and maintained for many years; it is less useful in linear (waterfall) life cycle models.

Empirical studies show that 60% of information systems' problems would be possible to identify and rectify by self-testing approach [13].

### D. *Embedded business processes*

Development of software engineering tends to devote more attention to precise modelling and designing of information systems instead of extensive programming. Some researchers are even predicting development of information systems without programming at all in very near future. Business process modelling is a compulsory initial phase of every information system development project according to this concept [16].

Workflow based information systems is the area where business process modelling is an essential component for functioning of information systems. Business process of organization is described by a workflow model containing sequential business process steps – activities - together with performers of the activity, deadlines, the actual state of the object in the workflow etc. Documents and reports can also be created during the workflow execution, and this should be included in business process descriptions.

It is common to describe business processes using modelling languages. There can be used universal modelling languages or domain specific languages (DSL). When DSL is chosen, it must ensure two important features: a) the language should be easy understandable for the majority of users, b) it should include all necessary information for automated execution of workflow steps.

The first step in development of information systems is to describe business processes to be supported (see Fig. 4). A set of graphical diagrams are created using DSL, and it serves as business process model. After the model is created the information from the diagrams can be transferred to the database of an information system. The business process descriptions are embedded into the information system, and the engine of the information system can interpret information from the diagrams. Embedded business processes ensure that the information system behaves according to the business process model.

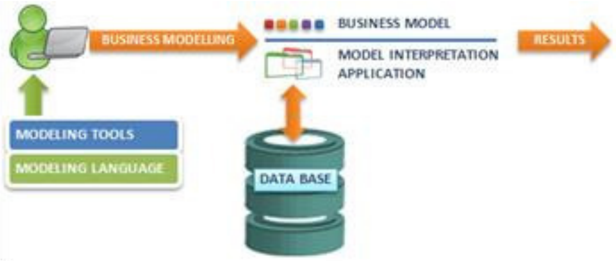


Fig. 4 Embedded business process

As practice shows [17], it is possible to create a special tool for transfer of model's data to executable application relatively quickly. The API of the graphical editor can be used to access the model's repository, to gather the information and to transfer it to applications database. This guaranties that the application operates according to the model developed in a graphical DSL. And the overall quality of the application – usability, reliability, performance etc. – is dependent on the application itself, not on the hypothetical ability of a code generator to create an application in the desired quality.

The authors have created the domain specific language BILINGVA [18] that is convenient for description of workflows. The approach was tested in practice, and particularly surprising was the positive feedback from users about the graphical representation and implementation of business processes. The diagrams served as some kind of information system's user manual that explained functioning of the information system in a more precise and understandable way than the conventional (written) user manuals.

### *E. Business process runtime verification*

From the beginnings of information systems the topical issues were: does the information system operate correctly?, are the system's results adequate?, and is the information system in the correct state in terms of the relevant business? Sometimes processes must be stopped as soon as possible after inadequate situation has occurred, otherwise more serious problems could rise [19].

Inadequate situations can be caused by many conditions. They can be caused by heterogeneous systems, which are developed at different times and used in a variety of companies. Problems can arise due to poor software quality and lack of testing. Problems may also occur due to incorrect user actions: incorrect execution of business functions, a breach of the input restrictions, or the timing and sequence of process steps.

For example, if warehouse system is not updated with payment information from accounting system timely, goods cannot be issued to customer. On the other hand, this situation is unacceptable for the customer, who has done payment according business process. Obviously, in this case there is no reason to look for errors in information system, but a person should monitor that payment data are imported timely. This is basic task of runtime verification – to verify systems execution in their runtime.

The authors [19] propose a solution for business process runtime verification (see Fig. 5). The basic idea of the solution is to run a separate verification process for each controllable business process (further – base process). Verification processes are described in DSL that has been developed in conjunction with the solution. A base process typically is executed by information systems, while verification processes should be executed on the basis of independent and external controlling software (further – a controller). The steps of the verification process are linked to base process steps and are described by events that acknowledge the execution of the each base process step.

Base processes can be executed manually or automatically by computer, and verification processes are executed independently. Each of base process steps makes some changes in the process "memory" (usually stored in database or file system). The verification controller receives acknowledgement from event agents about base process memory modifications, therefore identifying inconsistencies between the received information and the description of the verification process. If inconsistencies are detected, then they are reported to the support staff.

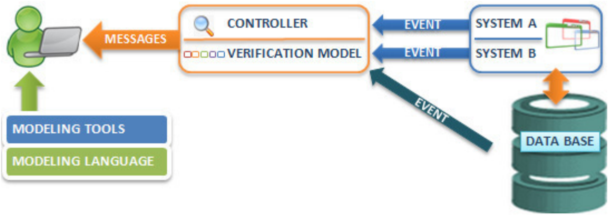


Fig. 5 Runtime verification

The solution provides a number of interesting possibilities, which bring us closer to the goal defined by ideas of autonomous computing:

- runtime verification can be done without modifications of the base process
- process verification can be added dynamically to legacy systems
- verification does not depend on modelling language used for process description, it depends only on possibility of verification agents to identity events of the base process.

Likewise, some solution limitations must be taken into account: verification mechanism can detect only those base process steps which leave some modifications in the systems „memory". Otherwise verification agents cannot work as external process, but must be incorporated into the base process.

However, it must be stressed that the proposed solution can significantly reduce monitoring load of information systems' operational staff. It automates business process runtime verification that typically is done manually and not continuously.

## IV. Conclusion

There were spent several years on research to achieve goals similar to autonomic computing – facilitating the use, maintenance and development of systems by including support components in them. The conclusions are as follows:

- several components, created using smart technologies, can provide good support in use, maintenance and development of information systems which are easy enough to implement for a small/medium size organization;
- there are quite many functions, which could be supported by respective smart technologies, for instance, data quality control, confidentiality control, built-in privacy protection [20], performance monitoring, availability monitoring, selecting environments for software compatibility testing [21], automatic testing of WEB services [22] and others;
- smart technology enabled systems are currently not very common due to the fact that these ideas are not popular enough yet; with increasing complexity of information systems, smart technologies will surely

grow in importance and will help to deal with complex system development and maintenance issues.

## References

[1] Kephart J., Chess D. The Vision of Autonomic Computing. Computer Magazine, IEEE, 2003, DOI=10.1109/mc.2003.1160055

[2] KEPHART, Jeffrey O. Autonomic computing: the first decade. In: ICAC. 2011. p. 1-2., DOI=10.1145/1998582.1998584

[3] Lalanda P., McCann J. A., Diaconescu A. Autonomic Computing: Principles, Design and Implementation. Springer, 2013, 288 p., DOI= 10.1007/978-1-4471-5007-7

[4] Bičevska Z., Bičevskis J. Smart Technologies in Software Life Cycle. In: Proceedings of Product-Focused Software Process Improvement. 8th International Conference, PROFES 2007, July 2-4, 2007 (Münch, J., Abrahamsson, P., eds.), Riga, Latvia, vol. 4589/2007, 2007. pp.262-272, DOI= 10.1007/978-3-540-73460-4_24

[5] Nami M. K., Bertels. A. Survey of Autonomic Computing Systems. In: ICAS '07: Proceedings of the Third International Conference on Autonomic and Autonomous Systems, 2007. p.26, DOI= 10.1109/conielecomp.2007.48

[6] Sterritt R., Bustard D. Towards an autonomic computing environment. In: Proceedings of 14th International Workshop on Database and Expert Systems Applications (Maríк, V., Retschitzegger, W., Stepánková, O., eds.), Prague, Czech Republic, 2003. pp.694 – 698, DOI= 10.1109/dexa.2003.1232103

[7] Herrmann K., Muhl G., Geihs K. Self management: the solution to complexity or just another problem? Distributed Systems Online, 2005, 1, vol. 6, DOI= 10.1109/mdso.2005.3

[8] Bičevska Z., Bičevskis J. Application of Smart Technologies in Software Development: Automated Version Updating. In: Scientific papers, vol. 733 (Bārzdiņš, J., Freivalds, R.-M., Bičevskis, J., eds.), University of Latvia, 2008, pp.24 -37.

[9] Rauhvargers, K. On the Implementation of a Meta-data Driven Self Testing Model. In: Software Engineering Techniques in Progress (Hruška, T., Madeyski, L., Ochodek, M., eds.), Brno, Czech Republic, 2008, pp.153-166.

[10] Arnautovic E., Kaindl H., Falb J., Popp R., Szep A. Gradual transition towards autonomic software systems based on high-level communication specification. In: Proceedings of the 2007 ACM symposium on Applied computing, 2007, pp.84-89., DOI= 10.1145/1244002.1244024

[11] Orso A., Jean M., Rosenblum D. Component Metadata for Software Engineering Tasks. In: EDO '00: Revised Papers from the Second International Workshop on Engineering Distributed Objects, London, vol. 1999, 2001, pp.129-144, DOI= 10.1007/3-540-45254-0_12

[12] Rauhvargers K., Bicevskis J. Environment Testing Enabled Software – a Step Towards Execution Context Awareness. In: H.-M. Haav, A. Kalja (eds.), Databases and Information Systems, Selected Papers from the 8th International Baltic Conference, vol. 187, IOS Press, (2009), pp. 169–179.

[13] Diebelis E., Bičevskis J. Software Self-Testing. In: Proceedings of the 10th International Baltic Conference on Databases and Information Systems, Baltic DB&IS 2012, July 8-11, 2012, Vilnius, Lithuania. IOS Press, vol. 249, 2013, pp. 249 – 262

[14] Bichevskii YY, Borzov YV. Prioriteti v otladke bolsih programmnih sistem Programmirovanie, 1982, vol. 3, pp. 31-34 (in Russian).

[15] Chengying M., Yansheng L., Jinlong Z. Regression testing for component-based software via built-in test design. In: Proceedings of the ACM symposium on Applied computing, March 11 - 15, 2007, Seoul, Korea, 2007. pp.1416-1421, DOI= 10.1145/1244002.1244307

[16] Draheim D. Business Process Technology: A Unified View on Business Processes, Workflows and Enterprise Applications. Springer Berlin Heidelberg ISBN: 978-3-642-01587-8 (Print) 978-3-642-01588-5 (Online), www.springer.com (2010), DOI= 10.1007/978-3-642-01588-5

[17] Bicevskis J., Cerina-Berzina J., Karnitis G., Lace L., Medvedis I., Nesterovs S. Practitioners View on Domain Specific Business Process Modeling. In: Databases and Information Systems VI. Selected papers from the Ninth International Baltic Conference DB&IS 2010, IOS Press, 2011, pp. 169-182.

[18] Cerina-Berzina J., Bicevskis J., Karnitis G. Information systems development based on visual Domain Specific Language BiLingva. Selected Papers from the 4th IFIP TC 2 Central and East Europe Conference on Software Engineering Techniques, CEE-SET 2009, Krakow, Poland, LNCS 7054 Springer, 2011, pp. 124-135., DOI= 10.1007/978-3-642-28038-2_10

[19] Oditis I., Bicevskis J. Asynchronous Runtime Verification of Business Processes. In Proceedings of the 7th International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), Riga, 2015, pp. 103-108.

[20] Nai-Wei L, Alexander Y. Danger Theory-based Privacy Protection Model for Social Networks In Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, 2014, pp. 1397–1406., DOI= 10.15439/2014f129

[21] Pobereznik L. A method for selecting environments for software compatibility testing In Proceedings of the 2013 Federated Conference on Computer Science and Information Systems pp. 1343–1348

[22] Bluemke I., Kurek M., Małgorzata Purwin M. Tool for Automatic Testing of Web Services In Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1553–1558., DOI=10.15439/2014f93

DOI: 10.15439/2015F60
ACSIS, Vol. 5

# Knowledge Management and Decision Support in Adaptive Case Management Platforms

Łukasz Osuszek

IBM Polska
Email:
lukasz.osuszek@pl.ibm.com

Stanisław Stanek

The General Tadeusz Kosciuszko Military
Academy of Land Forces, Poland
Email: s.stanek@wso.wroc.pl

*Abstract*— **The paper sets out from a central proposition that the concept of adaptive case management (ACM) bears on the evolution of business decision support and knowledge management in modern businesses. While presenting the state of the art in efforts to blend enterprise resources planning/business process management (ERP/BPM) systems with knowledge management systems (KMS) and decision support systems (DSS), the authors observe that the classical platform combining ERP/BPM with KMS and DSS was based on the interaction of three separate layers/subsystems and that, throughout the past decade, that approach proved satisfactory. However, in the last few years it has been increasingly felt that the approach to business process management and enterprise resource planning, as well as to their integration with knowledge management and decision support, needs to be modified. The dynamic and adaptive nature of some business processes poses challenges that the classical BPM approach cannot adequately address. Adaptive case management has been developed to better cope with such challenges. It makes it, on the one hand, easier to align a business to rapidly changing requirements and conditions, and, on the other, it allows organizations to more effectively exploit the potential inherent in organizational knowledge and information resources. The paper discusses the evolution of KMS and DSS from the perspective of their application in ACM environments.**

## I. The Origins of, and Rationale for, Adaptive Case Management

WHAT has long been demanded of information systems is that they move away from the prevalent control flow perspective, commonly adopted in the BPM area, toward the data perspective[1]. Attempts to meet these expectations have led to the emergence of a new class of information systems built around the approach known as adaptive case management.

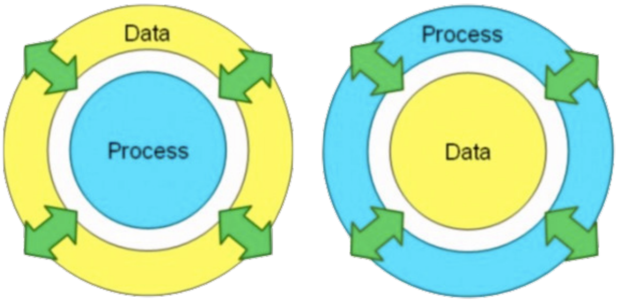


Fig. 1 The classical BPM approach vs. modern ACM.

The term “case” represents a generalization of any activity by so called “knowledge workers”, whose efficiency Peter Drucker sees as a principal challenge of the 21st century [2].

Unlike in classical BPM, under adaptive case management processes are of dynamic character, since they are not defined until at runtime. To master the unpredictability of processes and hence facilitate process management in contexts where processes are mostly complex and where relevant decisions are affected by a large number of factors, more and more organizations choose to switch to adaptive process management systems [3].

ACM allows perfect visibility and full control of each specific case, whether it is handled by a predefined or an ad hoc process, or by a combination of the two.

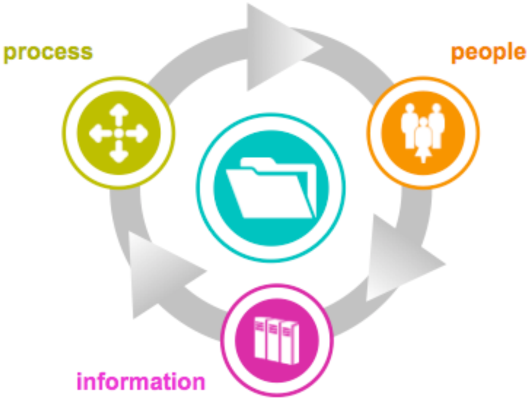


Fig. 2 A typical ACM implementation environment.

This work was not supported by any organization

In a dynamic process management environment, operators/managers, i.e. knowledge workers, can be creative and innovative in performing their work, in that way contributing to organizational knowledge management and creation.

To distinguish the role of process operator from that of temporary process participant, under ACM the former has been redefined and termed as the knowledge worker.
Van der Aalst uses the "blind surgeon" metaphor to illustrate the differences between the two [4]. Under the traditional approach to processes, a participant has a partial view of the whole process, usually limited to the step in the process at which the participant is supposed to make a business decision.

The knowledge worker, on the contrary, has a complete insight into information on the case or process. Knowledge workers constitute a new category of specialized staff whose job is, in the first place, to utilize and exchange knowledge in a productive manner. They are responsible for the generation and implementation of new ideas that enable organizations to align their strategies with the increasingly rapid changes taking place in the business environment; they do so, primarily, through searching, exchanging, combining and utilizing knowledge inside as well as outside the organization.

An enterprise that is run in compliance with the ACM concept will be able to seamlessly combine its core activities with an ability to generate and verify innovations daily [5]. Allowing operators to dynamically modify their processes (and business rules, too), the enterprise management system as a whole opens up to creative initiatives from staff at large, while at the same time preventing chaos that could arise as an outcome of unharnessed changes made to the operating properties. In addition, since it possible to examine the outcomes of changes as they emerge, information on which practices and solutions produce the best results and which yield the worst can be appended to organizational collective knowledge. This stands for actual day-to-day improvement and adaptation of business processes, relying on the best knowledge of a large portion of personnel and getting validated through feedback from customers.

The greatest benefit in deploying dynamic ACM-based business process management is that large enterprises can regain agility and responsiveness that makes them capable of operating and competing in a rapidly changing marketplace. Making it possible to actually delegate work and responsibility to process operators without the risk of losing control of the currently running processes, ACM permits large enterprises to manage their knowledge on an everyday basis through:

- creative, proactive experimenting based on continuous, even if modest, changes, introduced by a number of process operators and leading to gradual accumulation and dissemination of knowledge,
- validation of existing knowledge and elimination of outdated information that no longer meets customers' requirements or competitive challenges.

As a precondition, organizations must be able and ready to adjust their policies and operating properties on an everyday basis as well as to continually update their knowledge on the actual and likely needs of their customers. Adaptive management of business cases is an extension of classical process-based management and an attempt to bring it together with the concept of a learning organization.

One of the cornerstones of ACM is the overarching belief that any organization should continuously expand and process its knowledge on the mechanisms governing its business environment and that this management model is not only more effective but indeed a prerequisite for its ability to keep pace with the unprecedented dynamics of changes in present-day markets and customer expectations. It is often alleged that ACM aims to create a learning organization. Clearly, it streamlines the processes inside an enterprise at several levels, affecting managers as well as personnel.

In course of its business activity, a company creates, accumulates and validates knowledge, which is then used to evaluate and support business decisions. By this token, an ACM system can use business processes in the following ways:

1. As knowledge sources;
2. As a space for organization-wide, innovation-driven knowledge creation and limited experimentation;
3. For knowledge preservation and in database building to bypass the need to set up and operate another system.

## II. The legacy reference architecture

In business practice, the most common enterprise-level system architecture consists in mounting decision support and knowledge management subsystems on top of business process management and enterprise resource planning platforms.

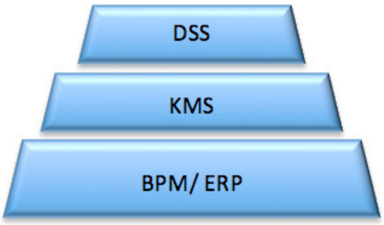


Fig. 3 The classical corporate work platform

Platforms designed in this way would make it possible to use information resources (business rules, workflow mappings, standardized management procedures, etc.) in order to build explicit knowledge and, consequently, develop knowledge management systems and address better support to decision making processes.

This approach involved several problems, the information flow being the fundamental one. Each layer had to be integrated with the others, which required months of analytical and design work focusing on a metadata model enabling the exchange of information between the component layers. For each layer, constituting a subsystem on its own, a specific conception had to be developed for its integration with the other components.

Another feature that users emphasized fairly often was the static mode in which knowledge and decision support were built. Where BPM processes had been defined and designed in advance, the values added by the DSS that the users mentioned the most often were decision process automation (the addition of another layer in the workflow) and instant updates to process flows to keep pace with external regulatory changes. Just like the DSS, beyond the stabilization phase of the underlying BPM/ERP system the knowledge management subsystem did not further contribute to increasing the corporate intellectual potential by expanding its knowledge base.

What users most complained about was that systems so designed were not adaptable enough. Decision support systems perform best in dynamic environments where decisions are complex and the response path cannot be defined in advance. However, in the architectures described above, most decisions are known at a very early stage (the BPM process map pre-defines all options to choose from, leaving no room for dynamic processing and decision making).

The same was, to a large extent, true about organizational knowledge building. In an architecture raised on a static foundation, a knowledge base is built most effectively at initial stages of the system's functioning, i.e. right after it has been put in place, while its subsequent use does not contribute much value to the knowledge base as no new, unique business cases come up.

The paper aims to demonstrate that that the ACM concept can successfully replace the classical approach involving a separate DSS and KMS mounted on top of a static BPM/ERP platform

A. Figure 3 shows architecture composed of three separate layers, which can be considered as the most mature and most tested solution to date.

## III. ACM FROM THE DECISION SUPPORT PERSPECTIVE

Just like the decision making process itself, the decision context and the process context will change dynamically in the course of decision making as the cases are being handled.

From a business perspective, DSS is often regarded as part of ACM. Fischer et al.[6] define the key DSS features as follows:

- developing socio-technical environments that support and empower users to engage in the process of system development not only at design time but also at use time,
- supporting social creativity by providing the technical and social conditions for the exchange of ideas during discussions, debates, brainstorming, co-creation sessions, and other forms of vivid collaboration,
- combining art and design in the processes of self-realization,
- use of meta-analysis for comparing, combining, synthesizing, summarizing, specifying, and generalizing of previous studies.

A popular definition of ACM asserts that it is "a collaborative process of assessment, planning, facilitation and advocacy for options and services to meet an individual's holistic needs through communication and available resources to promote quality cost-effective outcomes" [7]. Under this definition, ACM can be perceived as a platform comprising a decision support system.

Clyde Holsapple, one of the fathers of the DSS concept, commented that "… DSS architecture does not define what DSS is; rather, it functions as an ontology that gives a common language for design, discussion, and evaluation of DSS."[8] This perception of DSS corresponds with Lambert's opinion that "… the architectural design should set a common level of understanding among technical, non-technical and management participants." [9] Suresh Basandra, on other hand, believes that "… system architecture is the process of partitioning a software system into smaller parts." [10] These insights bring us to Holsapple's 2008 definition of the DSS architecture: "DSS architecture is a general framework that identifies essential elements of a DSS and their interrelationships." [11]

Substantial research conducted by the authors at the request of several business enterprises employing ACM systems indicates that most such systems have similar business goals concerning decision support:

- to support the knowledge worker in making optimal decisions in each and every case,
- to deliver faster and more accurate case resolution,
- to improve agility by following business rules in deploying decision support.

Within existing decision support systems, the control function is performed via meta-knowledge subsystems (like norms, axioms, ontologies, etc.). The development of a meta-knowledge subsystem is driven by the double loop pattern of knowledge development. The primary feedback loop, which is characteristic of adaptive learning, involves detection and rectification of deviations from operational norms. The secondary loop, found in the so-called generative learning, is responsible for creative modifications to operational norms.

Likewise, ACM system users will build up corporate knowledge using IT tools and social mechanisms to bring tacit knowledge (i.e. the staff's expertise and individual experience) to broader use in case processing.

It is essential in any DSS project to thoroughly analyze interactions among business owners. As Frederic Adam has it, "… [M]anagers are not seen as atoms but as active, purposeful agents. It is possible to visualize the Decision Making Network (DMN) and to investigate what happens within the networks as the organization tackles a Decision Situation." [12]

ACM could be viewed as an IT platform including an integrated decision support system. The DMN can be then identified with a process map that visualizes all possible case states and provides process managers/leaders with a profound insight into the business.

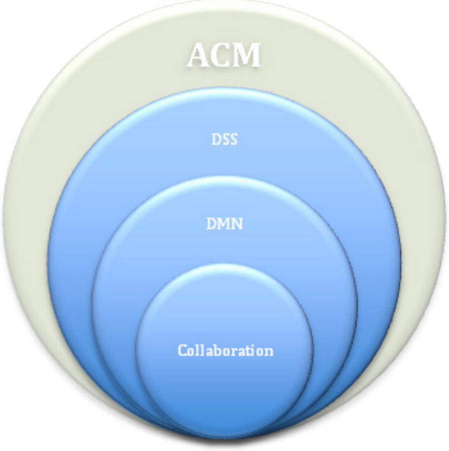


Fig. 4 A model of relationships between ACM and DSS

Those having to cope with less structured decision problems will normally need to have a good understanding of the problem solving process and to be familiar with the applicable techniques. Without this know-how, users situated beyond the operational level might not be able to use the system resources efficiently: even if an expert system is activated to provide them with support in choosing the most suitable tools (models) for their problem, the choice has to be ultimately made by the user. Observation reveals that the most common reason why some systems are not used in tactical or strategic problem solving is not the technology itself but the relatively high demand they put on users' competence (knowledge base).

ACM helps manage the unpredictable by enabling knowledge workers to effectively cooperate and share their knowledge, thus improving the functionality of a decision support system.

Users engaged in solving tactical and strategic problems will rather expect the system to become a "partner in problem solving." Interestingly enough, we have found that the lowest skill levels are associated with the highest expectations from the system, including a proactive attitude in assisting the user. Conversely, the expectations of most advanced and creative problem solvers are limited to being offered an efficient technology and a rich collection of presentation tools.

What is expected from the system in such circumstances is, in the first place, adaptability and expandability through appending new decision models. Not only does the DSS have to offer the requisite decision modeling tools but it also needs to be able to instantly integrate (owing to bidirectional data interchange) with dedicated external systems tackling specific business problems.

The ACM model typically includes a special resource containing knowledge on the business processes utilized in decision making (decision workflows). Identifying the key business processes and analyzing the decision making processes intrinsic to them makes it possible to accumulate knowledge needed to discover and assess relationships between decisions and their outcomes. This appears critical, in the light of our research, for decision analysis at all levels.

Our survey indicates that the most frequently used creative problem solving tools include:

- context-sensitive help along with access to historical data and similar cases,
- group work support tools, such as discussion forums or (widely popular) instant messengers.

## IV. ACM FROM THE KNOWLEDGE MANAGEMENT PERSPECTIVE

A company's body of knowledge is partitioned and distributed among staff and across worker groups; before it can be brought to productive use, it has to be properly organized. Any modern enterprise needs to have a knowledge management system, which can be described as a complex blend of understanding and experience, explicit and tacit knowledge, material and social technology [13]

The following are the principal goals of knowledge management in organizations:

- To make the most of the knowledge that is already available within the organization,
- To create new knowledge, and
- To increase the understanding of knowledge.

It has taken some time for companies and researchers to realize that, besides data as such and besides information that can be interpreted by humans, there exists another vital resource that becomes increasingly crucial to a company's performance but cannot be captured and managed via standard information management methods – and that is knowledge. It can be either explicit knowledge, readily accessible e.g. from an internet portal, or tacit knowledge that resides in the staff's minds and originates in their individual experience, training and talent.

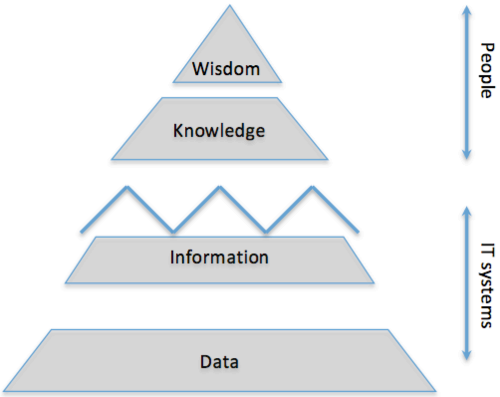


Fig. 5 A model of interrelationships between organizational resources.

Nonetheless, efficient knowledge management alone does not ensure success in business. David Pollard argues that businesses survive and achieve success not because they give a lot of attention to managing their knowledge but because they manage their operations better than their competitors do [14]. The ACM concept, and the instruments it offers, can help organizations optimally utilize their knowledge repositories as a resource and to run their business adaptively, thus increasing their competitive edge.

ACM tools, such as the IBM Case Manager or Pega SBPM, offer knowledge workers a wealth of opportunities to communicate their ideas, observations and suggestions, and to establish best practices, at the same time increasing their autonomy in the workplace.

A survey of industry leaders such as Ford Motor Company, PWC, LLP, Hewlett-Packard or Arthur Andersen, conducted by the American Productivity & Quality Center (APQC), has revealed the presence of processes and features typical of learning organizations [15]. Of these, the following five elements seem to be of critical importance:

Knowledge sharing that occurs in solving problems or with a view to delivering better business results. Different approaches are pursued; in some companies, executive management is involved, while in others preference is given to the use of internet-based knowledge bases or to knowledge sharing among staff during problem resolution.

This aspect of knowledge management has become a central element of the ACM concept. Information on each business case is stored in the system and retrieved by staff engaged in handling cases as they collaborate in case processing and closing. Modern information technologies can provide support for group work through communicators, messengers, social networks, and corporate portals, helping exchange insights, comments, and case processing information – thus facilitating knowledge creation and knowledge sharing.

There is a high degree of awareness of the relationships between knowledge acquisition, knowledge sharing, and achieving business objectives; the staff appreciate the importance of knowledge sharing and understand its contribution to the accomplishment of their company's strategic goals.

It should be noted that, from the ACM perspective, information (e.g. process-related data) is construed as a tool for knowledge creation rather than part of knowledge itself [16]. This is a fundamental difference that has practical consequences for the approach adopted toward knowledge management. Gilbert Probst, Steffen Raub and Kai Romhardt assert that knowledge is a collection of information and capabilities employed by individuals in problem solving [17].

Under ACM, the conception of knowledge management conforms to the definition adopted by the NASA: "Knowledge management is getting the right information to the right people at the right time, and helping people create knowledge and share and act upon information in ways that will measurably improve the performance of an organization and its partners." [18]

The company's principal values are closely related to self-development, cooperation and knowledge sharing; knowledge management is well deliberated and embedded in the organization's business strategy.

At a time of economic downturn, new challenges will arise for business enterprises, forcing them to seek improvements to flexibility, innovation, and responsiveness;

it is also necessary to take a new outlook on human capital – each company's prime asset – with its competencies, autonomy, and responsibility. Adaptive case management systems respond to companies' real needs by offering them dynamic handling of business processes on the one hand, and effective communication and knowledge management mechanisms on the other. In the world of today, success is founded not so much on a careful orchestration of corporate processes, but rather on an organization's ability to instantly adapt, make quick decisions, and use intuition and knowledge management in adjusting its process flows to changing business conditions.

In his 1999 typology, Skyrme proposes a distinction into six types of knowledge:

- Know-**how** – i.e. individual skills and familiarity with procedures,
- Know-**who** – the kind of knowledge that triggers access to required resources; acquaintance with people who can help find the answer or perform a task,
- Know-**what** – or structural knowledge, e.g. acquaintance with pertinent patterns,
- Know-**why** – or "deeper knowing" that enables one to interpret the information one already has and understand the broad context of whatever one does,
- Know-**when** – a sense of the right timing and rhythm for an action,
- Know-**where** – an awareness of the place and context in which an action would be best performed [19].

The latest ACM systems make use of all the above-mentioned types of knowledge. Within such systems, the knowledge of how, what, where and when is readily available and easy to access and modify.

The business owners of respective processes define instructions and types of documents required in process handling, as well as tasks and roles to be performed in solving business cases. Depending on the type of case being processed, relevant properties are displayed to the case manager. The know-how of every assignment (i.e. how it can be best completed within a specified time) is available throughout the system, easy and intuitive to find for each ACM user.

It is different with knowing **who** and **why**. ACM supplies knowledge workers with ample functionalities supporting e.g. task delegation or, notably, the creation of a hierarchy of experts whose assistance is instrumental to resolving a given case. Owing to social networking services (e.g. setting up a team room for experts designated by the system to sort out a specific problem), the knowledge of who can help process a case and why a particular decision should be taken is broadly utilized and dynamically expanded.

The knowledge management style, as well as its structure and language and the extent to which it has been formalized, are aligned with the organization's culture and work environment.

ACM software packages contain rich collections of personalizable tools providing whatever organizational and technical means it takes to raise organizational competence, improve the staff's education and learning capability, and boost up collective intelligence. It supports the development and use of state-of-the-art mechanisms for semantic content analysis and industry-specific glossaries aiding communication among knowledge workers within an organization. Owing to enhanced text analysis techniques, it makes it possible to discover trends, patterns and relationships in unstructured data as well as in related structured data. The resulting observations become part of organizational knowledge and can be used in decision making, forecasting, and setting business targets. In ACM environments such as the IBM Case Manager, the user interface and the system vocabulary are customizable and can be adapted to the language specific to a given professional/business area (e.g. medical or other discipline-specific terminology).

Managers engage in promoting attitudes and behaviors involving cooperation and knowledge sharing.

In organizations that are managed in line with the ACM concept, core activities go hand in hand with a drive for daily innovation. Given the process operators' ability to dynamically adjust and fine-tune their processes, the enterprise management system as a whole becomes open to creative initiatives from most staff while at the same time avoiding the risk of disorder that might result from spontaneous changes to operating norms. In addition, being able to observe the outcomes of changes, the organization can incorporate findings on best and worst practices into its body of collective knowledge. This stands for actual, day-to-day improvements and adjustments to business processes based on knowledge contributed by a large number of staff and on feedback from customers. Instead of hierarchical relations and work division, ACM offers autonomy that enables staff at lower levels to create knowledge and transfer it to higher levels, where new solutions are assessed and approved by senior executives in recognition of their staff's initiative and enthusiasm.

From the perspective of knowledge management, most ACM systems are characterized by several fundamental properties. Depending on specific organizational conditions and requirements, the system's design should embrace socially or technically oriented knowledge management tools. An ACM system framework must aim to integrate all of the company's functional areas as well as all of the existing management subsystems.

Knowledge Management
Collaboration
Adaptive Case Management

Fig. 6 A model of relationships between ACM and KM.

Benefits that can be expected in relation to the introduction of a support system for knowledge management (e.g. an ACM-compliant one) include improvements to organizational knowledge distribution, innovation management, and knowledge transfer (sale). Deloitte & Touche point out another two rewards:

- enhanced internal efficiency – dissemination of best practices, innovative concepts and valuable experiences,
- increased loyalty – establishing and strengthening ties between staff, customers, shareholders, and suppliers [20].

One can easily endorse Ernst & Young's definition of ACM as "a framework or system designed to help companies capture, analyze, apply, and re-use knowledge to make faster, smarter, and better decisions and achieve competitive advantage." [21]

## V. THE NEW ORDER.

The application of an ACM framework substantially increases the capabilities of platforms conforming to the classical architecture delineated in Chapter 2. Adaptive case management effectively combines BPM/ERP components with a dynamic approach to business case management.

In a world where business processes are unpredictable due to their complexity and the number of factors affecting decisions, more and more organizations choose to shift to adaptive process management systems such as adaptive/advanced case management [22] or dynamic case management [23].

The above can be illustrated by a business process involved e.g. in construction projects or loan application processing. Each must be tackled differently, in a manner specific to the type of project or suited to a customer's unique requirements. Since it is impossible to predict and model all types of projects and all customers' likely needs, there has to be a way to dynamically adjust a business process to specific contexts and individual requirements. Process management cannot be therefore reduced to merely reiterating a process according to an established routine, even if the process has been perfectly optimized. Attention to customers' individual tastes and requirements means that processes in an organization must be individualized, too. The advantages exercised by small, innovative and fast-adapting businesses over large international corporations demonstrates that the key to success is no longer in process optimization alone but that it should be sought in an ability to dynamically adjust a process to customers' requirements and address effective support to organizational decision making and knowledge management. Large companies spending enormous amounts of money on introducing new management methods and complex information systems cannot dream of the flexibility that comes as natural to small family businesses.

To overcome this increasingly evident paradox, it is necessary to extend the classical architectural model of a corporate platform described in Chapter 2. Given the need to deliver swift performance demanded by customers and the number of processes running parallel, process owners should not bear the sole responsibility and powers for analyzing and adjusting their processes. With ACM, that responsibility is shifted toward knowledge workers. Within their delegated powers, knowledge workers are allowed to make changes to the processes they are engaged in, just like process leaders would. This functionality, coupled with the integration of decision support and knowledge management subsystems into the ACM architecture, provides a vantage point over companies whose operations are based on the classical BPM/ERP+ DSS + KMS model.

Under the traditional approach, the staff tackling a process are forced to execute an algorithm developed on the basis of a standard process map, i.e. designed in accord with the knowledge workflow that was in place at the time the map was made. As the conditions for process execution will vary in practice (e.g. there do not exist two identical consulting or development projects) and no two processes are in fact performed in the same way, it is advisable that standard processes be adjustable to implementational requirements by their direct executors[24]. The classical process improvement cycle administered by process leaders, involving such subsequent steps as process modeling, performance monitoring, formulating conclusions and, eventually, utilizing the findings to improve the process, is far too slow and therefore inadequate. What is more, in the event that certain clients have conflicting expectations, it might be impossible to design a "universal" process that could be accepted by all the stakeholders.

The ACM approach, incorporating a DSS and a KMS working parallel, is particularly useful in designing processes that are not known in advance, which is what occurs e.g. in court proceedings or in medical treatment. In such processes users have to be given the capability of dynamically modifying and creating tasks within each process. These users are therefore referred to as knowledge workers. Owing to ACM, not only organizational knowledge is built in the KMS, but also the DSS becomes a much more valuable tool, helping make even the most difficult decisions concerning a wide variety of business cases.

A dynamic management strategy allows concentration and intensification of efforts directed at supporting decision processes and organizational knowledge building, while catering to the most unique expectations of clientele.

Under the classical three-layered architecture (BPM/ERP, DSS, KMS), platforms typically suffer from delays arising from the necessity to involve process leaders in making process-related decisions and from the very limited possibility to establish systemic, institutional links between the company's core activity and knowledge management via a process based system. Given the current pace of changes, top executives or process leaders are increasingly unlikely to be able to react timely.

At the heart of ACM is the belief that an organization must never stop expanding and refining the knowledge it already has on the mechanisms governing its business environment, and that management methods founded on this belief are not only more effective but simply constitute a precondition of an organization's responsiveness to the incredibly robust changes in present-day markets and in customers' expectations. Another distinctive feature of ACM, perhaps just as important, is that it can effectively and efficiently support knowledge workers in making business decisions. It is frequently stated that ACM enhances organizational learning to the extent that it represents a step toward the building of a learning organization, since internal process improvement takes place at several levels, affecting both personnel and managers. It should be also underscored that ACM comprises fully-fledged and fully integrated, tailor-made decision support and knowledge management subsystems that do not need to be integrated through individual system integration projects.

The benefits of ACM and the development potential of this methodology have already been recognized by software vendors (e.g. IBM Advanced Case Manager). A major emergent trend in software development is to remove inflexibilities attributable to the posture whereby processes, once identified, are treated as a binding blueprint for action. ACM is different in that it allows users of corporate software applications to take initiative in performing actions that have not been envisaged at identification and design stages. As a result, dynamic ad hoc processes are simpler and easier to model.

ACM systems account for better results in knowledge management and decision support owing to:

- reducing the process description stage to preliminary modeling or to the modeling of known, invariable elements, subject to subsequent supplementation and expansion via the learning mechanism built into the knowledge management system;
- significant acceleration of process improvement by allowing both process operators and process leaders to modify processes, whether in steps or leaps;
- enabling instant dissemination of the knowledge inherent in process innovations through the integrated knowledge management system.

When fully exploiting the potential of ACM-class systems, companies can extract process-related knowledge that has so far been tacit or hidden and in that way find out about processes that cannot be structured by defining straightforward, clear-cut algorithms (process maps). This makes it possible to substantially accelerate the acquisition of knowledge about processes, thereby reducing or eliminating the time required for their formal identification. In effect, organizations should be better prepared to seize emerging opportunities and make optimal decisions, which will affect not only their current performance, but their ability to survive in the market as well. Therefore, it seems that the classical approach is likely to be soon abandoned or modified, and business optimizations will be chiefly conducted using ACM.

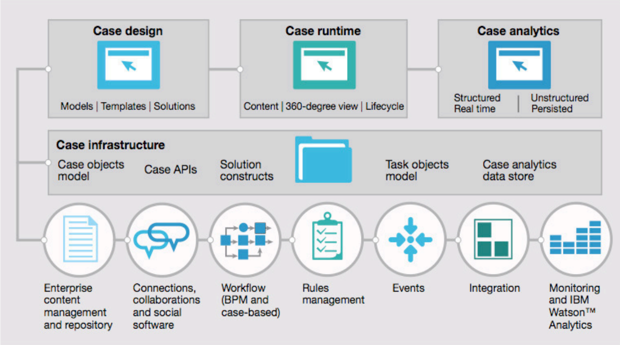

Case design
Models | Templates | Solutions
Case runtime
Content | 360-degree view | Lifecycle
Case analytics
Structured
Real time
Unstructured
Persisted
Case infrastructure
Case objects model
Case APIs
Solution constructs
Task objects model
Case analytics data store
Enterprise content management and repository
Connections, collaborations and social software
Workflow (BPM and case-based)
Rules management
Events
Integration
Monitoring and IBM Watson™ Analytics

Fig. 7 The architecture of the IBM Case Manager.

As shown in Figure 7, elements accounting for group work, knowledge sharing and knowledge discovery (Connections) as well as those responsible for the definition and optimization of decision making rules/procedures (Rules Management) form an integral part of an ACM solution.

## VI. A CASE STUDY

JM Family Enterprises, Inc., founded in 1968, is a diversified automotive company ranked by Forbes as the 27th-largest privately held company in the U.S. The introduction of an advanced case management platform has affected a number of its core businesses, including World Omni Financial Corp. – a major US car finance company.

At World Omni ACM has, in the first place, improved customer service by enabling service associates to access all relevant information concerning a customer's case and hence provide a quicker and more reliable response regardless of where the loan/lease servicing request is initiated. ACM helps the company manage information and knowledge as the customer requests arriving from multiple sources, such as phone, fax, mail or e-mail. When a loan/lease request is received, an associate initiates the process by completing appropriate electronic forms or templates and electronically attaching any associated documents. Such approach structures the case data and improves decision process. Inputting information into electronic forms removes the need to re-key data and, owing to integration, makes it possible to share data with mainframe and client-server systems that are part of the loan/lease servicing process.

At the same time, the elimination of paper-based documents has further streamlined the loan/lease process and resulted in a corresponding "green benefit" meaning that approximately 168,000 pages of paper are no longer printed annually. Moreover, it translates directly into the optimization of knowledge workers' daily duties. 360 degree view for the case information and data is used for decision support in key process steps.

Following is an overview of the key benefits obtained by JM Family in particular business areas through the deployment of an advanced case management system:

- **Customer service**. At its locations in Mobile, Alabama and St. Louis, Missouri, the World Omni call center employs more than 700 customer service associates and processes tens of thousands of automobile loans and leases annually. Originally, the customer service process was complex and involved sizeable volumes of paper documents, frequent interaction with customers, and multiple information sources and systems. Each loan/lease servicing transaction had to pass through a specified sequence of steps engaging many associates and including a number of approval and audit levels. Now all the necessary customer data, along with the requisite knowledge (information regarding a customer's case, transaction history, staff's comments and shared expertise), are available in each single case, providing for an ability to deliver a quick and reliable response/decision regardless of where the loan/lease servicing request is initiated. This is example for using ACM system for better knowledge utilization and making more accurate business decisions.

- **Better punctuality and efficiency** corresponding to shorter response times in handling customer inquiries. The implementation of case management has resulted in improved access to information and knowledge pertaining to a customer case. As a result, knowledge workers can make better-informed and timelier business decisions. Electronic document management and distribution enhances productivity by removing the delays arising from the need to wait for information required to complete a task.

- **Enhancing business outcomes**, which translates into increased customer satisfaction. Working with IBM, JM Family has achieved its primary goals of improving customer service and raising productivity. Customer service associates can now quickly determine the status of a loan or lease service transaction – a task that previously required significant manual effort. A comprehensive view of each customer's case allows associates to update account data and have changes applied consistently across all systems involved in the process. In addition, the ACM system lets knowledge workers collaborate and exchange their knowledge about optimal processing (automatically contributes to organization's knowledge base). As a result of business process automation, loan/lease transactions are now finalized more quickly. JM Family has also achieved its secondary goals relating to: improvements in data quality, paper reduction, better compliance with corporate reporting requirements, and the ability to monitor process performance metrics. JM Family estimates savings at approximately $202,000 annually and anticipates further savings as additional business processes become automated.

- **A better organizational culture**, underpinned by the characteristics of **lean manufacturing**. All of the customer content has been organized into dedicated information silos. This move has been applied consistently across the entire organization, and the idea of a single central backbone for information has led to the emergence of a better, more mature organizational culture. The benefits are reflected in reduced operating costs and increased efficiency that have been accomplished by improving the management and distribution of relevant information and knowledge. Not only has JM Family improved the loan and lease processes, but the company's progressive IT culture has motivated it to strive for continual process improvement with ACM.

- **Market forecasting, costs of decision prediction, business optimization**. Case managers need an insight into workloads and processes to optimize case handling and to be able to find out whether service associates are achieving key performance metrics. To support these objectives, JM Family utilizes the IBM FileNet Process Analyzer to record loan/lease servicing performance metrics. These data are then used by IBM BI software to generate daily performance reports, allowing continual monitoring of the automated loan/lease servicing process as well as of individual customer service associates participating in the process. Such approach is an example of directly contribution to DSS and to business performance overall.

- **Smart, rich data/content analysis**. The ability to dynamically analyze customer content means that it becomes possible to catalog and organize content from multiple sources by exploring and aggregating JM Family information. Content exploration leads to new knowledge and understanding for making instant decisions about business value, relevance, disposition and category schemes in order to take action. "Decommission what's unnecessary" means to cut down on costs and reduce risk by eliminating obsolete, over-retained, duplicate, and irrelevant content, along with the infrastructure that supports it.

**Employee productivity** has been improved by enabling service associates to access all relevant information and enterprise knowledge regarding a customer's case, which accounts for quick and reliable decision regardless of where the loan/lease servicing request is initiated.

**Annual projected savings** of approximately $202,000; faster and more accurate response to customer inquiries; much less paper used to print out documents, which results in significant environmental benefits; improved compliance with corporate records requirements. ROI: 64%, payback: 1.6 years.

**Improved data quality and reduced audit costs.** By eliminating many processes that involved re-keying of data across systems, JM Family has reduced data errors to help ensure consistency and accuracy of data that are transferred between systems. New business processes have been designed to automate World Omni approval procedures and implement firm controls, thereby reducing the amount of auditing required. JM Family estimates annual savings attributable to cost cuts in audit steps at $68,000.

**Savings in case progression and resolution**. JM Family's reputation for technical innovation implied that advanced case management could be a key enabler and driver in improving the quality of customer service as well as the productivity of customer service associates. Additionally, it was recognized that ACM could help the staff improve knowledge management, decision support and other business processes throughout the organization. JM Family has chosen to use IBM software to gain the advanced case management insights that it needed to make the customer experience seamless. The software contributes an integrated process approach that combines electronic forms, business process integration and systems integration to optimize process automation. A flexible framework provides customer service staff with access to all relevant content for each case at every step regardless of source.

**Improved business visibility and transparency** to further enhance compliance. JM Family has designed a process to automatically record the status of key processes as well as of the participants involved in reviews and approvals. JM Family employs the IBM Enterprise Records software to automatically capture and retain loan/lease documents in accordance with corporate compliance policies.

**Continuous improvement with case analytics**. The ACM system implemented at JM Family offers high performance analysis of customer content, hence better monitoring of the automated loan/lease servicing process as well as of individual customer service associates engaged in the process. The continuous improvement policy, coupled with process knowledge retrieved from case analytics, has led to better business decisions resulting in significant savings.

## VII. Conclusions

The paper has tried to highlight the reasons why adaptive case management could be considered as a valuable tool in both decision support and knowledge management. Admittedly, ACM is currently shaping trends in the evolution of both KMS and DSS – by adding vital dynamics to knowledge acquisition and management as well as to decision support.

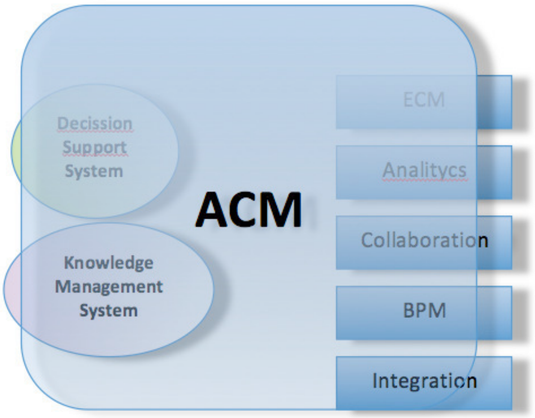


Fig. 8 ACM vis-à-vis organizational resources.

The new business model associated with ACM proves more effective and capable of satisfying most of the requirements of modern businesses.

A dynamic management strategy permits concentration and intensification of efforts aimed at supporting decision processes and organizational knowledge building, standing for an ability to respond to even the most extraordinary customer expectations.

The classical platform, whose architecture is divided into three layers (BPM/ERP, DSS, KMS), is clearly inferior to ACM in terms of efficiency. Furthermore, it offers limited possibilities of establishing systemic, institutional ties between a company's core business activity and knowledge management via the system's process-orientedness. Its adaptability is, on the other hand, increasingly compromised by executive and line managers', as well as process leaders', inability to deliver a timely response to rapid changes in the business environment.

ACM represents the most recent model that comprises integrated subsystems responsible for knowledge management and decisions support. Its very design accounts for superior performance over the traditional approach based on a combination of distinct ERP/BPM, KM and DS systems. Its primacy has been demonstrated by a number of business case studies.

## REFERENCES

[1] Van der Alst, V.M.P, Berens, P.J.S., *Beyond Workflow Management: Product-Driven Case Handling*. In: S. Ellis, T. Rodden, I. Zigurs (Eds), International ACM SIGGROUP Conference on Supporting Group Work (GROUP 2001), New York: ACM Press, pp. 42-51.

[2] Peter F. Drucker, *Management Challenges for the 21st Century*, New York: Harper Collins, 1999, p. 157.

[3] K. D. Swenson (Ed.), *Mastering the Unpredictable: How Adaptive Case Management Will Revolutionize the Way that Knowledge Workers Get Things Done*. Chapter 1. The nature of knowledge work. Keith d.Swenson Tampa: Meghan-Kiffer Press, 2010.

[4] W. M. P. van der Aalst, M. Weske, D. Grünbauer, *Case Handling: A New Paradigm for Business Process Support*, Data & Knowledge Engineering 53 (2005), Elsevier, pp. 129-162.

[5] M. White, "Delivering Case Management with BPM in the Public Sector: Combining Knowledge with Process." In: L. Fischer (Ed.), *2009 BPM and Workflow Handbook: Spotlight on Government*. Future Strategies Inc., 2009.

[6] G.Fischer, E. Giaccardi, "Meta-Design: A Framework for the Future of End-User Development." In: *End User Development – Empowering People to Flexibly Employ Advanced Information and Communication Technology*, H. Lieberman, F. Paterno, V. Wulf (Eds), Dordrecht, The Netherlands: Kluwer Academic Publishers, 2004.

[7] F. Marfleet, R. Barber, S. W. T. Trueman (Eds), *National Standards of Practice for Case Management, 2013 edition*. Melbourne: Case Management Society of Australia, 2013.

[8] C.W. Holsapple, *DSS Architecture and Types*. In: F. Burstein, C.W. Holsapple (Eds), *International Handbooks on Information Systems. Handbook on Decision Support Systems 1 – Basic Themes*. Berlin-Heidelberg: Springer, 2008, pp. 163-190.

[9] B. Lambert, "Data Warehousing Fundamentals: What You Need to Know to Succeed". *Data Management Review*, March 1996.

[10] S. Basandra, *Software Architecture, Data Structures, Algorithms, Programming and Testing Questions and Answers*. California: Basandra Books, 2013.

[11] C.W. Holsapple, *DSS architecture…*, op. cit.

[12] Adam F., "Experimentation with Organisation Analyser, a Tool for the Study of Decision Making Networks in Organisations." In: *Implementing Systems for Supporting Management Decisions*, P. Humphreys , L. Bannon, A. McCosh, P. Migliarese and J-C. Pomerol (Eds), London: Chapman & Hall, 1996, pp. 1-20.

[13] E. Skrzypek, "Wpływ zarządzania wiedzą na wartość firmy." In: E. Urbańczyk (Ed.), *Zarządzanie wartością przedsiębiorstwa w warunkach globalizacji. Wybrane zagadnienia*. Szczecin: Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, 2001.

[14] D. Pollard, *Becoming Knowledge-Powered: Planning the Transformation*. In: Y. Malhotra, *Knowledge Management and Virtual Organizations*. Idea Group Publishing: Hershey, PA, 2000, pp. 196-213.

[15] Z. Kwaśnik, W. Żukow, *Współczesne problemy ekonomiczne jako wyzwanie dla zmieniającej się gospodarki*. Radomska Szkoła Wyższa: Radom, 2010.

[16] I. Nonaka, H. Takeuchi, *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford: Oxford University Press, 1995. Polish edition: *Kreowanie wiedzy w organizacji*, Warszawa: Poltext, 2000, pp. 80-82.

[17] G.J.B. Probst, S. P. Raub, K. Romhardt, *Managing Knowledge: Building Blocks for Success*. London: John Wiley & Sons, 1999. Polish edition: *Zarządzanie wiedzą w organizacji*. Kraków: Oficyna Wydawnicza, 2002, p. 35.

[18] S. Łobejko, *Systemy informacyjne w zarządzaniu wiedzą i innowacją w przedsiębiorstwie*. Warszawa: Oficyna Wydawnicza SGH, 2005. p. 35.

[19] D. Skyrme, *Knowledge Networking: Creating the Collaborative Enterprise*. Oxford: Butterworth-Heinemann, 1999, p. 46.

[20] A. Kowalczyk, B. Nogalski, *Zarządzanie wiedzą. Koncepcja i narzędzia*. Warszawa: Difin, 2007.

[21] R.Blunt, *"Knowledge Management in the New Economy",* Lincoln, Writers Club Press, 2001.

[22] Cf. e.g. www.adaptivecasemanagement.org

[23] C. Moore, C. Le Clair, R. Vitti, *Dynamic Case Management – An Old Idea Catches New Fire*. A report by Forrester Research, December 28, 2009.

[24] M. Szelągowski, *IT jako wsparcie dla procesów*. "CIO Magazyn Dyrektorów IT", 4/2005. Available from www.klubcio.pl/artykuly/313661/IT.jako.wsparcie.dla.procesow.html. Accessed on: October 22, 2014

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1551–1556

DOI: 10.15439/2015F277
ACSIS, Vol. 5

# Levels of the Use of Electronic Communities in the Management of Regions

Cezary Stepniak
Czestochowa University of Technology
ul. Dąbrowskiego 69, 42-201 Częstochowa, Poland
Email: cstep@zim.pcz.pl

Tomasz Turek
Czestochowa University of Technology
ul. Dąbrowskiego 69, 42-201 Częstochowa, Poland
Email: turek@zim.pcz.pl

***Abstract*—The article describes the levels of use of electronic business communities in the management of the city. The study was prepared on the basis of data, observations and investigation at Czestochowa City council, in April and May 2015. The research are related to the development of the concept of Regional Business Spatial Community (RBSC). City Council, along with service providers and providers of media for the residents work for regional development. Five basic levels were identified: informative, tender, negotiative, the level of projects and of processes. In addition to the characteristics of the levels, a qualification of the applied solutions and tools to specific levels has been made. The analysis shows that in the region of Czestochowa, relations at the level I to III occur. The summary indicates directions for further research on the developement of electronic business community.**

## I. Introduction

MORE and more new competencies are assigned to modern local self-government bodies. They affect the operation of more and more areas of local life. To accomplish the tasks assigned to the offices, IT tools are becoming increasingly necessary. One of the effects of the use of IT tools is the ability to create electronic communities.

Appointment of electronic communities in cooperation with the local administration aims to create a network of cooperation between different actors operating in the region. These communities can cover various thematic areas, can be open or closed, can serve to establish cooperation or continuous improvement of it.

This study takes the theme of development of electronic communities to support local development. Creating a regional electronic communities the assumption to facilitate the planning, organization and implementation of various projects. The assumptions of creating these communities is to encourage the increase of the activity of different types of entities functioning in the region [1].

In this consideration of these concepts for electronic development levels of the community were proposed. The arrangement of the shown levels is intended to indicate how they should evolve the built communities [2]. The proposed system of levels is the result of the analysis of several years of cooperation of the authors with Czestochowa City Council for the creation of these communities in order to implement selected urban projects.

## II. Idea of electronic communities

The essence of electronic communities is to join entities, enterprises, institutions and offices in one circle of interest with the use of ICT solutions (information on communication technologies). This idea reminds the use of social networking sites (e.g. Facebook, Twitter, Instagram) by individuals. These people voluntarily create an account in order to communicate and exchange their views. These portals also allow for focusing on specific tasks, objectives and events. Users shall provide one another with information and, this way, form interest groups.

In a similar way, around specific projects and initiatives, enterprises, institutions and offices can be focused. Due to the geographical spread and growing importance of information resources in modern economy and business processes, emerging electronic communities are electronic-virtual. Obviously, they will not use the already existing social networking sites for this. It is necessary to build specific environment - an electronic platform (e.g. Spatial Regional Business Community RBSC - see more [3]).

The basis of the platform of collaboration among electronic communities is an Internet service. Access to the site can be open or closed. This means that only entities that meet certain criteria can use it. Access to the website may also require an invitation by another entity or have the appropriate recommendations. Users of the service can create within its interest circles. These interests may affect ongoing projects, investments, management of the region, cooperation, etc. The implementation of these tasks requires to supply the system with appropriate data and information. The purpose of the scheme is therefore the integration of data from different sources and from many entities. The collected data, according to specific tasks and access rights, would be made available to participants in the electronic community in cloud computing technology.

The basis for the internet system agglomerating online communities can be data from government offices and local government, such as maps, records, information about entities operating in a given area, information about auctions, investments held, etc. On the base constructed this way users would have the option of applying their own information and data (which from the point of view of the community could help to increase the value of the service), from their systems, such as:

978-83-60810-66-8/$25.00©2015, IEEE

- ERP - selected information about the transactions, collaborators, owned resources [4],
- GIS - Spatial information on the nature concerning, for example, distribution of real estate, infrastructure course, scattering branches, commercial premises, the work carried out, etc. [5],
- CRM - selected information about contractors, collaborators, needs in terms of cooperation, missing resources for specific investment or project, etc. [6].

The created website, focusing on electronic communities in a given region becomes a kind of data warehouse. The system integrates data from different sources and in different formats. Then, after establishing rules for access integrated and aggregated data are made available to other participants in the community.

The sharing of data may be controversial and it can raise concerns. Data, information and knowledge are an important part of the organization's resources. They are protected, as this determines competitive advantage. Sharing of information resources within the electronic community therefore requires an appropriate approach and policy. So it is only possible to share such resources, which, in effect, will constitute value added for the electronic community, and, at the same time, can bring tangible economic effects, for example, will allow to gain a business partner, shorten the time of the investment, will help to identify inconsistencies or barriers.

## III. Levels of application of electronic communities

The terms of reference of local government are appointed by civil laws. In them they are contained lists of the areas whose functioning is the responsibility of specific levels of local government. The purpose of these offices is to organize the principles of functioning of the individual, identified areas of life, the creation of entities or cooperation with relevant organizations. In addition, in certain parts of the act, they govern the responsibilities and roles of acceptable administrative units. It is worth remembering that the tasks of the local government bodies are not limited to regulation of specific areas of social life and making administrative decisions, but they can also stimulate local activity, both social and economic. The aim is to stimulate local entrepreneurship in various fields of human activity.

The proper functioning of local government administration should be focused on cooperation with the environment. Electronic media may become one of the main channels of both cooperation and mobilization of potential partners. With it one can gain data about the needs occurring in the region, as well as collect the opinions of residents about the planned projects. It is also possible to keep appropriate information in terms of policy and activities of the office, as well as all types of projects carried out in the area. Electronic media can be an important channel of communication between the office, residents and all stakeholders in the area.

When planning the rational use of IT tools in the activities of the local administration and in the broader regional management thread, the application of criteria should be defined. In these considerations mainly two criteria for the areas of local community life and levels of cooperation were taken into account.

The basic tasks of the governments of cities with county rights in Poland include [7] :

- ensuring spatial order,
- estate management,
- environmental protection and nature conservation and water resources, construction of roads,
- building streets and bridges, traffic organization,
- protection of public transport,
- water supply and sanitation,
- tasks in the field of geodesy, cartography and cadastre,
- education, health and care of the third century people,
- culture (libraries, theatres and other institutions),
- protection of monuments of culture and nature,
- public safety,
- promoting the city,
- social activation of residents,
- creating the city's development strategy.

There are definitely more of the mentioned tasks. However, this citation is to indicate how many different tasks public administrations of cities with district rights have to carry out. To carry out these tasks, the office can create special units for specific tasks or order them to external actors. In some areas, the office can stimulate the construction of infrastructure, which is meant to enhance local entrepreneurship. It happens that in some cases the local government must define a strategy of development of the administrative unit, which may cause stimulation of one sector of the economy at the expense of another (for example, the conflict between cumbersome industry and tourism - especially characteristic for places situated in an attractive tourist area) [8].

In view of the fact that local authorities are elected, making decisions of a strategic nature for the development of the entity must be consulted with representatives of the local community and gather the opinions of residents, take into consideration local needs and ideas for the development of the unit. Very often local authorities are held accountable for actions they have taken for the development of the region and the results they have.

Therefore, their activities must be carried out in different areas and at different levels of cooperation. Analysing the contemporary principles of functioning of local governments, the following levels of cooperation can be indicated (compare to [9]):

1. Information.
2. Tender.
3. Negotiation.
4. Projects.
5. Process.

At the level of information offices collect data about the needs occurring in the region and inform themselves about their own activities. This may involve both information about the activities of local governments (e.g. The information of the topics discussed at sessions of the municipal), as well as planned and the measures taken in the city in various areas of social life. For example, BIP (Public Information Bulletin) can be used for this purpose.

At the level of tender subject matter exchanges of information refers mainly to the planned projects in the city. Offers can flow in two directions. On the one hand, the offer may flow from potential investors who plan to invest in the city and for this purpose are probing interest in the city and its residents planned investment. On the other hand, the local government may announce emerging ideas and plans for different types of initiatives to the residents to hear their opinions. In this case, you can use the systems to collect and transmit documents electronically, as well as the possibility of collecting public opinion through electronic methods in various vital issues residents.

The level of negotiation is related to the organization of projects implemented in the city. The role of the office may be threefold. The authorities may initiate the project (e.g. construction of the road). Using IT tools may carry out auctions, and then negotiate, it may also organize consortia made up of representatives of many entities, which in turn has to lead to take the project. The second type of the role of the office is to support people's initiatives to be implemented throughout the city. In this case, the city can be a shareholder or, in a variety of ways, support the project implemented, including seeking partners for cooperation. What it can also be accomplished by using IT tools. The third role is to control, which is on one side of verifying all documentation relating to the operation, on the other hand may consist of facilitating and promoting investments.

The level of projects is to create tools which will support the implementation of specific projects. This is particularly true for large projects that will require the cooperation of many entities. In this case, IT tools can be used to establish a communication medium between all the partners involved in the project.

Referring to the contemporary concepts of process management, many projects can be applied to specific procedures for implementation. Using the tools for process design, it is possible to design informatics support for processes carried out in the framework of the project [10]. This can apply both to investment processes and then to the control or management of entities arising as a result of the investments the city [11]. In this way, not only a tool for electronic document flow can be created, but also social tools can be used to create the development of the project's areas.

The above levels of cooperation are the basis for the creation of electronic communities. The thing is that created communities are not homogenous. Hence, it is unlikely to create a single community. As a part of the regional communities, may occur many community sectors (e.g. related to education, health services and municipal investments). Furthermore, many traders may belong to different communities. These communities may be closed and open. In addition, different rules may be accompanied by the possibility of joining the framework of the community. This may depend both on the specifics of the industry, where a community was established, as well as the level of cooperation.

Construction of electronic communities is a long process and can take many forms, which over time can be converted together with their development. The formation of regional electronic community usually takes place according to certain phases. A significant impact on its shape has also the purpose for which they arise. For the purposes of regional governance can be invoked various communities, which seeks different objectives and designed to satisfy the diverse needs of society.

Bearing in mind the diversity of the community and the objectives of their use in regional management in these deliberations levels of use of electronic community in accordance were adopted with accepted earlier levels of cooperation.

Information level is a kind of substitute for the community. Its aim is the presence of different information or initiatives. Information may involve various important issues in the life of the city. In contrast, initiatives on one hand inform about emerging initiatives, on the other objective is to obtain an opinion, and, above all, support for relevant projects. Generally, the information appears at the integrator (e.g. the town hall), which aims to send a message to the widest possible population. Usually, messages are transmitted to open community, though not excluded is the creation of news compasses closed. Information level may be interdisciplinary.

The level of bidding has been generally assumed by discipline. It refers to specific sectors of social life. It also requires interaction between the initiators of projects, and those who will give an opinion, in the long run, may take part in them. It is worth remembering that not always the initiators are entities that organize the project and in the future will be integrators. The level of bidding electronic community is usually open. Although they may also receive the offer (in particular of a business), which can be routed to a closed community.

The level of negotiating refers to specific planned projects. An electronic community will have to negotiate in order to organize the project from preliminary discussions by submitting comments and offers, and ending with the completion of electronic auctions and having conversations to the conclusion of a relevant investment agreements have in the future supplies. Generally, the community may be mentioned at the outset generally open, but with the implementation of further preparatory activities for the implementation of the project can be closed. It is also necessary to note that many similar communities can be invoked, each for a different project.

With the closure of community at the negotiating level, exclusively to the entities involved in the implementation of the project, it can be assumed that the community moves to the level of the project. At this level it is important to provide adequate communication to the smooth implementation of the project and its management. At this level you will be required a certain level of integration of information systems managed by participants in the community [12]. These communities typically will be closed. Also in this case there may be many different communities that support various projects.

The level of the process is a further development level of cooperation. It is based on the fact that the project is quite strictly formalized. First of all, procedures for implementation have been strictly defined. Thanks to it, it is possible to control the implementation of all processes within the project [13]. The condition for achieving this level is the right level of integration of information systems of all participants in the community, the availability of tools for modelling processes and their application. Communities are closed. Time may be limited for the time of investment or for the duration of the operation of the project. Comprehensive description of the various levels of use of electronic community are presented in Table. 1.

## IV. Technology used within electronic communities

Implementation of the indicated levels of cooperation within the electronic community requires a set of ICT tools. These tools require proper integration so as to form one coherent system. The functioning of such a system must be network service (Web Service) in database technology.

The simplest in the implementation seems to be the level of information. Technological solutions that can be used is news in the form of blogs, forums and other Web 2.0 tools. They allow the posting of information on existing activities, investments, tenders, etc. Other participants of the community can give opinions and comment on them. Even at this level of the system (such as on social networking sites) should enable the creation of groups and circles of interests around specific themes. Similar approaches may be used as a proposal. However, it is necessary to supplement the system with the possibility of submission of tenders in an automatic manner, e.g. Through a system of forms and attachments. Sent offers should be reliable, which is why the system can be completed with the need for registration, and in some cases it may be required to sign an electronic certificate.

Implementation of negotiation level builds the idea of groups and circles of interest. In the framework of specific initiatives, circles of interests are filled with participants. They become cooperators and collaboration platform offers the possibility of negotiations in the virtual space. The necessary technological solutions in this area appear to be advanced communication tools, video conferencing, workflow systems and groupware systems (e.g.. Microsoft SharePoint or Novell Open Workgroup Suite).

Implementation of joint projects within the electronic community web platform needs to be supplemented with tools to coordinate and assist in the definition of business factors, making the right decisions about funding and display the status of projects and resources across the enterprise. This type of software helps manage projects and allows collaboration from anywhere using tools for project managers, project teams and decision-makers. An example of such a platform are applications such as Microsoft Project version running in cloud computing technology - MS Project Online.

Support for business process level requires implementation of solutions allowing for the design, implementation and development processes in the context of groups and circles of interest. Applications supporting Business Process Management BPM can also be handled in the cloud. An example of such a solution may be IBM software. As declared by the manufacturer it allows for:

- optimizing operations, providing them with an excellent view of the ongoing work and ongoing tasks based on continuous monitoring and analysis of processes;
- faster execution of tasks through comprehensive collaboration tools;
- secure management of changes and taking them under intuitive supervision;
- offering customers valuable forms of interaction with the inclusion of mobile environments to business processes;
- continuous analysing the business through the integration of business processes with the basic enterprise systems.

Diagram of the functioning of the web service for electronic communities is presented in Figure 1.

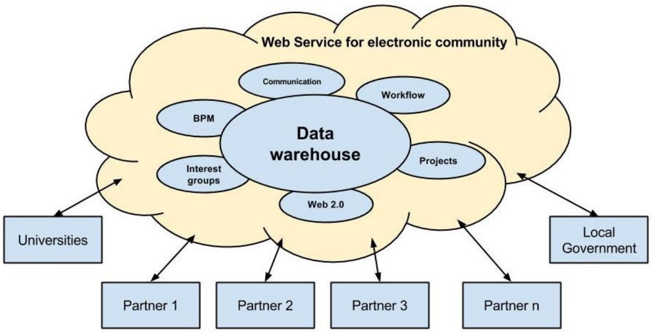


Fig. 1 Scheme of Web Service for electronic community
Source: Own study

Table I.
Characteristics of Cooperation Level in Electronic Communities

| Level | Availability | Participants | IT Tools | Range |
|---|---|---|---|---|
| 1. Information | Open, unidirectional | All actors in the region | Web 2.0, Web Services | Regional |
| 2. Tender | Open, two-way | Entities interested in a particular industry | Web 2.0, Web Services, e-forms, | Industry |
| 3. Negotiative | Open/closed | Entities interested in the project | Web 2.0, Web Services,, MS SharePoint, etc | Interested in the project |
| 4. Projects | Closed | Entities carrying out the project | ERP, GIS, CAD/CAM, CRM, – MS Project Online | Stakeholders of the project |
| 5. Process | Closed, formalized | Entities carrying out the project | ERP, GIS, CAD/CAM, CRM BPMN or BPEL | Stakeholders of the project |

An important element of web service is a data warehouse. Principles of the data warehouse is presented in Figure 2.

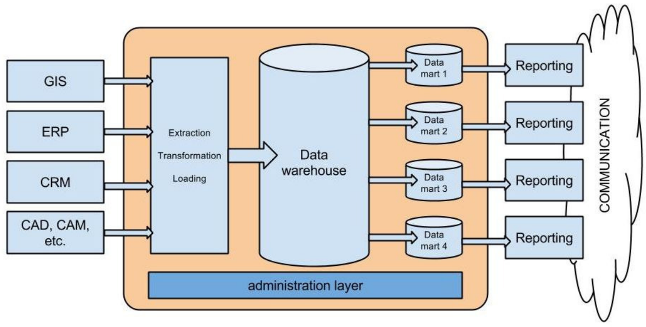


Fig. 2 Data warehouse model for electronic community
Source: Own study

## V. Current status of the functioning of electronic community

In the first quarter of 2015 year, a number of discussions and consultations with the Office of Czestochowa, which concerned the current state of the electronic community were conducted. The aim of the study was also to identify potential technological and organizational solutions that can contribute to the development of this type of concept.

The Office, as the body responsible, and strongly interested in the development of the city and the region could potentially become an integrator and leader of this kind of technology. The more that has a large number of information resources, which could become the basis for the initiation of the said web service and data warehouse.

Most of the solutions operating at the moment in Czestochowa's offices, and which can be regarded as the nucleus of the electronic community function on the first level - information. That is, they operate on the principle of communicating information. Offices, institutions and firms have the ability to publish information about their activities, planned investments or needs. Some data is published in the activation of specific social groups. Web 2.0 technologies allow for interaction, feedback.

The most important solution in this aspect is the website of the City Council www.czestochowa.pl It has a thematic areas allowing you to create electronic germs communities in sections: economy, education, culture, sport and tourism,

Web service municipal office is also integrated with other thematic parties, e.g.:

- Public Information Bulletin (bip.czestochowa.pl),
- The electronic public administration services platform – (epuap.gov.pl),
- System of Electronic Communication in Public Administration (sekap.pl).

City Czestochowa also has solutions aimed at electronic integration of citizens through a system of public consultation, support local initiatives - konsltacje.czestochowa.pl

Due to the advantages of tourism and pilgrimage, Czestochowa has a developed system of City Tourist Information Service - info.czestochowa.pl.

Part of systems, in particular those which relate to areas of the economy and investment, has the characteristics of a second-level electronic community - level proposals. On the biznes.czestochowa.pl is a search system of investment offers. (Business and Real Estate Marketplace). It consists of a main areas: maps listings (map deals), offers base, make an offer and ask for offer.

This system could be a first step for potential investors looking for potential areas for investment location.

A similar function can be performed by geo investor (http://e.czestochowa.pl/geoportal-inwestora). Information about the city, the existing administrative procedures, planning and fiscal policy, as well as investment offers in Czestochowa can be found there.

Part of the systems managed by the Office of Czestochowa has elements of Level 3 - negotiation. In particular, it can be seen in GIS, whose task is to record geodetic resources of the city. These systems are at the moment there are models that allow for automation and virtualization negotiation processes, but in the case of the occurrence of such needs, there is the possibility of their development and software.

The study of information systems and web solutions managed by the City and discussions with officials revealed that at the moment no system supports electronic community at the level of implementation of projects and processes. This idea, however, aroused interest. Employees magistrate indicated their willingness to participate in the potential projects in this area.

## VI. Directions of development and conclusion

According to research conducted in the City Hall of Czestochowa it is known that a part of the management office is interested in creating electronic communities, which would facilitate contacts with potential petitioners, investors, stakeholders cooperating and various other organizations and residents. The electronic media allow not only the flow of documentation (eg. taking samples of the documents and their electronic submission), but also allow making contact with various actors. Already, employees of the relevant departments of the Municipal Office in Czestochowa are in contact with the representatives of various organizations, invite to participate in various training meetings and other events. Under mailing lists hides a kind of network of cooperation, especially when many of these entities interactively respond to sent invitations.

It seems that the problem lies in building a tradition of cooperation in the use of IT tools. It also seems that the participants of these forms of cooperation could take a more active part themselves by inspiring their own events, and invite representatives of other entities, including the town hall. In this way, it could be possible to create more circles of the electronic community.

In continued discussion it was noted that communities may be based not only on electronic communication, but also on data exchange. This data is usually registered in closed or only partially open systems of real or potential participants in the proposed community. In principle, from negotiating the level of communication between participants of the community can be enriched by the exchange of data which may facilitate subsequent cooperation.

At the level of projects, in particular data exchange processes between systems of the users already seems to be rather essential. However, achieving this requires the fulfilment of several conditions. The most important ones include:

- the existence of a genuine desire for cooperation, including the exchange of knowledge and resources between members of the community,
- building ventures involving many actors, where IT tools will become an essential part of management,
- supporting process management in organizations and shaping attitudes among managers of community participants, as well as rank and file workers conducive to the implementation of formalized business procedures
- awareness among others on the electronic community, proposed and submitted socio-economic initiatives, business process modelling tools.
- creating rational data collection systems to support the planning and implementation of regional projects.
- training of specialists in the construction of models describing the processes and selected regional issues, the use of which will require the use of data already collected.

These requirements have been defined basing on research conducted at the office of the city of Czestochowa and on the basis of contacts with companies supplying utilities (eg. plumbing companies, a provider of electricity and gas, etc.)together with other civil society organizations operating in the city of Czestochowa.

By analyzing the obtained results it can be stated that the office fulfils the obligations assigned to it by law. However, there is a will to activate the community of the town, which until then struggling among others with problems such as depopulation and a relatively high unemployment rate in the whole province of Silesia. Therefore, it seems that the establishment of regional electronic community can be one of the factors increasing the activity of residents and reverse adverse trends in development.

## References

[1] D. Feehan, M.D. Feit, “Making Business Districts Work: Leadership and Management of Downtown, Main Street, Business District, and Community Development Organizations”, Published by Routledge, 2006.

[2] M.R. Parks, Social Network Sites as Virtual Communities. In *A Networked Self: Identity, Community, and Culture on Social Network Sites*, edited by Z. Papacharissi, Routledge, New York 2011,.

[3] D. Jelonek, C. Stepniak, T. Turek, The Concept of Building Regional Business Spatial Community, ICETE 2013. 10th International Joint Conference on e-Business and Telecommunications. Proceedings. 29-31 July, Reyklavik, Iceland, SCITEPRESS, p. 83-90, 2013.

[4] J. Wieczorkowski, P. Polak., “Analysis and Implementation Phases in the Two Segmental Model of Information Systems Lifecycle, in 2012 Proceedings of the Federated Conference on Computer Science and Information Systems FedCIS, PTI, IEEE, Wrocław 2012, pp. 1041 - 1046.

[5] F. Harvey, A primer of GIS. Fundamental Geographic and Cartographic Concepts. The Guilford Press. New York London, 2008.

[6] J. Dyche, The CRM Handbook: A Business Guide to Customer Relationship Management. Published by Addison-Wesley, 2002.

[7] Ustawa z dnia 5 czerwca 1998 r. o samorządzie powiatowym. Dz. U. 1998 nr 91 poz. 578 (*The Act of 5th of June 1998. on county government. Acts. Laws 1998 No. 91 item. 578*).

[8] C. Stępniak, “Metodologiczne i organizacyjne aspekty budowy strategii gmin” (*Methodological and organizational aspects of the construction of municipal policy*). In: „*Zarządzanie publiczne. Uwarunkowania - kierunki – techniki*”. Monografia pod red. Arnolda Pabiana. Wyd. Wydziału Zarządzania Politechniki Częstochowskiej. Częstochowa 2010.

[9] E. Ziemba T. Papaj J. Będkowski, Egzemplifikacja e-government w Polsce – analiza porównawcza SEKAP i ePUAP (*Exemplification of e-government in Poland - comparative analysis of SEKAP and ePUAP*), Roczniki Kolegium Analiz Ekonomicznych nr 29/2013, p. 430.

[10] Freund, B.R. Rucker, “Real-Life BPMN: Using BPMN 2.0 to Analyze, Improve, and Automate Processes in Your Company”. Published by Createspace 2013.

[11] J. Wieczorkowski, “Narzędzia modelowania procesów biznesowych w aspekcie wytwarzania i wdrażania systemów informatycznych”. (*Business process modeling tools in terms of production and implementation of information systems*). In: “*Gospodarka elektroniczna - wyzwania rozwojowe*” t.1, red. Jacek Buko, Henryk Babis, Roman Czaplewski, Zeszyty Naukowe nr 702 Ekonomiczne Problemy Usług nr 87, Uniwersytet Szczeciński, Szczecin 2012, pp.522-531.

[12] C. Stepniak, T. Turek, Integration of Spatial Information Resources on the Example of Utility Companies in Częstochowa Region, Online Journal of Applied Knowledge Management, Vol 2, pp. 97-108, 2014.

[13] I. Weber, Semantic Methods fo Execution-level Business Process Modeling. Modeling Support Through Process Verification and Service Composition. Ed. By Springer.-Verlag Berlin-Heidelberg 2009.

DOI: 10.15439/2015F68
ACSIS, Vol. 5

# Moral Hazard in IT Project Completion. A MultipleStudy Analysis Case

Bartosz Wachnik
Warsaw University of Technology,
Faculty of Production Engineering
Institute of Organization of
Production Systems
ul. Narbutta 85, 02-524 Warszawa,
Poland
Email: bartek@wachnik.eu

***Abstract*—Implementing management support information systems with the use of outsourcing is the prevalent method of completing this type of project in Poland. Agency theory is one of the significant categories of theories used in the analysis of IT outsourcing. Literature studies indicate a research gap concerning the phenomenon of moral hazard in IT projects consisting in the implementation of management support information systems. The scope of this article is to present research results on the phenomenon of moral hazard based on the case study method. The research results may be interesting for theoreticians of business informatics and for practitioners completing IT projects both in enterprises and government agencies.**

## I. Introduction

According to J. Lee [1], IT outsourcing means managing a company's IT infrastructure through administration mechanisms exercised in cooperation with external organizations. A particularly interesting area of IT outsourcing is the purchase of implementation services as part of the completion of IT projects consisting in management information system implementation. The author's research has shown that IT outsourcing linked directly to the implementation of management support information systems, i.e. ERP, BI, CRM and DMS, was used in 78% of Polish enterprises, while the remaining 22% completed this type of project themselves.[1] Consequently, the issue of outsourcing in this type of project is clearly visible in the majority of Polish companies and it requires further research by theoreticians of business informatics in Poland. Internationally, issues linked to IT outsourcing have been the subject of research by business informatics theoreticians for many years. J. Dibbern [2] analyzed 84 articles on IT outsourcing published between 1992 and 2000. 10 categories of theories used to analyze the question of IT outsourcing were identified, which were then divided into three groups: strategies, economy and social-organizational groups. Agency theory was ranked third in terms of the number of publications,[2] which shows that it is frequently used in research on IT outsourcing. Agency theory stems from the need to explain the behavior of participants in relations client-contractor, where both parties have different goals. The risk is known as agency problem, relations – as contract, and the parties of the relation – as principal and agent. Agency theory addresses the phenomenon of dependency, where the principal delegates tasks to the agent. Agency problem is considered in two aspects [3]. First, it refers to the contradictory goals of both parties, which result in different expectations dependent on classifying the risk priority differently – the so-called risk preferences, linked to the agent's activities carried out on the principals' orders. Both sides of the relation try to minimize the risk. Different goals may result from opportunism, where both sides enforce the realization of their own goals over common goals. The second problem refers to the mechanism which influences specific behaviors of the agent, according to the principal's expectations [4]. The unit which constitutes the basis of theory analysis is the contract defining the rules of cooperation between the principal and the agent. Important concepts introduced by the theory are: moral hazard, adverse selection and programmability. According to the research carried out by Dembe and Boden [5], the term "moral hazard" dates back to the 18th century and was widely used by British insurance companies until the end of the 19th century. The early usage of the term has negative connotations, linked to fraud or other immoral behavior – usually on the insuree's part. Renewed scientific interest in the phenomenon of moral hazard appeared amongst economists in the 1960s, and it was not applied to instances of fraud or immoral behavior. The term would be used to describe ineffectiveness, which could

[1] The research was carried out between 2012 and 2014 amongst 300 companies based in Mazovia, Greater and Lesser Poland, as well as Upper and Lower Silesia, which completed 370 projects consisting in management support information system implementation. The selected enterprises met the following criteria: number of employees between 80 and 1000, own IT department, minimal income: 40mln PLN.

[2] The ranking of theories applied to research of IT outsourcing with the use of 84 articles conducted by J. Dibbern, T. Goles, R. Hirscheim and B. Jayatilaka [2]. Theory of transaction costs—16, Strategic management theory—14, Agency theory—10, Group of resource allocation theories—9, Group of social exchange theories—7, Game theory—4, Theory of power group—2, Diffusion of innovations theory—2, Other theories, e.g. of knowledge management, risk management, psychological contract —13.

978-83-60810-66-8/$25.00©2015, IEEE

appear in case of bad risk management, rather than in the context of ethics.

We need to stress here that the agency theory, alongside institutionalism, the theory of contracts and the theory of transaction costs, constitutes a component of the new institutional economy [6], which is based on the idea of institution understood in three categories: *A*. social, defining the norms of human relations, *B*. legal, regulating the creation, duration and termination of legal relationship between legal entities and *C*. organizational, regulating the functioning of formal organizations, safeguarding certain norms.

J. M. Buchanan [7] confirmed that a new contractarian paradigm of a new institutional economy is forming, which he described by saying that "economics comes closer to being a 'science of contract' than a 'science of choice' [...] The maximizer must be replaced by the arbitrator, the outsider who tries to work out compromises among conflicting claims." This statement, as well as the concept of treating an enterprise as a cluster of contracts and making the project's success dependent upon the quality of contracts, fully confirm the rationale of a new institutional economy as a central element creating a model of IT services in the form of outsourcing.

The scope of the article is to present research results concerning the incidence of moral hazard in IT project completion as part of outsourcing. In the article, I focus on projects consisting in the implementation of management support information systems, i.e. ERP, BI, CRM and DMS. The article stems from an attempt to identify the conditions influencing the effectiveness of completing a given group of IT projects based on outsourcing. In my opinion, the factors influencing the effectiveness of IT project completion have been changing over the years and their character has become increasingly nuanced. This results from many factors, i.e. the rapidly changing technology, the evolution of project completion methods, the fast-growing saturation of IT system markets, the hypercompetition amongst suppliers, the behaviors and practices of suppliers during the process of sale or project completion. This is why I believe that researching the problem of outsourcing in IT projects is important, where the results will be beneficial both for theoreticians of business informatics and practitioners working in companies and government agendas. The article belongs to a cycle of over a dozen articles that I wrote to present the results of my research on IT project completion as part of outsourcing.

## II. Moral hazard in IT project completion

According to Y. Lichtenstein [8], agency theory describes the methods of completing IT projects during the whole life cycle of a project, i.e. as part of the following three stages:

*Stage 1*—Organization and completion of a tender aimed at selecting a management support information system and a company to implement it.

*Stage 2*—IT project completion.

*Stage 3*—Information system operation on the basis of an SLA contract.

Agency theory presents the contract for the service of implementation of a management support information system between the supplier and the customer as a relation between an agent and a principal. During the first stage of a project's life, an outsourcing agreement for the provision of IT system implementation services is established, and later on a Service Level Agreement is negotiated. Y. Lichtenstein [8] indicates that in case of negotiating IT project agreements, and later their implementation, an incompatibility of both parties' interests occurs. The client's goal and interest is the completion of an IT project within the planned time frame and budget, considering the Total Cost of Ownership (TCO) and meeting all the organizational and technological requirements designed for the project. The supplier's goal and interest lies in achieving a planned income in the whole life cycle of the project, meeting the planned quality of completed work. Table I (Source: Own Study) presents the contradictions between the client's and supplier's interest in case of completing an IT project consisting in the implementation of a management support information system.

Y. Lichtenstein [8] points out that a high level of information asymmetry between the supplier-agent and the client-principal means that at every stage of the project's life cycle, the supplier-agent may be prone to the abuse of trust: moral hazard, which results directly from the high level of information asymmetry concerning both the ERP, CRM, DMS and BI-class software as such, and the method of implementing the project. Especially in the case of projects completed on the basis of a fixed budget, the supplier-agent may have a strong motivation to push the cost below the planned budget, which can have an influence on the quality of services.[3]

[3] The research was carried out in 2013. It covered 500 enterprises where 895 IT projects consisting in the implementation of ERP, CRM, BI, DMS, BI and E-learning-class management support information systems were completed. The chosen enterprises were based in Mazovia and Lesser Poland and the research was conducted among enterprises employing less than 400 people. Research results showed that in the management support information system market in Poland, the structure of agreement types is as follows: 68%—fixed price contract, 27%—time and material contract, 5%—cost-reimbursable contracts.

TABLE I.
GOALS AND INTERESTS OF BOTH TRANSACTION PARTIES

| Client's goals and interests | Supplier's goals and interests |
|---|---|
| Fulfilling business goals<br>Fulfilling technological goals<br>Completing the project within the time frame<br>Completing the project within the maximum planned budget<br>Minimizing the total cost of software ownership (TCO)<br>Achieving functional and technological solutions creating a temporary competitive edge | Completing the project according to the planned quality<br>Maximizing the project's profitability in its entire life cycle, i.e. implementation and usage<br>Obtaining reference information after the completion of the project, which will allow the supplier to sell services to other projects |

The supplier-agent possesses all the information about their current activities and all their plans linked to project completion, while the client-principal has only a limited knowledge in this area. Knowledge concerning the completion of IT projects consisting in the implementation of management support information systems is increasingly robust and dynamically changing, which means that the client-principal is not able to effectively negotiate a favorable contract, and then supervise this type of project on their own. Due to the fact that both parties' interests are not equal, the supplier-agent may be prompted to take inappropriate action, i.e. abuse trust (moral hazard), in a situation where they are not controlled because the knowledge and experience of the client-principal limit the possibilities of effective control. The potential action of the supplier-agent aimed at securing the realization of their particular interests may result from their opportunism. Inappropriate behavior of the supplier-agent manifesting itself in trust abuse in the second stage, which is completed on the basis of a fixed budget, and in the third stage of a project's life cycle may lead to:

*A.* Guaranteeing full project profitability for the supplier-agent when it turns out that, already after signing the implementation agreement, the scope of the project is larger than originally foreseen at the bidding stage. This situation may occur when the client-principal was only guided by the price in their choices, while the competition amongst the suppliers was high. In the current situation of hypercompetition on the market, it may happen that this type of projects may be offered at a lowered price. In this situation, paradoxically, it is the client-principal who should make sure that the project price is not too low, with the assumption that they have given the supplier-agent all the requirements for the implemented IT system.

*B.* Increasing project profitability for the supplier-agent compared to the earlier plan of completing the project at a lower cost.

The result of the supplier-agent's trust abuse in the project's whole life cycle may be the occurrence of one, or a combination of several organizational complications impacting the effectiveness of an IT project, i.e.:

1. Lack of complete or partial fulfillment of project goals.

2. Unjustified exceeding of the budget in the completion of the second stage, as well as the third stage, the so-called Total Cost of Ownership.

3. Unjustified exceeding of the planned time frame for the completion of the second stage.

An important organizational aspect of the contract between the supplier-agent and the client-principal is an attempt to construct a mechanism which would effectively eliminate the possibility of trust abuse in the second and third stage of the project life cycle. Designing such a mechanism may turn out to be costly and difficult to complete [4]. My research results will allow us to indicate and describe selected attempts to abuse trust by the supplier-agent in IT projects consisting in the implementation of management support information projects, i.e. ERP, CRM, BI, and DMS.

## III. RESEARCH METHODOLOGY

In my research, I have used the multiple case study method. Four enterprises which implemented and are currently using management support information systems, i.e. ERP, CRM and DMS, constituted the subject of research. The goal of the case study is developing the agency theory, and especially a better understanding of the notion of moral hazard in management support information systems.

I analyze the case study, as it allows us to develop the existing theory and provide explanations of phenomena unrecognized before, such as moral hazard in IT projects. I focus on the client's perspective during the whole life cycle of an IT project in an enterprise, i.e. from the bidding stage to the operation of management support information systems. My choice of research method – case study, is motivated chiefly by two circumstances [9]:

1. The early stage of knowledge development in the given research area, i.e. agency theory in a specific group of IT projects.

2. Lack of recognition of moral hazard in real conditions.

As part of the multiple case study analysis, I would like to pose the following research question:

In what behaviors does moral hazard in the relation between the principal (client) and the agent (supplier) manifest itself in case of IT projects consisting in the implementation of management support information systems based on outsourcing?

The choice of studied cases was carried out through purposive sampling. According to B. Flyvbjerg [10], there are five main criteria of case selection. Table II (Source: [10]) presents the criteria along with their characteristics in the context of conducted research.

TABLE II.
FIVE MAIN CRITERIA OF CASE SELECTION

| **Criterium** | **Information on the fulfillment of the criteria** |
|---|---|
| Data availability | Guaranteed |
| Distinctiveness of the case, clearly illustrating studied patterns | Projects that ended in partial failure, but not interrupted during the implementation |
| Variation in the analyzed cases | Variation in the analyzed cases is expressed in the selection of:<br>- IT projects consisting in the implementation of four management support information systems, i.e. ERP, CRM, DMS<br>- Client profile<br>- The results of project implementation |
| Critical character of the phenomenon allowing to formulate a general statement | The incidence of moral hazard between the agent (supplier) and the principal (client) during the whole life cycle of project implementation influences the results of project implementation from the client's perspective. |
| Metaphor allowing to point the researcher's attention towards a specific course of the studied phenomenon | Aiming to analyze the phenomenon of moral hazard in the entire project life cycle, I selected cases that could be studied at the stages of: bidding, contract negotiations, implementation and information system operation. |

## RESEARCH RESULTS AND THEIR INTERPRETATION

Table III (Source: Own Study) presents the results of multiple case study research.

I analyzed the implementation of ERP, CRM, BI and DMS-class management support information systems completed as part of a contract based on a fixed budget. All the analyzed implementation projects ended in partial failure, although their implementation was not interrupted. I diagnosed three major types of supplier's behavior, through which moral hazard, resulting from the opportunism[4] of the supplier-agent and a high level of information asymmetry between the client (principal) and the supplier (agent) in IT projects, manifested itself.

*Behavior 1.* An attempt to minimize the supplier's cost of completing stage 2 and 3 of the project by:

- Engaging specialists with low competences and little experience, which is linked to their low salary.
- Lowering the actual workload of consultants working on specific tasks in the project in comparison to the workload declared in the agreement regulating the outsourcing of implementation services.

*Behavior 2.* An attempt to sell implementation services concerning stage 2 at a lowered price, which will allow them to win the tender in the conditions of hypercompetition in the market. The supplier's intention is to compensate for the possible losses in stage 2 with profits in stage 3.

*Behavior 3.* An attempt to intentionally complete chosen types of project tasks in stage 2 poorly, e.g. acceptance tests, training and project documentation in order to create a competence gap concerning system configuration amongst client's application users and make it impossible for them to introduce changes to the configuration independently, e.g. in stage 3, which will increase the number of client's orders from the supplier.

To sum up, in their entire life cycles, the analyzed projects did not end in full project success, however we need to stress that the projects were not interrupted and abandoned. Two main results of project complications were observed, i.e. exceeding the budget both in stage 2 and 3, and exceeding the time frame without justification while completing stage 2. In each case, the clients engaged external consultants, who in the first two stages of the project played the part of the "client's advocate", indicating the dangers linked to the supplier's behavior and the possible consequences of their materialization. Another role of the external consultants was suggesting counteractions to the client, in order to minimize the danger of the project ending in a total failure. Most probably, it was the engagement of external consultants that led to the projects' partial success and prevented them from failing completely. In spite of this, the respondents from the client's side clearly indicated that the supplier's behavior, through which the moral hazard manifested itself, was the source of project complications. Amongst the researched IT projects, there were three types of behavior that demonstrated moral hazard on the supplier-agent's side. In-depth analytical workshops with the representatives of the client-principal showed that the members of client's project group diagnosed the described behavior already

[4] *Polish Language Dictionary* (Warszawa: PWN, 1996), defines opportunism as concentrating on one's personal interest while ignoring common goals and generally assumed codes of behavior.

during stage 2 of the project. Moral hazard at this stage of the IT project leads to a deep crisis of trust and decreases the chances of a successful project completion. The mutual loss of trust influences the increase of transaction costs [11], especially the cost of monitoring the agreement implementation, the adjustment cost and the cost of terminating the agreement by the client, which can on its own lead to a complete failure of the project. In analyzed cases, increasing client's transaction costs was visible mostly in the increasing workload of the tasks. During in-depth workshops, the respondents on the client's side made suggestions that they would like to pass on to the future client's project managers, who will be responsible for the implementation of stages 1 and 2 of the IT projects:

- Engaging external consultants as client's "advocates" with the aim to support the completion of stages 1 and 2.
- Professional preparation for project implementation by the workers from client's project group:
  - o Specifying precisely (ex ante) the functional requirements of the system.
  - o Specifying precisely (ex ante) the direct and indirect benefits of the project.
  - o Specifying precisely (ex ante) the requirements for the supplier.
  - o Specifying precisely (ex ante) the requirements for the IT system.
- Designing an implementation agreement, which could eliminate the morally hazardous behavior of the supplier.

We need to stress here that the presented types of supplier's behavior are universal and they concern all the relations between the supplier and the client, also in case of contracts signed in public administration.

## IV. Conclusions

The respondents stressed that the agreements of software licensing purchase, service purchase and Service Level Agreement are not simple and even in the current economic structure there are very few specialized lawyers who, linking their legal and MIS knowledge, could secure the company's interests. During in-depth analytical workshops, the respondents indicated that many lawyers de-

Table III.
The results of multiple case study research

| | Company X | Company Y | Company Z | Company A |
|---|---|---|---|---|
| Client's (principal's) profile | Bank | Manufacturing company | Service company | Legal firm |
| Type of purchased IT system | BI | ERP | CRM | DMS |
| Supplier's (agent's) profile | Reseller of software designed by the market leader | Reseller of software designed by the market leader | Reseller of software designed by the market leader | Reseller of software designed by the market leader |
| System operation period | 5 years | 4 years | 7 years | 7 years |
| Implementation results | Project not completed on time without justification, stage 2 completed within the budget, operational costs increased significantly in stage 3 compared to the original estimate, not all the business goals completed. | Project not completed on time without justification, stage 2 not completed within the budget, operational costs increased significantly in stage 3, not all the business goals completed. | Project not completed on time without justification, stage 2 not completed within the budget, operational costs in stage 3 according to the plan, all the business goals completed. | Project not completed on time without justification, completed within the budget, not all the business goals completed. |
| Type of implementation services | Fixed budget | Fixed budget | Fixed budget | Fixed budget |
| How does the moral hazard in the principal (client) – agent (supplier) relation manifest? | The budget for the completion of stage 2 was calculated by the supplier on the basis of a lower hourly wage for the consultant than stipulated by the maintenance service budget in the SLA agreement regulating the system maintenance services in stage 3, i.e. the operation stage. The change of rates took place directly after the completion of stage 2. Thus, the total system maintenance cost (TCO) increased in relation to the estimates completed in stage 1. | Engaging people with low qualifications and little professional experience in the area of a specific ERP system in the given manufacturing company. Thus, the client does not obtain the benefits resulting from the knowledge transfer and good practice.<br>Decreasing the engagement of supplier's consultants in some implementation tasks, e.g. completing acceptance testing with system calibration in comparison to the obligations included in the agreement concerning the workload of this task. | The low quality of knowledge transfer carried out by the supplier, both regarding the configuration and operation of the system by its end users, with the aim of creating the client's competence gap.<br>The supplier's consultant-programmer programs the system to limit the workload during the completion of stage 2, e.g. by introducing constants instead of variables in the code, which results in eliminating the possibility of configuring a given function simply and without an excessive workload in stage 3. | Engaging people with low qualifications and little professional experience in the area of a specific CRM system implementation. The people engaged by the supplier as key consultants had only graduated from university 2-3 years beforehand. Decreasing the engagement of the supplier's consultant in certain implementation tasks (e.g. ensuring the quality of data migrated from the old system to the new system, training key system users) in comparison to the obligations included in the agreement regarding the workload of the task. |

signed the agreements trying to eliminate the moral hazard, struggling between two extreme approaches, i.e. case-specific agreement, which was supposed to foresee and eliminate the majority of dangers resulting from cooperation in stage 2 and 3, and a simplified agreement regulating only the basic aspects of the cooperation. In case of case-specific agreements, transaction costs increase dramatically[5]. The presented research results point towards three basic behaviors of the supplier where moral hazard occurred in IT projects, resulting from the opportunism of the supplier-agent and a high level of information asymmetry between the client (principal) and the supplier (agent). All the three types of behaviors increased the risk of an unsuccessful IT project implementation and influenced the transaction costs, also through the loss of mutual trust. Identifying the unwanted behavior of the supplier allowed the client's project group to react and to minimize the risk of the project's failure.

New institutional economy treats enterprises as a cluster of contracts and it makes the success of the projects dependent on the quality of agreements. We need to underline that institutional economy focuses not only on analyzing the available choices and making those choices, but also on solving conflicts as part of the enterprise's stream of contracts. Considering the current and projected saturation of the market with management support information systems, we need to stress that a cluster of contracts regulating IT outsourcing constitutes one of the most important agreement groups in a large group of enterprises. Until now, IT projects were rarely analyzed from the perspective of institutional economy by Polish theoreticians of business informatics, thus creating a research gap. Special attention in this area is owed to the work by J. Auksztol [3]. I believe that tightening the knowledge gap and promoting the knowledge, as well as practical skills in this area, may increase the effectiveness of implementing IT projects based on outsourcing, both in enterprises and government agencies.

## References

[1] J. N. Lee, S. M. Miranda, Y. M. Kim, "IT outsourcing strategies: universalistic, contingency, and configurational explanations of success," *Information Systems Research,* vol. 15, no. 2, pp. 110–131, 2004.

[2] J. Dibbern, T. Goles, R. Hirscheim, B. Jayatilaka, "Information systems outsourcing: a survey and analysis of the literature," *The Data Base for Advance in Information Systems,* vol. 35, no. 4, pp. 6–102, 2004.

[3] J. Auksztol, *IT Outsourcing in Management, Theory and Practice (Outsourcing Informatyczny w Teorii i Praktyce Zarządzania).* Gdańsk: University of Gdańsk Press, 2008, pp. 50–51.

[4] K. M. Eisenhardt, "Agency theory: an assessment and review," Academy of Management Review, vol. 14, no.1, pp. 57–74, 1989.

[5] A. E. Dembe and L. I. Boden,"Moral hazard: a question of morality?," *New Solutions,* vol. 10, no. 3, pp. 257–279, 2000.

[6] T. Gruszecki, *Modern Enterprise Theories (Współczesne Teorie Przedsiębiorstwa).* Warszawa: PWN, 2002, p. 193.

[7] J. M. Buchanan, "A contractarian paradigm for applying economic theory," *The American Economic Review,* vol. 65, no. 2, pp. 225–230, 1975.

[8] Y. Lichtenstein, "Puzzles in software development contracting," *Communications of the ACM,* vol. 47, no. 2, pp. 61–65, 2004.

[9] R. Yin, *Case Study Research: Design and Methods.* Thousand Oaks, CA: Sage Publications, 1984.

[10] B. Flyvbjerg, "Five misunderstandings about case-study research," in *Qualitative Research Practice,* C. Seale, G. Gobo, J. F. Gubrium, D. Silverman, Eds. London-Thousand Oaks: Sage Publications, 2004.

[11] O. E. Williamson, "Market and hierarchies: some elementary considerations," *The American Economy Review,* vol. 61, no. 2, pp. 112–123, 1971.

[5] In one of the analyzed cases, the case-specific implementation services agreement for stage 2 was over 300 pages long, it was negotiated for 4 months, while the implementation project lasted 3 months and did not exceed 50 000 PLN.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1563–1568

DOI: 10.15439/2015F365
ACSIS, Vol. 5

# Big data as a business opportunity: an Educational Perspective

Ilona Pawełoszek
Częstochowa University
of Technology,
Management Faculty
Poland
Email:
ipaweloszek@zim.pcz.pl

Jędrzej Wieczorkowski
Warsaw School of Economics
Institute of Information Systems
and Digital Economy
Poland
Email:
jedrzej.wieczorkowski@sgh.waw.pl

*Abstract*—**The paper presents considerations on the concept of big data. The aim of the paper is to confront the attempts of defining big data with its common understanding by different groups of users. In this research the group of respondents are students of Warsaw University of Economics. The authors advocate that the student's opinion and attitude to big data can be important regarding the fact that they will probably become managers - business users of big data solutions. Another aim is the analysis of present educational offerings of Polish universities in the area of big data and suggesting the directions of the development of educational curricula to better fit the marked demand for big data professionals.**

## I. Introduction

THE attempt to demystify Big Data may on one hand lead to numerous definitions that can be found in academic literature, corporate and individual blogs of companies and people interested in latest advances, algorithms, and practical applications in this field [1] [2]. On the other hand some publications regarding better common understanding of Big Data and its social aspects can be found [3].

The term 'big data' is very simple in its construct, it is a combination of two frequently used words 'big' and 'data'. The simplicity of the term makes even non-specialists have their own idea on what big data really is. The term suggests gathering, processing and using large volumes of data.

'Big' is a relative term which changes its meaning along with technology development. Particularly performance and capacity challenges that are associated with big data systems evolve over time in their nature, scale and scope. Thus it is not surprising there is much confusion surrounding the term.

The phrase 'big data' has been extensively used in the recent years. For example analysis of keyword popularity with the Google Trends tool reveals the constant, systematic and regular growth of number of searches with the keyword 'big data'. This phenomenon has been taking place since the year 2012 (see. Fig.1). However the question about context of the searches stays open so the further research is sorely needed.

Over the past five years, the emergence of huge data sets and data-intensive science are fundamentally changing the way researchers work in virtually every scientific discipline. This phenomenon can be easily observed while analyzing bases of scientific publications (see. Fig.2.).

Big Data is currently understood as the term increasingly used to describe the process of applying serious computing power to seriously massive and often highly complex sets of information [4], data sets whose sizes exceed the capacity of

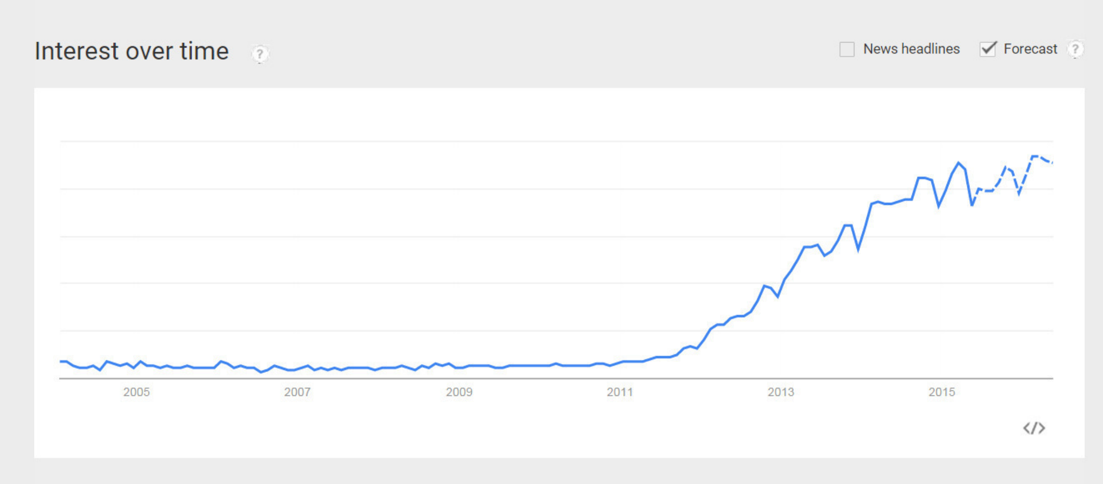


Fig. 1 Big data search term – interest over time with forecast, Source: Own elaboration

978-83-60810-66-8/$25.00©2015, IEEE

conventional database tools for gathering, storing, managing and analyzing data [5]. The above definitions are timeless and universal because they are not domain-specific and do not specify the typical big data features in any point of time, however they emphasize the change that has happened as compared to 'conventional database'. This point of view is highlighted for example by Dumbill [6] who states that: Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of database architectures.

Other examples of definitions underline the unstructured character of data, i.e. accordingly to Rouse [7] big data is a general term used to describe the voluminous amount of unstructured and semi-structured data a company creates - data that would take too much time and cost too much money to load into a relational database for analysis.

Model 3V of META Group [8] is usually considered a starting point for describing big data in the modern sense. The big data notion was not considered then, but its 3 basic features were taken into account: Volume (very large volumes of processed data), Velocity (variability and dynamics of processed data), Variety (variety of processed data). Later various authors tried to choose other features characteristic for big data, which could enhance the 'v' model, particularly: Value – the processed data is valuable, Veracity – another way 'accuracy' or 'truthfulness', Visualisation – the possibility to present the vast amount of data in a way that would be comprehensible, easy to understand and read.

The 'Volume' feature is the most difficult to verify among the original 3V because the amount of data significantly increases with the IT development. The volume of large datasets may differ depending on the area of application. The remaining 'v' features seem very important: the possibility of real-time analyses of particular variables, such as stream data, sensor data (velocity) and analysis of various unstructured data (variety). These features are characteristic to contemporary big data in contrast to the previous approaches to massive data processing.

Attempts to define the term big data are as well very general and timeless as bringing up the technological aspects and trends which on one hand describe the concept in a more detailed way but on the other hand go obsolete very quickly. At the same time, because of dynamics of the discussed term different opinions can be found agreed that the concept is not yet fully definable.

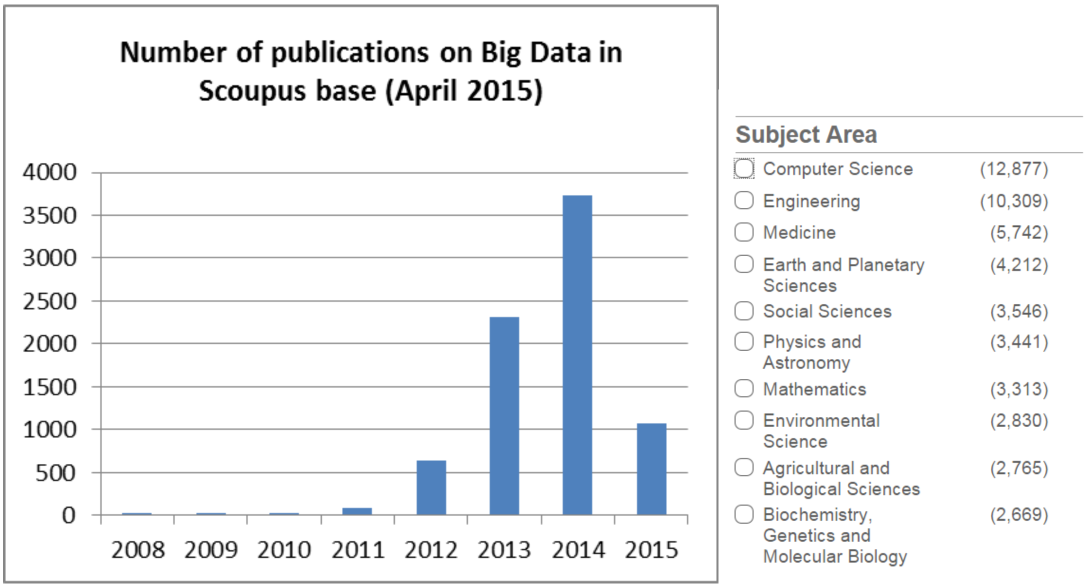
## II. THREE-ASPECT APPROACH TO BIG DATA. THE AUTHOR'S WORK ON UNDERSTANDING OF BIG DATA CONCEPT

The lack of unified definition of the term big data coupled with the high recognition and popularity as well in scientific papers as in popular publications induced the Authors to investigate its common understanding by different professional and social groups.

The Authors conducted research on the context of using the term 'big data' in press articles from daily newspapers [3]. The analysis of online edition the newspaper of Agora (one of the biggest media companies in Poland) was conducted twice: in March and September 2014. The thematic context of 10 of the top search results for the key word 'big data' had been analyzed after excluding the incidental cases where the document's context was not relevant to the research subject.

The Authors proposed "three-aspect approach", identifying three essential aspects of big data: technological, business, social.

**The technological aspect** in the above classification represents a focus on the methods of big data analysis and information technology used. Considering the methods of analysis the big data concept is based first of all on well-grounded statistical methods, artificial intelligence, machine learning and data mining. Similarly to the latter, the aim of big data analysis can be exploration of implicit associations and patterns in large data sources. The big data approach is to analyze all the accessible transaction data rather than a representative sample. Such approach does not require a predefined model and hypothesis, but enables discovering



Subject Area

| | |
|---|---|
| ○ Computer Science | (12,877) |
| ○ Engineering | (10,309) |
| ○ Medicine | (5,742) |
| ○ Earth and Planetary Sciences | (4,212) |
| ○ Social Sciences | (3,546) |
| ○ Physics and Astronomy | (3,441) |
| ○ Mathematics | (3,313) |
| ○ Environmental Science | (2,830) |
| ○ Agricultural and Biological Sciences | (2,765) |
| ○ Biochemistry, Genetics and Molecular Biology | (2,669) |

Fig. 2 Number and subject areas of scientific publications on big data in Scopus base, Source: Own elaboration

previously unknown patterns in all the accessible data. Moreover the big data approach gives a possibility to analyze unstructured data such as texts from internet pages.

**The Business aspect** of the described classification focuses on applications of big data. Considerably shrinking cost of data gathering and processing makes it worth to process the data which were previously not possible or economically not feasible. The new approach can be the solution for information overload which refers to the situation when there is enough accessible and relevant data but there are no capabilities to process or draw conclusions on the basis of the data [9]. The big data can be considered the next stage in evolution of business intelligence and predictive analysis. In this context the possibility of using transactional data directly for analysis seems to be very important qualitative change. This change is especially viable for implementation of new organizational solutions and methods. As Płoszajski [10] notices in history it was always so that when the prevalence replaced rarity, the necessity for new business models arose.

Part of the new applications are old but improved solutions for data mining, risk modelling, fraud detections, churn predictive modelling, load forecasting for electric power systems, stock technical analysis, optimization of logistics [11]. Combining the concept of big data with other ideas (e.g. GIS, cloud computing) allows many new benefits, for example in supporting public administration, regional electronic communities [12].

**The social aspect** of big data is in the Authors' opinion associated with social consequences of data processing results. These matters are often overlooked in definitions of big data concept. Alongside the clear and unquestionable advantages for society, there are also some threats, the first of which is associated with massive processing and using of personal data. In contemporary world every person leaves large amount of electronic traces such as records in the data bases or different archives of unstructured data. These are the data describing the activities of the internet user, in particular e-mails, posts on social networking websites, pages visited, keywords and phrases input into search engines, billings and geolocation data from mobile networks, data on financial transactions (payments with credit cards, bank transfers), data from loyalty programs, purchases, medical sensitive data, video from monitoring systems. The abovementioned data may have great commercial value (for example enable personalized targeting and displaying custom ads) and public value (i.e. protection of public security). The essential issue in question is the required level of privacy which can be different in commercial and public applications of big data processing.

Table I. Research on the concept of big in the press articles and notes

| | Aspects | | |
|---|---|---|---|
| Period | technological | business | social |
| March 2014 | 1 | 5 | 8 |
| September 2014 | 3 | 5 | 7 |

Source: Own elaboration

The Author's researches on the concept of big data were focused on assigning the context to the press articles and notes. Because some of the papers were focused on several aspects, the total number of references in each text was over 10. The summarization of research results is presented in table I.

It should be highlighted that the research encompassed only nonprofessional press. Moreover what is puzzling is a very small number of references to technological aspect. The articles which were strictly economical were focused on the examples of big data applications. Most of the references were to the social aspect. The potential threats of privacy were often discussed especially in the contest of data processing by various institutions, the control of the internet and social networking websites by commercial units.

It is generally can be seen in the press articles that particularly interested for journalists is the social aspect, to a lower degree applications and least interesting were the technological and analytical issues. The reason for this lack of interest in technological aspects is probably the lack of expertise and competencies in technology and data processing. Most of the journalists were trying to follow popular, trendy topics. Popularizing big data the journalists often emphasize the influence of technological developments and economic changes on society.

## III. RESEARCH ON KNOWLEDGE BIG DATA CONCEPTS

The Authors decided to explore the knowledge and understanding of big data among students. The aim of the research was also to verify the hypothesis that the social aspect of new technologies is particularly important for the respondents. The studied community was a group of students from the Warsaw School of Economics.

Apart from openness to the new solutions, which is the typical feature of young people, it is vital to be broadminded and realize the potential of the new technologies applications. In particular the students of economic schools should have well-grounded, good idea about ICT applications in widely understood economy. This issue is especially important for this research. The future users of big data – based solutions will probably be today' students of economics, statisticians and analysts (so called data scientists). A broad group of big data users will probably be the management professionals from companies and institutions who will commission studies of big volumes of data. Therefore the group of respondents was intentionally chosen to draw conclusions about the desirable new directions for higher education in the area under consideration.

The group of 140 students underwent the survey using traditional paper questionnaires with closed questions. Among the surveyed persons 75 were first-year students and 65 second of third-year students. In the Warsaw School of

Economics the first-year students implement the standard study program including, inter alia, basic courses on technology and statistics.

In the next years of studies the students have more possibilities to choose the courses and their sequence depending on their interests. For this reason in the research procedure students are divided on two groups – the first year of study and higher (II and III) years of studies. The respondents were asked to state whether they are acquainted with the term 'big data' and more detailed questions regarding the understanding or the considered term (11 questions) and its characteristic features (8 questions) were asked. The data were gathered during the educational activities not relevant to big data concept. Such a form of survey ensured virtually 100% response rate.

The assumption was made that the students have enough knowledge on the basic level to understand the capabilities of modern information technologies and their applications (in particular in business), and also they should realize the social consequences of using big data.

For the question „Have you ever met the term 'big data' in contest of new possibilities of massive data processing?" there were 58% of positive answers. It should be noted that the number of positive answers is increasing. Among the students of the first year there were 40% of positive answers, while on the higher years of studies there was 77%.

Although it cannot be claimed clearly that the subject of big data is included in the obligatory courses for bachelor studies, the teachers may mention about it during the obligatory course of 'Statistics' on the first year of study. Moreover the concept can appear in the content of other courses however it is not fully discussed.

In this case the understanding of big data term is of special interest. The further analysis of survey results will only take into account the students who have acquaintance with the term big data. A clear majority of respondents (68%) understands big data concept very broadly, as every processing of large amounts of data. This group does not think of big data in terms of theoretical considerations and it can be supposed that they understand the term 'big' in a purely common way. However regarding the continuous theoretical considerations on the term big data and the lack of uniform definition it cannot be assumed that this 'shallow' understanding is not correct.

The students' faith in the efficiency of typical big data methods can be observed. Most of them (65%) claim that the characteristic feature of big data is quality of data processing: 67% of the first year students, and 62% of the higher years of study. In consequence most of the students do not see the problems of data quality which are in fact typical to big data. And again the awareness of importance of this problem increases along with the year of studies (17% totally, in which 7% for the first year and 24% for the second and third year).

The majority of the respondents stated that great business value of processed data is characteristic for big data (83% of all the respondents, 77% first year of study, 86% higher year of study). Because the respondents are students of economics it seems natural that they notice the significance of detailed data in management. However it can be noted that the students' attitude to the massive data processing tends to be slightly indiscriminate. And again the criticism increases long with the education level and is probably caused by improving knowledge. Considering the field of study, it can be assumed that the business aspect of big data is more clear for the respondents than the technological one. The technological issues are not explained in detail, similarly the analytical methods are discussed only on the selected study specializations on master degree.

The survey results on social aspect of big data are also interesting. According to most of the students a threat to privacy when using massive data processing for surveillance, monitoring of individuals and identification are not associated with big data (45% of positive answers). This is also the case with privacy law and personal data processing with IT tools (35%). Generally the number of positive answers is not low. Compared with the previous research on journalists, the students pay more attention to possibilities of using big data for economic purposes than on threats to privacy. It can be assumed that young people somehow used to partial loss of privacy, due to the fact they extensively use internet, in particular social networks that is necessarily associated with privacy issues. However to prove the hypothesis it would be necessary to conduct specific studies.

The above discussed results of the survey are related primarily to questions about the opinion: how wide is the concept of big data, which is practical importance and what could be the negative consequences of applications. Other questions allow for verification of the correct understanding of the concept. Most respondents correctly classified technology of high-performance computers into big data idea (62%) and very large amount of processed data (91%). These questions do not relate to technical terms and the answer seems to be intuitive. On the other hand, there is no knowledge of IT terminology associated with big data.

The results showed poor understanding of problems of big data (particularly in IT) among students of economics. There is also an increase in knowledge at the later years of study, mainly in statistics and the possibility of use of big data solutions.

## IV. Big data and education

The widespread use of Internet, mobile computing and sensor networks are generating massive amounts of data available for decision making. As the Internet of Things grows the data can be collected from more sources. The increasing storage capacity and cloud storage solutions allow the data to be gathered across longer periods of time, and a greater number of variables can be monitored, at a relatively low cost and with less effort than ever before.

This situation brings opportunities and challenges to many domains, so the educational institutions should definitely focus on this emerging field to satisfy the market demand for professionals in big data analyses, applications and technologies.

Given the broad scope of the issues covered by big data domain, higher education in economic schools should provide an overview of the possible uses of big data in business and administration institutions. The key aspect is to develop students' ability to critically evaluate the technologies, and consequences of their application. The Authors while browsing information about educational offering were attempting to identify academic courses which are related to any aspects big data.

Table II presents an overview of university courses that strictly or somehow referred to big data. The faculty name in the second column of the table provides a context for the course name. The short research on available courses was made by analyzing the online resources of the USOS systems of different universities where the detailed syllabi of the courses are available.  The syllabi contain content of subjects for each classes. The criteria to choose the course as somehow related to big data were that at least one of the topics has the keyword ‘big data’.

The 15 example courses have been identified among different levels of education: bachelor or master degree, post-diploma and also studies founded from Erasmus program. It is worth to notice that, the range of faculties and their specializations is quite broad from social sciences, through business oriented to strictly technological ones. Five of them refer directly to the phrase ‘big data’ in their titles and all of them are offered by information technology or economic faculties or specializations. These technical and economical faculties also some other courses related to big data such as programming, visualization of information, social networks etc. The term ‘big data’ also appears in descriptions of strictly humanistic subjects, for example offered by Faculties of Polish Studies in the context of social media.

Table II. Overview of university courses that referred to big data

| no. | Course name | University / faculty or institute |
|---|---|---|
| 1 | Big data | Warsaw School of Economics |
| 2 | Big data | Wrocław University of Economics / Management IT and Finance |
| 3 | Big data and business analytics | University of Zielona Góra / IT |
| 4 | Big data processing | University of Warsaw / Mathematics, Informatics and Mechanics |
| 5 | Big data processing | University of Wrocław |
| 6 | Business intelligence systems | Nicolaus Copernicus University in Toruń / Management |
| 7 | Concurrent and Distributed Programming | University of Warsaw / Mathematics, Informatics and Mechanics |
| 8 | Culture 2.0: social media | Jagielonian University in Kraków / Polish Studies |
| 9 | Design and visualization of information | Nicolaus Copernicus University in Toruń / History |
| 10 | International Forecasting and Simulations | University of Warsaw / International Relations |
| 11 | Landscape of today's Polish Internet | Kazimierz Wielki University in Bydgoszcz / Administration and Social Science |
| 12 | Mass media | Universitas Opoliensis / Polish Studies |
| 13 | R and Big data | Warsaw University of Technology / Mathematics and Information Science |
| 14 | Social networks and multi-agent systems | University of Zielona Góra / IT |
| 15 | Social portals - the role and importance | University of Social Sciences / IT |

Source: Own elaboration

The presented research shows that the term ‘big data’ is applied in broad range of domains. The technological aspect (in the context of applied IT solutions) is mainly visible on IT oriented fields of study and in the context of analytical methods, also can be met on IT and economic studies. The business aspect of the analyzed courses is mainly considered from the social media on humanistic studies.

The social aspect of big data is also considered. It should be noted that among the identified courses commercial applications or public administration point of view are not present. We can suppose that there are some courses that deal with business data processing but this is not exposed in the course syllabi as ‘big data’. However it cannot be said that the term ‘big data’ is used only on technical studies.

There are first attempts to create the fields of study addressed to future big data specialists. One of those efforts is the master-level study in SGH (Warsaw School of Economics). “Advanced Analytics – Big Data” has interdisciplinary character and it is placed in scientific disciplines: economics, management sciences, mathematics and informatics. Competencies of the graduates are focused on the problems of acquiring, ordering, storing and analysis of large data volumes using advanced tools. The competencies of the graduates are planned to be the focused on business and economic phenomena and processes.

The new field of study will educate the specialists in data acquiring – as well structural as unstructured – from many resources (i.e. data bases, data warehouses, internet, text files, sensor data and geolocation data) and their analysis which encompasses searching for hidden associations, knowledge extraction, creating prognostic and simulation models and interpretation of the results.

Among the specialization courses apart from detailed economics, there are some courses on data bases and data warehouses, decision rules, optimization methods, graph and network theory, data mining, text mining, data visualization and concept of big data. Moreover the large number of computer laboratories was foreseen with SAS software. [13]

It should be noted that the field of study described above is aimed at educating relatively few analytics and big data professionals. Another issue is education of many specialists of management, finances, administration and other fields, who should have general knowledge on big data, directed on the ability to use this concept.

## V. Conclusions

Above considerations show the multiaspect and relative immaturity of big data concept. The concept is constantly evolving due to the very relative term 'big' and can be understood in many different ways. The research of popular paper articles, survey among students and research on educational curricula of Polish universities also confirm this hypothesis on multiaspect and nonuniform understanding of the big data concept. As the research shows, in spite of high recognition of big data term, its understanding among students of economic studies is mainly intuitive and associated with everything that is related to data storage and processing of large data volumes. The specific big data features poorly recognized by students and often highlighted by big data professionals are: lack or poor structuring of data and their processing for analytical purposes in a real-time.

Apart from poor understanding of the concept, the interesting aspect is the respondents' faith in the business utility of big data solutions. The business aspect for this group is most important. Such attitude is differs from the attitude represented by most of the journalists for whom the business aspect in most cases is an occasion for discussing and describing potential problems of privacy violation and its social repercussions.

The analysis of academic curricula related to big data also indicates its broad conceptual scope. Depending on the field of study various aspects are discussed. At present the students of economic studies are supposed to be future managers. Once on the labor market they will probably expect to have information technologies and analytical methods to bring them valuable data to help them improve business performance. This can bring positive effects in the form of increased IT investments for data analysis and big data applications. In this situation the interest in big data technologies should be used as a chance to improve the level of education in this area, so the future managers could understand real capabilities offered by massive data processing. The education in this field is necessary not only for small group of data analysis specialists/data scientists but also (to some certain degree) to all the students of economic studies. Simultaneously, apart from providing technological knowledge the care should be taken to understanding the social aspect of big data. Improper use of personal data may lead to society's reluctance and in consequence for example to establishment of legal regulations significantly limiting the applications of new technologies of massive data processing what could be harmful for the country economy.

## References

[1] D. Boyd, K. Crawford, "Critical questions for big data" in *Information, Communication & Society*, Volume 15, Issue 5, 2012, pp. 662-679.

[2] M. Tabakow, J. Korczak, B. Franczyk, „Big data – definicje, wyzwania i technologie informatyczne" in Business Informatics, 1 (31) 2014, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, pp. 138-153.

[3] J. Wieczorkowski, P. Polak, "Big data: Three-aspect approach" in *Online Journal of Applied Knowledge Management*, Volume 2, Issue 2, 2014, International Institute for Applied Knowledge Management, pp. 182-196.

[4] http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/04/15/the-big-bang-how-the-big-data-explosion-is-changing-the-world.aspx, 2013.

[5] McKinsey Global Institute, "*Big data: The next frontier for innovation, competition and productivity"*, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, 2011.

[6] E. Dumbil, "What is big data? An introduction to the big data landscape.", http://radar.oreilly.com/2012/01/what-is-big-data.html, 2012.

[7] M. Rouse, "Big data", http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data, 2011.

[8] D. Laney, "*Application delivery strategies"*, META Group, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[9] J. Wieczorkowski, M. Dałek, „Problem przeciążenia informacyjnego a integracja systemów informatycznych", Zeszyty Naukowe nr 762 Ekonomiczne Problemy Usług nr 104, Uniwersytet Szczeciński 2013, pp. 439-448.

[10] P. Płoszajski, „Big Data: nowe źródło przewag i wzrostu firm", *E-mentor*, nr 3 (50) / 2013, pp. 5-10.

[11] R. Szupiluk, *„Dekompozycje wielowymiarowe w agregacji predykcyjnych modeli data mining"*. Warszawa: Oficyna Wydawnicza SGH, 2013.

[12] D. Jelonek C. Stepniak, T. Turek, *"Barriers in Creating Regional Business Spatial Community"* in Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, pp. 1243-1250, DOI: 10.15439/2014F264

[13] *„Program kształcenia na studiach drugiego stopnia w Szkole Głównej Handlowej w Warszawie na kierunku Analiza danych – big data"*, Szkoła Główna Handlowa w Warszawie, 2014.

DOI: 10.15439/2015F93
ACSIS, Vol. 5

# An analysis of the opportunities and challenges connected with utilization of the cloud computing model and the most important aspects of the migration strategy

Janusz Wielki
Opole University of Technology
ul. Luboszycka 7, 45-036 Opole,
Poland
Email: Janusz@Wielki.pl

*Abstract*—**This paper is devoted to the cloud computing model and the opportunities and challenges connected with its utilization by business organizations, including the most significant issues related to the cloud migration strategy. It is composed of five parts. In the first, some facts concerning cloud computing as a new technology and the reasons for organizations using this type of computing model are provided. Next, the notion of cloud computing is briefly defined and the most significant opportunities and possibilities linked with cloud-based solutions are presented and discussed. The following part is focused on the analysis of the most important problems and challenges connected with the cloud computing model. The penultimate part of the paper deals with the issues connected with the cloud migration strategy, focusing on the most important elements. In the final part of the paper, the most significant conclusions and suggestions are offered.**

## I. Introduction

THE dynamic development of solutions available using the cloud-based model and the growing interest in them can be noticed taking place over recent years. Cloud computing is perceived as one of the components of the next-generation IT infrastructure (NGI) of contemporary organizations and a very important technology of the future [1]. According to the report by the McKinsey Global Institute, it is regarded as one of twelve potentially disruptive technologies that will transform business and the global economy [2].

The importance of cloud computing as a technology has been confirmed by various studies. The results of the 18th Annual Global CEO Survey (1322 company leaders from 77 countries) conducted by PwC indicate that cloud computing is considered by CEOs as one of the six most strategically important developments in digital technologies [3]. According to the results of the 2014 Technology Innovation Survey, conducted by KPMG among almost 800 global technology industry leaders, cloud computing is a top technology that will have the greatest impact in driving business transformation for enterprises [4]. In addition, the results of a survey of ICT professionals, carried out by the German Federal Association for Information Technology, Telecommunications and New Media (BITKOM), confirms the importance of this technology. 64% of those surveyed consider cloud computing as the leading information and communication technology trend [5].

The significance of cloud computing additionally increases because of the fact that it enables other highly impactful technologies, including: mobile Internet, automation of knowledge work, the Internet of Things, and Big Data [2], [6], [7].

So, in the context of the importance of this technology and its transformational potential, it is pertinent to analyze a few of the key issues connected with it, which this paper will proceed to do. They include the following aspects:

- the opportunities and benefits connected with cloud computing utilization,
- challenges and concerns connected with cloud computing usage,
- the cloud computing strategy and its key elements.

## II. Opportunities and benefits connected with cloud computing utilization

There are many diverse benefits and opportunities arising from the adoption of a cloud computing model and different mixes of the delivery (SaaS, PaaS, IaaS) and deployment (public, community, private and hybrid clouds) models. Although cost is very often indicated as a key benefit, ([8], [9]) in fact there are much more important advantages associated with this approach. According to the results of the above mentioned 2014 Technology Innovation Survey, there are four main categories of benefits connected with this technology. They include [4]:

- improved business efficiencies/productivity (37%),
- cost reductions (22%),
- faster innovation cycle (11%),
- accelerated time to market (10%).

Other reports also often indicate considerations such as [2], [10], [11]: increased agility, opportunities connected with the implementation of new business models, improved collaboration among business units and partners, geographic expansion, creating new opportunities for SMEs. If the issues connected with an organization's productivity are concerned, there are two main considerations. They are [2]:

978-83-60810-66-8/$25.00©2015, IEEE

- infrastructure and operating expenses,
- application development and packaged software.

According to estimates by the McKinsey Global Institute, in relation to the first of these, productivity gains could reach 20-30%. They result from [2]:

- reduced infrastructure and facilities footprint,
- high task standardization and automation.

In the second case the McKinsey Global Institute estimates that productivity gains could reach 10-15%. They result from [2]:

- standardization of application environment and packages,
- faster experimentation and testing.

There are many aspects of the cloud computing model which can result in lower costs. For example, if the physical IT infrastructure is considered the most important aspects include [2], [8], [9], [12], [13]:

1. Reduction or elimination of waste related to the low level of hardware utilization.
2. Reduction of costs connected with hardware maintenance.
3. Lowering costs related to energy consumption.
4. Possibilities for the permanent analysis of costs and the selection of the optimal service level.

In the case of software utilization, the cloud computing model can lead to the reduction or elimination of costs connected with:

1. Purchase and installation of software, its maintenance and upgrade.
2. Purchase of wrongly selected software (see [14]).
3. Low level of software usage.
4. Developing and testing of applications.

There are many new factors connected with cloud computing which have an impact on faster innovation cycles. Easy and cheap access to tools for the development and testing of new products or services, e.g. cloud-based Big Data tools, is a good example in this context [7]], [15]. Big Data tools are already being used for this purpose by a growing number of companies [16].

Simultaneously, the implementation of cloud-based solutions leads to a diminishing demand for IT department employees to be responsible for the maintenance of an organization's physical IT infrastructure and, correspondingly, the possibility to release some of the budget previously allocated for this purpose. This allows for the reinvestment of the savings on innovative products or services. According to the results of the earlier cited survey conducted by the Manchester Business School and Vanson Bourne, this occured in the case of 62% of the companies surveyed [9].

Use of the PaaS model is a good example as far as accelerated time to market is concerned. In this case companies which develop their own software instead of creating their own environment can instantaneously use ready-made tools for the application building process, delivered to them as a service.

If company agility, understood as the capacity of an organization to identify and capture opportunities more quickly than competitors, is considered, cloud computing significantly increases the possibilities of companies in this respect, due to the fact that utilization of the cloud-based model considerably broadens opportunities for the quick and flexible adjustment of an organization's IT infrastructure to new needs or new market situations. Such situations can require the implementation of new applications, adding new services or increasing computational capacity. In addition, using cloud-based solutions can be quicker than using a company's own staff [2].

Utilization of the cloud computing approach also provides organizations with numerous new opportunities to implement new business models. In many cases, these business models would not be feasible without usage of this computational model. An innovative business model called car sharing, implemented by the Zipcar company, is an example of such a situation. This business model is based on a complicated management system of a single set of cars which are shared by many users, which would not be possible without an advanced IT system where one of the key elements is the utilization of the cloud computing model [17].

Improved collaboration among business units and partners is made possible by the provision through the cloud of easily accessible, continually developing, applications. This aspect, combined with the above mentioned opportunities to build and implement new business models, provides organizations with new possibilities for geographic expansion.

The cloud computing model also provides small and medium enterprises (SMEs) with significant opportunities, especially in respect of costs. In the case of the smallest SMEs or start-ups with only small levels of capital at their disposal, this is not in relation to the reduction of costs previously incurred for IT infrastructure but rather opportunities to access hardware and software which would not be achievable in the traditional computational model, owing to financial barriers, particularly around the purchase of hardware and software and the employment of skilled IT workers to cover maintenance. The cloud model and the associated possibilities of "hiring" services connected with physical IT infrastructures or applications, enables smaller firms to more effectively compete with large organizations [8]. They can also access sophisticated solutions such as the above mentioned Big Data tools or the programming environment in the PaaS model.

## III. The most important challenges and concerns connected with cloud computing utilization

As is the case of all IT solutions, those offered in the cloud model bring not only benefits and opportunities but also problems and challenges. According to the results of the above mentioned 2014 Technology Innovation Survey, there are three main groups of challenges connected with the cloud computing model. They include [4]:

- security (23%),
- technological complexity (16%),
- risk management (15%).

Security anxiety is undoubtedly the key concern connected with this technology, and is confirmed by the results of other studies [2]. According to the Cloud Security Alliance there are various types of security concerns including such aspects as: data loss/leakage; account, service, and traffic hijacking; shared technology vulnerabilities or insecure application programming interfaces [18]. Of course the level of the potential risk depends on many factors, including the type of cloud being used, the service provider, the technologies being used (including data encryption) as well as the procedures it uses, or the procedures applied by a client-organization. In the latter case, it also relates to the phenomenon called shadow IT, which is the use by employees of cloud-applications not approved by their IT departments for business purposes [19].

The spread of the cloud computing model should also reduce the issue of technological complexity, due to the fact that service providers are likely to do everything to make their solutions as simple and as easily manageable as possible. Such a trend is already perceivable (see [20]).

The third issue which causes the biggest anxiety among managers is risk management connected with the utilization of the cloud computing model. Apart from the above mentioned security-related issues, some of the most import challenges are those connected with the availability of services. The deeper the dependency of an organization on cloud-based solutions, the more important this issue is. In spite of numerous publicized cases of problems with the availability of cloud services from well known providers, availability of this type of service is generally at a very high level. This fact is confirmed by the results of various studies (see [21]). Regardless of the level of availability of cloud services, every organization which utilizes the cloud computing model has to have appropriate procedures and technological solutions in place in case of problems with access to services. The same relates to the management of other types of risks.

Apart from the above mentioned issues, there are other challenges and limitations connected with the implementation and utilization of the cloud computing model. In the case of technical issues, one of the most important, and often underestimated issues, is network capacity. This is due to the fact that cloud-based technology is deployed through massive data centers that necessitate high-capacity bandwidth [2]. Simultaneously bandwidth requirements can significantly differ depending, for example, on the type of cloud-based application (basic, intermediate or advanced) [6].

The next element can constitute a significant hurdle in the process of adoption of a cloud computing approach, which is a reservation regarding the usage of the cloud-based model (lack of trust in cloud-based solutions). Such reservations are typically connected with issues such as concerns about placing sensitive data on third-party servers somewhere in the world. An important element of these concerns is the earlier mentioned issue relating to the reliability of cloud-based solutions. In spite of improvements in cloud technology, high-profile downtime accidents continue to take place. As a result, they affect public perceptions concerning the reliability of cloud-based solutions.

The next significant challenge which can constitute an important barrier to the implementation of cloud-based solutions are structural issues and cultural resistance in organizations' IT departments. It is connected with the fact that usage of such computation models causes deep changes in IT management practices and the functioning of IT departments. It can, and in many cases does, lead to raising concerns about loss of control and the lowering of significance and position of these departments in companies. In fact, in the new "cloud realities", their role will be significantly different, moving to that of broker of IT services [22].

The newly required skill sets are the next issue which can be a source of fear and which can cause resistance. It relates to the skills of the employees of IT department and not only of technical nature. In this context a demand for a new type of manager, called "true program managers", has been indicated. They should be leaders who know how to cooperate with internal technologists and third-party vendors in order to pull together a particular solution and deliver it as an overall program [1]. Another significant factor is connected with the complexity of migrating enterprise IT systems to the cloud [2].

There are also many legal challenges. They relate to such aspects as: regulations concerning the place of data storage and access to that data, data ownership, privacy and data protection issues, the applicability of the law connected with data protection and the scope of vendors' responsibility (including liability for data residing in a particular online location). These issues have yet to be settled by policy makers, and a significant barrier is the fact that the law in many countries does not address these issues. An important constraint in the ability to take advantage of some of the benefits of cloud-based solutions (especially those connected with public clouds) is the fact that in many countries data protection laws restrict the possibility of storage and transfer of some types of data outside their borders [2], [23]. A further legal issue constitutes a significant concern in the case of the utilization of public clouds. It relates to inquiries from governmental agencies (e.g. the National Security Agency in the U.S.). Namely, public cloud providers can be obliged to provide information concerning their customers, but may be legally barred from informing them about this fact [1].

## IV. The most important elements connected with the cloud strategy

In spite of the growing interest of organizations of various sizes in adopting cloud-based solutions it does not mean that they have a cloud strategy. According to research conducted by IDC only 17.9% of young organizations (fewer than five years of operation) and 10.2% of mature ones (more that five years of operation) have an optimized cloud strategy. That is:

- a broadly implemented, cloud-first strategy which is proactively managed,
- clearly driving business innovation while improving IT operational efficiency [24].

Although one could perceive the utilization of the cloud computing model as a relatively simple issue, this is in fact not the case. There are numerous potential problems and challenges connected with its implementation and many aspects have to be carefully analysed and planned. It is obvious that implementation of cloud-based solutions is simplest in the case of companies without any previous “burdens” and legacy systems i.e. start-ups. But the bigger the company is, with many complicated business processes, the greater the scale of the challenges. Because of this fact it is especially important for organizations to have clear and comprehensive cloud strategies, including the cloud migration plan, in case they decide to utilize a cloud-based model.

Generally a strategy for the utilization of cloud-based solutions and migration to this computation model should be based on three key phase (see Fig. 1).

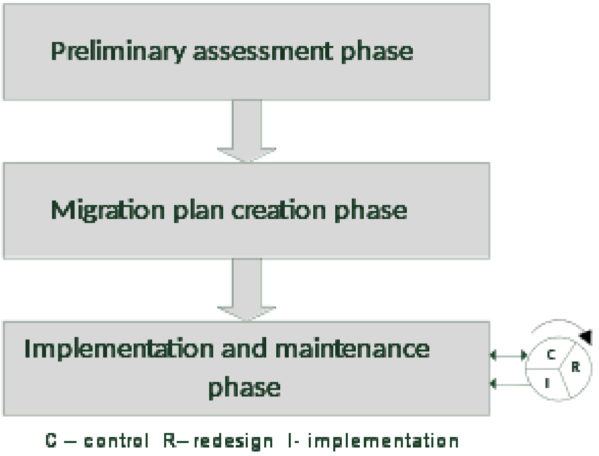

Preliminary assessment phase
Migration plan creation phase
Implementation and maintenance phase
C
R
I
C – control R– redesign I- implementation

Fig. 1 Cloud utilization strategy and its three phases (source: own source)

In the preliminary assessment phase a management board, or especially established special committee, should make a preliminary assessment of the usefulness of using a cloud computing model in the context of its impact on the organization’s functioning.

This relates to such issues as:

- assessment of whether the cloud computing approach aligns with the organization’s culture e.g. in terms of  outsourcing or not outsourcing any of its own operations (a culture of risk avoidance [25]),
- assessment of whether the cloud computing approach aligns with the organization’s objectives and what would be the migration goals (improvement of productivity, cost reduction, increased agility, new business models implementation etc.),
- appraisal of the risk connected with the migration process in the context of the potential impact on it of internal and external factors,
- assessment of which key stakeholders would be impacted by the migration process and how.

In the context of the above mentioned issues, it is evident that many organizations, for various reasons, do not decide to move to the cloud. According to the results of Computerworld’s Forecast survey conducted among 194 IT executives in May and June 2014, almost one third of those surveyed declared that their companies had no such plans (see Fig. 2).

And there is one more significant element of this phase - a comprehensive assessment of the readiness of the organization for the migration process. It is concerned with issues such as technical, human and organizational aspects. According to the results of the Chief Infrastructure Technology Executive Roundtable (CITER) organized by McKinsey, there are six key elements important in the context of an organizations’ “cloud readiness”.

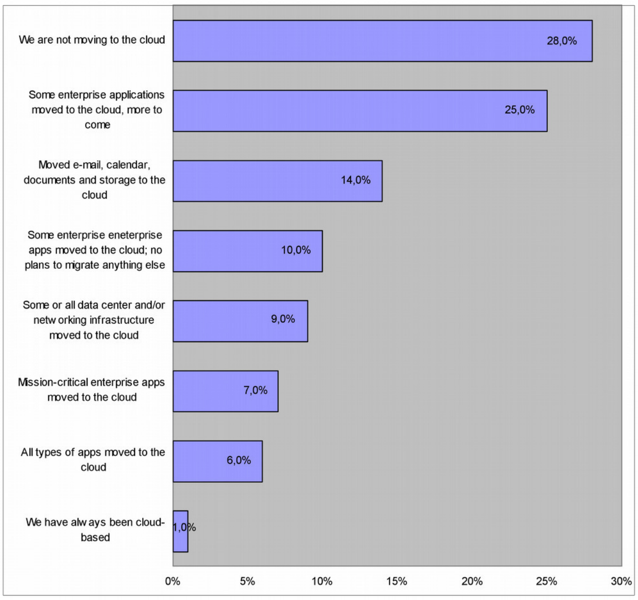


Fig. 2 Organizations’ strategies for cloud computing (source: [11])

They include such issues as [1]:

- well defined and well understood workflows to support cloud offerings,
- highly automated IT infrastructure,
- the right technical skills and training to support cloud of company’s IT department,
- effectively broken down/collapsed by company's IT department technological silos (e.g., Windows, UNIX, storage, networking),
- extensive self-service options (e.g., development/test servers) offered by the company’s IT infrastructure,
- a strong understanding of the issues connected with forecasting the demand for cloud offerings.

In the case of a positive assessment of the sense and advisability of using a cloud-based solution, it is then necessary to move to the second phase - development of a migration plan. In its scope it is necessary to analyze the more detailed issues (including technical, legal and organizational ones) and make final choices, such as:

- determination of the final migration goals and choosing the business processes which should be cloud-supported,
- selection of the deployment model (cloud type) which will be used by the organization,
- selection of the delivery model(s) (SaaS, IaaS, PaaS) which will be applied and determination of the scope of their utilization (type of cloud-based applications, elements of the infrastructure which will be moved to the cloud etc. – see Fig. 2)
- selection of the approach for adopting cloud-based solutions.

In the case of this last issue there are two basic approaches which can be chosen by an organization i.e. “brownfield” and “greenfield”. The first one concerns the piloting of individual technologies and next deploying them in the current IT environment in order to replace solutions used so far. The second one relates to situations in which an organization deploys a separate standalone environment for new applications, and legacy applications are migrated to this new environment [1]. In the context of decisions concerning deployment and delivery models organizations should remember that such choices have a significant impact on issues relating to the level of direct control of the organization over the solution and risk connected with it (see Fig. 3).

Knowing the specific elements connected with the planned delivery and deployment model, the organization can then start the process of selection of a provider of cloud services. In this context an important aspect is to determine the provider’s role and scope of engagement in the migration process as well as the parameters which should be fulfilled by the services delivered in the cloud mode. As far as the first aspect is concerned, according to the results of the above mentioned Chief Infrastructure Technology Executive Roundtable (CITER) organized by McKinsey, there are four important roles of vendors in the process of cloud migration.

They include such issues as [1]:

- implement and integrate solutions,
- provide managed services (e.g. hosting, storage, incident monitoring/response),
- provide IT staff augmentation for ongoing operations,
- design a company's architecture and/or IT infrastructure environment.

As far as the second aspect is concerned, in this case it is necessary to determine the parameters which should be fulfilled by the services delivered in the cloud mode and working out the metrics (Key Performance Indicators) as precisely as possible, allowing for their control [10]. It is also extremely important to determine requirements for such issues as: data security, back-up procedures, location of data, ownership of data or the scope of the provider’s responsibility.

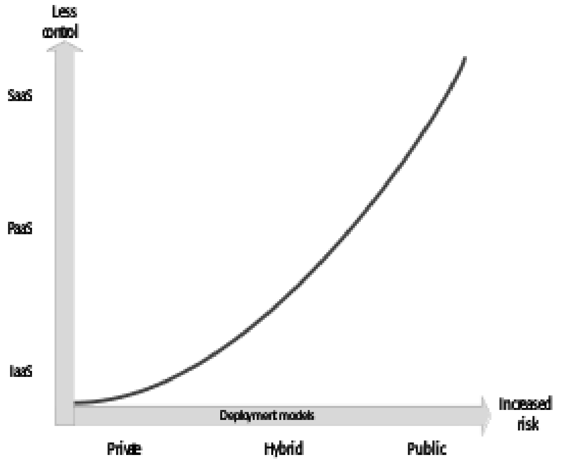


Fig. 3 Inherent risk relationships with cloud service delivery and deployment models (source: [25])

Other important issues which have to be planned include:

- the means of integration of the cloud-based solutions with those which will remain functioning in the traditional way,
- determination of a disaster recovery plan and procedures including a risk management program and incident management,
- organizational changes (especially in the IT department) i.e. their scope, their implementation and overcoming potential cultural resistance,
- necessary training and its scope,
- a cloud governance model.

The third and final phase of activities connected with the introduction of cloud-based solutions to an organization is the implementation phase, the maintenance and monitoring of their functioning in the context of the performance of the organization as a whole. The key issues which have to be realized in this phase include:

- final selection of the service provider and signing the contract including a Service Level Agreement (SLS),
- elaboration, together with the vendor, the migration plan based on the selected approach of adopting cloud-based solutions, including the scope and dates of necessary trainings,
- execution of the planned trainings,
- implementation of the planned cloud-based services and their integration with those functioning in the “traditional” way, as according to the plan,
- testing the functioning of the implemented solutions and making required corrections,
- monitoring the functioning of the implemented solutions based on a previously prepared cloud governance model and making necessary improvements and corrects.

## V. Conclusion

The results of numerous surveys clearly indicate the growing interest of organizations of varying types in IT solutions available using the cloud computing model. This fact is also confirmed by the data concerning the dynamic growth of the cloud computing market and cloud data center IP traffic [6].

Undoubtedly, utilization of cloud services provides organizations with many opportunities. They are not only connected with such issues as the reduction of IT-related costs or productivity improvements, but they also have significant influence on a faster innovation cycle, increased agility or implementation of new business models. Also more and more organizations discern that without cloud-based solutions it would be more difficult to be able to store, analyze and use the rapidly increasing amounts of data critical for their market success and development.

But utilization of cloud-based solutions also brings numerous potential challenges and concerns related to such issues as the reliability of cloud services, data security, privacy or cloud-platforms compatibility. Also the migration process to cloud solutions includes many aspects which have to be well thought out and carefully planned. It relates not only to purely technical issues but also to legal and organizational ones. In the context of the last aspect, one should remember that implementation of a cloud-based approach, especially when it is a larger scale program, will be a type of reengineering project, as it is connected with a deep rethinking of the way an organization conducts its own operations. In this case problems of the various types of potential resistance (especially connected with the IT department and its changing role) and overcoming it, should be carefully addressed. An important challenge is also the lack of so called "best practices" which could be applied by organizations in their migration process (see [11]). In this situation properly planned and executed cooperation with vendors is a very important aspect and one of the success factors.

Taking into consideration the above mentioned remarks, it is necessary to stress that an incorrect and ill-considered approach to the cloud computing phenomenon can lead to a situation where, instead of expected benefits, numerous problems arise, negatively or even destructively influencing an organization's functioning and the realization of its goals. This results from the fact that cloud-based projects are connected to the IT infrastructure, a crucial component of every contemporary organization and the foundation of their functioning.

## References

[1] C. Gnanasambandam et al., "Next-generation IT infrastructure", http://www.mckinsey.com/insights/business_technology/~/media/mckinsey/dotcom/insights/business%20technology/nextgeneration%20it%20infrastructure/next_generation_infrastructure_feb2014.ashx, January 2014.

[2] P. Bisson et al., "Disruptive technologies: Advances that will transform life, business, and the global economy", http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Disruptive%20technologies/MGI_Disruptive_technologies_Full_report_May2013.ashx, May 2013.

[3] PwC, "18th Annual Global CEO Survey", http://www.pwc.com/gx/en/ceo-survey/2015/assets/pwc-18th-annual-global-ceo-survey-jan-2015.pdf, 2015.

[4] KPMG, "The Changing Landscape of Disruptive Technologies", http://www.kpmg.com/PL/pl/IssuesAndInsights/ArticlesPublications/Documents/2015/KPMG-Global-Technology-Innovation-Insights-Fall-2014-online-secured.pdf, 2014.

[5] eMarketer, "Germany's ICT Professionals Focus on the Cloud, Security and Big Data", http://www.emarketer.com/Articles/Print.aspx?R=1011987, February 5, 2015.

[6] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2013–2018", http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.pdf, 2014.

[7] Intel IT Center, "Big Data in the Cloud: Converging Technologies", http://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/big-data-cloud-technologies-brief.pdf, September 2014.

[8] European Commission, "Unleashing the Potential of Cloud Computing in Europe", http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0529:FIN:EN: PDF, september 27, 2012.

[9] T. Olavsrud, "How Cloud Computing Helps Cut Costs, Boost Profits", *CIO*, http://www.cio.com/article/2387672/service-oriented-architecture/how-cloud-computing-helps-cut-costs--boost-profits.html, March 12,.2013.

[10] Oxford Economics, "The Path to Value in the Cloud", http://www.windstreambusiness.com/resources/published-reports/ the-path-to-value-in-the-cloud-strategy, 2014.

[11] M. Pratt, "Cloud computing claims a pivotal role in 2015", *Computerworld*, http://www.computerworld.com/article/2843861/cloud-computing-claims-a-pivotal-role-in-2015.html?nsdr=true, November 10, 2014.

[12] N. Carr, "The End of Corporate Computing", *Sloan Management Review*, vol. 46, no. 3, 2005.

[13] H. Agarwal, "Managing the demand for IT infrastructure", http://www.mckinsey.com/insights/business_technology/managing_the_demand_for_it_infrastructure, 2014.

[14] Sage, „Europejskie firmy marnują 9,6 mld € rocznie", http://media.sage.com.pl/PressRelease.278985.po, 2014.

[15] T. Davenport, *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Boston: Harvard Business School Press, 2014.

[16] M. Aielloi, G. Pagani, "The Smart Grid's Data Generating Potentials", in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, Warszawa, 2014, pp. 9-16.

[17] S. Griffith, "Zipcar: selling cars, one ride at a time", http://whatmatters.mckinseydigital.com/internet/zipcar-selling-cars-one-ride-at-a-time, 2009.

[18] HP, "Five myths of cloud computing", http://www.findwhitepapers.com/force-download.php?id=23344, 2012.

[19] Stratecast, 'The Hidden Truth Behind Shadow IT', http://www.mcafee.com/cn/resources/reports/rp-six-trends-security.pdf, November 2013.

[20] M. Ricknäs, "Amazon offers troubleshooting tool to Web services users", *Computerworld*, http://www.computerworld.com/article/2883840/amazon-offers-troubleshooting-tool-to-web-services-users.html?phint=newt=computerworld_da, February 12, 2015.

[21] M. Gagnaire et al., "Downtime statistics of current cloud solutions", http://iwgcr.org/wp-content/uploads/2012/06/IWGCR-Paris. Ranking-002-en.pdf, 2012.

[22] Cisco, "Your Strategy for Cloud and Our Perspective", http://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/cloud-infrastructure/at-a-glance-c45-730763.pdf, 2014.

[23] P. Van Eecke, "Cloud Computing Legal Issues", http://www.isaca.org/Groups/Professional-English/cloud-computing/GroupDocuments/DLA_Cloud%20computing%20legal%20issues.pdf, 2013.

[24] Cisco, "Cloud Adoption", http://share.cisco.com/cloudadoption/, 2014.

[25] COSO, "Enterprise Risk Management for Cloud Computing", http://www.coso.org/documents/Cloud%20Computing%20Thought%20Paper.pdf, 2012.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1575–1583

DOI: 10.15439/2015F110
ACSIS, Vol. 5

# Risk factors framework for information systems projects in public organizations – Insight from Poland

Ewa Ziemba
University of Economics in Katowice
ul. 1 Maja 50, 40-287 Katowice
Poland
ewa.ziemba@ue.katowice.pl

Iwona Kolasa
University of Economics in Katowice
ul. 1 Maja 50, 40-287 Katowice
Poland
iwona.kolasa@ue.katowice.pl

***Abstract*—The aim of this study is to answer the question about risk factors for the information system (IS) projects in public organizations in Poland. These factors were identified based on a critical review of literature, practical collaboration, the case study and logical deduction. The paper continues as follows. Firstly, a relationship between risk factors and a project success is explained and risk factors presented in the literature are shown. Secondly, a methodology of examining risk factors for the IS project in public organizations is presented. Thirdly, the risk factors for the IS projects in public organizations in Poland are identified and the framework of risk factors presented in the literature is improved. In this framework the factors are classified into eleven groups, namely (1) top management support; (2) manage processes in organization; (3) involve end users; (4) manage information system development process; (5) make system requirement analysis; (6) plan the project; (7) manage, monitor and evaluate the project; (8) manage project team; (9) manage team experience; (10) manage team communication; and (11) public sector procedures and processes. This paper concludes with a presentation of the study's contribution and limitations, implications for the findings and the stream of future work**.

## I. Introduction

THE information systems (IS) projects are always connected to substantial risk. A considerable number of IS projects still use more resources than planned, take longer to complete and provide less quality and functionality than expected [1], [2]. The questions are, what are risk factors for IS projects and how manage risk in IS projects? Among some most common risk factors for IS projects are: unrealistic goals, inaccurate estimation of necessary resources, badly defined requirements, poor presentation of a project status, and unmanaged risk [3], [4]. It has been identified that a poor risk management (RM) of IS projects often leads to failure in IS projects both in public and business organizations [5].

Although some managers claim that they manage risk in their projects, there is evidence that they do not manage it systematically [6]. This shows that public and business organizations should improve not only their ability to identify, but also manage the risk associated with projects [7].

The existing studies mostly examine risk factors for IS projects in business organizations [8]-[10]. There are only few studies concerning risk factors for IS projects in public organizations [11], [12]. This portrays the need for studying risk factors influencing the success of IS projects in public organizations. Therefore, conducting research among Polish public organizations should contribute to greater understanding of risk factors for IS projects and should help fill the gap in the existing body of knowledge.

This article focuses on analyzing risk factors in IS projects in public organizations in Poland. Its aims are to: (1) indicate risk factors for IS projects in public organizations in Poland, and (2) define a risk factor framework for IS projects in public organizations.

The article is structured as follows. Section I is an introduction to the subject. Section II states the theoretical background of risk factors for IS projects. Section III describes a research methodology. Section IV presents the research findings on risk factors for two IS projects in Polish public organizations and the risk factors framework presented in the literature is enhanced. Section V provides the study's contributions and limitations, implications for the findings and the stream of future works.

## II. Literature Review

### *A. IS Project Success and Its Risk Factors*

Is there a relationship between risk factors and project success? This question has been considered relevant by people from both academic and practitioners' communities already for a long time, especially in the area of IS, where projects have a long history of failing [13]-[16]. What exactly is defined as risk? Risk is the occurrence of any event that has consequences for, or an impact on the success of an IS project [17]. Many authors define that all projects involve risk of some sort [18]. There is no project without a risk. Risk management (RM), therefore, is one of the main issues of a project. Its positive impact on planning, decision making, avoiding bad events, and giving a proper response to a risky situation is remarkable [19]. RM is both a science and an art for identifying the treats, assessing and controlling them by applying the most effective manner [20].

The success of IS projects is traditionally measured by time, budget and requirements criteria. Many researchers define a project success in terms of compliance with time limits, cost limits and meeting requirements [21]. The significant impact on projects success has RM [22]. It helps to identify and manage risk, and thereby prevent IS projects from getting off the track. RM involves identifying the potential risk, measuring, monitoring and controlling them

978-83-60810-66-8/$25.00©2015, IEEE

in an organization to meet its strategies and objectives, and leads to decrease the undesired effects in project life cycles.

There is a general consensus that effective planning and implementation of a RM methodology both positively affect the success rate of any project [23]-[26]. There are several methodologies of project RM that represent the course of actions required to manage risk during IS projects [27]. However, the main point is to identify the exact risk factors for IS projects [28]. We found only five papers published after 2010 defining risk factors. Characteristics of the publications are presented in Table I.

Based on the literature findings, there have been identified ten risk groups (RG). Namely, (1) top management support; (2) manage processes in organization; (3) involve end-users; (4) manage information system development process; (5) make business requirement analysis; (6) plan the project; (7) manage, monitor and evaluate the project; (8) manage project team; (9) manage team experience; (10) manage team communication. Each of risk groups is clearly defined by particular risk factors. The compilation of risk factors grouped into risk groups is presented in Table II.

### *B. Risk Factors for IS Projects in Public Organizations*

In the literature, researchers are conducting studies on identifying risk in public organizations. Patanakul [11] conducted research on large-scale IS projects in the public sector. There were defined the exact risk factors:

- system design and implementation;
- problems in requirement identification;
- project management and governance;
- problems in managing project risk;
- problems in project monitoring, control and managing changes;
- problems in project governance; and
- contract management.

Aritua, Smith and Bower [12] run research on risk factors in public sector in the UK. The research focused on public organizations' projects in general, not in the context of IS projects. According to them, rejecting the risk of a general nature, specific to certain sectors of the economy, the risk factors can be distinguished as follows:

- linking strategy and projects;
- difficulties in project delivery;
- skills shortage and resources;
- cash flow and funding problems;
- sustainability and environmental legislation;
- challenges of procurement;
- competition for contractors; and
- change in government policy.

Analyzing the above risk factors, it can be noticed, that some of them are the same as defined in the risk groups presented in Table II. However the risk, namely: contract management, challenges of procurement, and change in government policy are not included among the risk defined for business organizations.

### *C. Risk Factors Framework for IS Projects*

Risk usually comprises a lot of factors interacting with each other. Researchers have built several frameworks to classify the factors and present relations between them [34], such as the MIT90s framework of Morton [35], the project life-cycle framework of Markus and Tanis [36], the strategic-tactical framework of Holland and Light [37], and the process-control-information (PCI) framework of Bemelmans [38]. The literature presents, that the success of IS projects is dependent on the dynamics and interaction of the organizational and technical factors [39]-[40], [41]. The authors structure and classify the risk factors in the MIT90s framework which covers organizational as well as technical issues. The framework is simple and easily extendable and can, therefore, be used in different settings for multiple purposes [34]. For instance, the framework was applied for supply chain information systems critical success factors [34], [42].

The MIT90s framework contains the following dimensions [35]:

- Project strategy – project goals and how the organization fulfills these goals;
- Structure – process, functions, and structure of the project in organization;
- Individual and roles – the roles, skills, knowledge, social ties and attitudes of people;
- Management process - the management process that steers the implementation project; and

TABLE I.
CHARACTERISTIC OF THE PUBLICATIONS

| Publication | Research characteristics | Research result |
|---|---|---|
| S. Liu, L.Wang (2014) | survey, 26 respondents (IS managers) | identified 27 risk factors |
| S. Sundararajan, M. Bhasi, P. K. Vijayaraghavan (2014) | 1 case study | identified 20 risk factors |
| C. Lopez, J.L. Salmeron (2012) | interview, 12 respondents (IS/IT projects experts);<br>risk evaluation using IPA method | identified 46 risk factors |
| L.Jun, W. Qiuzhen, M. Qingguo (2011) | survey, 93 respondents;<br>the influence between factors were measured | identified 7 risk factors |
| P.K. Dey, B.T. Clegg, D.J. Bennett (2010) | 1 case study | identified 41 risk factors |

TABLE II.
RISK FACTORS FOR IS PROJECTS

| Group of risk | Risk factor | | Source |
|---|---|---|---|
| RG1<br>Top management support | R01 | Lack of top management commitment to the project | [29] [30] [32] |
| | R02 | Top managers make important IT decisions without consulting the others | [29] [32] |
| | R03 | Unrealistic projects outcomes | [30][33] |
| | R04 | Excessive project size | [29] [31] |
| | R05 | Change in ownership or senior management during the process of development | [30] |
| | R06 | Time too short/too long | [29] |
| | R07 | Unrealistic schedule | [29] |
| RG 2<br>Manage processes in organization | R08 | Resources shifted away from the project because of changes in organizational priorities | [29] [30] |
| | R09 | Major effect of project implementation on organizational structure | [30] [33] |
| | R10 | Mismatch between organization culture and required business process changes needed for new system | [30] |
| | R11 | Changes in organizational priorities | [29] |
| | R12 | Continuous changes in the organizational environment | [29] |
| RG 3<br>Involve end-users | R13 | Lack of user participation | [29] [30] [31] [32] [33] |
| | R14 | Users resistant to change | [29] [30] [32] |
| | R15 | Target users are unfamiliar with the technology and require additional training | [29] [30] [31] |
| | R16 | Users with negative attitudes toward the project | [30] [32] |
| | R17 | User is not committed to the project | [29] [30] |
| | R18 | Users constantly request further changes | [29] |
| | R19 | Conflicts between users departments | [29] |
| RG 4<br>Manage information system development process | R20 | High level of technical complexity | [29] [30] [31] [32] |
| | R21 | Immature technology | [29] [30] [32] [33] |
| | R22 | New technology and use of technology that had not been used in prior projects | [29] [30] [32] |
| | R23 | Lack of effective development methodology | [30] [32] [33] |
| | R24 | Large number of links to other system required | [29] [30] |
| | R25 | Inadequate system documentation; incomplete or non-existent | [29] [32] [33] |
| | R26 | Lack of proper tests | [29] [32] |
| | R27 | Lack of integration between systems | [29] [32] |
| RG 5<br>Make system requirement analysis | R28 | Continually changing scope and system requirements | [29] [30] [32] |
| | R29 | Unclear or incomplete system requirements | [29] [30] [33] |
| | R30 | System requirements not adequately identified | [29] [30] [32] |
| | R31 | Conflicting system requirements | [30] |
| | R32 | Failure to manage end-user expectations | [29] |
| | R33 | Lack of frozen requirements | [29] |
| RG 6<br>Plan the project | R34 | Poor project planning | [29] [30] [31] [32] |
| | R35 | Inadequate estimation of required resources | [30] [32] [33] |
| | R36 | Critical activities are not identified | [29] |
| RG 7<br>Manage, monitor and evaluate the project | R37 | Project progress not monitored closely enough | [29] [30] [31] [32] [33] |
| | R38 | Lack of an effective project management methodology | [29] [30] [32] [33] |
| | R39 | Ineffective communication | [29] [30] [32] [33] |
| | R40 | Inexperienced project manager | [29] [30] |
| | R41 | Project manager lacks required skills | [29] |
| RG 8<br>Manage project team | R42 | Lack of knowledge management | [33] |
| | R43 | Frequent turnover within the development team | [29] [30] [32] [33] |
| | R44 | Team members are unmotivated | [29] [32] [33] |
| | R45 | Inadequate composition of project team | [29] [32] |
| | R46 | Improper definition of roles and responsibilities | [29] [33] |
| RG 9<br>Manage team experience | R47 | Team members lack of specialized skills required by the project | [29] [30] [31] [32] |
| | R48 | Inadequately trained development team members | [30] [32] |
| | R49 | Team members are unfamiliar with the technology | [29] |
| RG 10<br>Manage team communication | R50 | Conflict and no cooperation between the team members | [29] |
| | R51 | Team member are in many localizations | [33] |
| | R52 | Inadequate team size | [33] |

- Technology – the information system being implemented.

The MIT90s framework indicates that the success of IS projects dependents on the interaction of the organizational and technical system. The framework (Fig. 1) provides opportunities for better understanding of dependency among risk factors. Firstly, risk factors can be grouped into five dimensions of the MIT90s framework which is easy to present from a management perspective. Secondly, the framework of risk factors also provides an understanding of the dynamics and cause-effect relationships of a complex IS projects. The arrows in Fig. 1 indicate that changes in one of the five interacting dimensions are influencing the other. The risk identified in one dimension will cause the higher probability of risk in the other dimension. For example, RG6 Plan the project – the factor R34 Poor project planning, will influence RG8 Manage project team, R45 Inadequate composition of project team. Poorly planned project

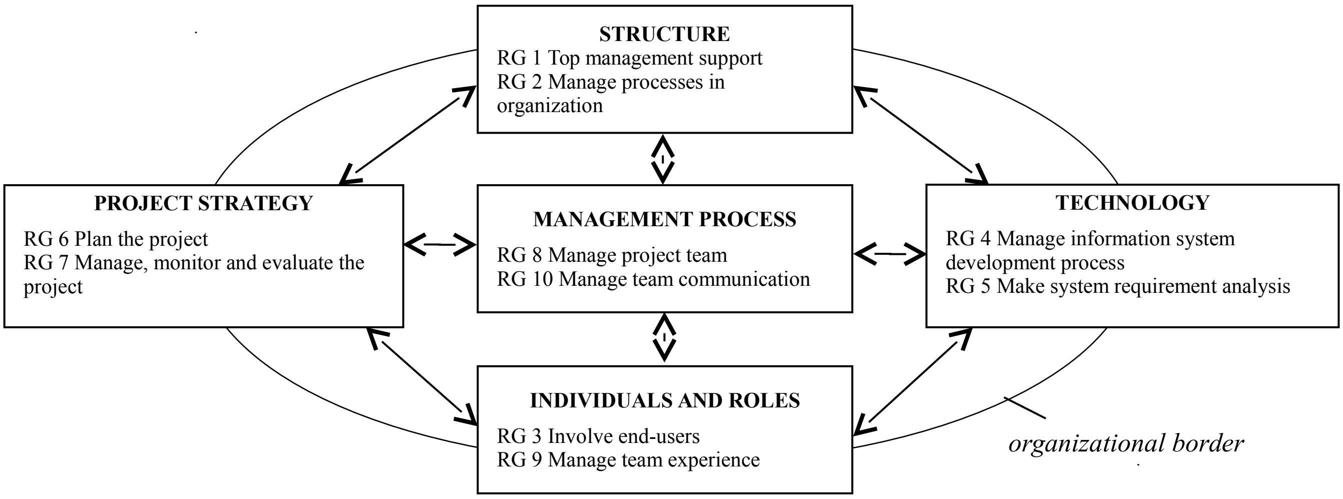

STRUCTURE
RG 1 Top management support
RG 2 Manage processes in organization
PROJECT STRATEGY
RG 6 Plan the project
RG 7 Manage, monitor and evaluate the project
MANAGEMENT PROCESS
RG 8 Manage project team
RG 10 Manage team communication
TECHNOLOGY
RG 4 Manage information system development process
RG 5 Make system requirement analysis
INDIVIDUALS AND ROLES
RG 3 Involve end-users
RG 9 Manage team experience
organizational border

Fig.1. Risk factors framework for IS projects based on Scott Morton [35]

generates risk of underestimation or overestimation of resources.

## III. Research Methodology

The goal of this research was to analyze risk factors for IS projects, described in the literature, in the context of public organizations and identify the most critical ones. The following research questions were posed:

1. What are the risk factors for IS projects?
2. What are the risk factors for IS projects in public organizations?
3. Is there any significant influence between risk factors?

Research methods included a critical review of literature, the case study, practical collaboration and logical deduction. The following steps were taken.

The first step. The empirical evidence was searched aimed at peer-reviewed journal publications from 2010 to 2015. The process was supported by the use of electronic tools for the search and selection of publications. The search included journals indexed in bibliographic databases, i.e. Ebsco, ProQuest, Science Direct. The search was conducted using a relevant set of keywords and phrases such as ''software project'' or ''information technology project'' and ''risk management'' or "risk factors", and ''project success'' included in paper abstracts in all possible permutations and combinations (taking into consideration the logical AND, and OR as appropriate). A search was done on the appearance of any combination of these terms, with a result of 933 hits. All hits of 4 pages or less were excluded and narrowed to reviewed academic journals in English. Then, a second selection was made by evaluating the abstracts of the publications selected in the first round of selection. This second round, it was necessary to make sure that the publications included all three topics: software/IT project, project success, and project risk management. The search process resulted in a total of 13 journal publications, published between 2010 and 2015.

The second step. Risk factors for IS projects were improved on the basis of practical collaboration the authors with IT companies that develop IS systems for business and public organizations.

The third step. After careful evaluation of the literature findings, practical collaboration and logical deduction, risk factors for IS projects were further refined, classified and presented based on MIT90s framework. In the framework, the risk factors were considered in five groups as (1) project strategy, (2) structure, (3) individual and roles, (4) management process, and (5) technology.

The fourth step. Using the case studies approach, the risk factors for IS projects in public organizations in Poland were defined. Moreover, semi-structured interviews with end-users and project team members were conducted as well as shareable documentations related to IS projects management were analyzed during the study. Data was obtained from documents and records such as statement of work, project plan, risk management plan, minutes of meetings, review meetings, reports, project overview presentations and project closure reports. This study was conducted in 2010 and 2013. It concerned IS projects in two Polish public organizations. The IS projects included development and implementation of integrated IS.

The fifth step. The risk factors included in MIT90s framework were evaluated and further developed. The framework was supplemented by additional risk factors defined for RM of IS projects in the public organization.

## IV. Research Findings

### *A. Case Studies of IS Projects in Public Organizations*

Public organizations in Poland, due to territorial scope of their operations are divided into public organizations at the state level, embracing the whole Poland, and public organizations at the local levels, district or county. The described case studies of IS projects refer to the state level, where project management took place and the local levels, where IS was implemented.

Two similar projects, one successful and one not, will be used to present the application of risk factors in IS projects [43]. Information about a project was gathered by participation in those projects and conducting series of semi-structured interviews. Data was obtained from documents and records such as statement of work, project plan, risk management plan, minutes of meetings, review meetings, reports, project overview presentations and project closure reports.

Table III shows that those two projects were similar in terms of scope and size. As a result, the outcomes of the projects were different. Project A ended only as a partial success. Finally, IS was implemented but it was not fully used by the end-users after 12 months. The completion of Project A was also significantly delayed. Project B was fully successful. IS was implemented and it is fully used by its end-users.

Project A was carried by a public organization at the state level. The aim of the project was to improve and automate government processes and to implement an integrated information system, i.e. an ERP system in sixteen public organizations at the local levels. The ERP system used to this point of time was out of date. The results of change were to centralize management of the organizational structure of all sixteen public organizations and automation of supporting government processes for finance and accounting, human resources management, payroll management, inventory management, and fixed assets management. The expected benefits of the project were to eliminate unnecessary documentation, systemize document circulation, ensure a smooth flow of information, and make information accessible (which is relevant, timely to appropriate users and in an appropriate form). A specifically set up project team of the central public organization was responsible for the implementation of the ERP system. The project team was composed of people from the departments of the central public organization, such as: accounting, human resources, payroll, fixed assets, and inventory management, and from the IT department. Moreover, the project team was supported by the members of IT company, especially business analysts, systems analysts, and project team leaders.

Project A was managed using PRINCE2 methodology, however only few documents were created. There was created a risk procedure, however the risk was never escalated to steering committee. The risk registry was fulfilled at the beginning of the project, but was not updated during the project. The project team was not properly instructed about necessity of risk reporting. The basic risk management approach was missing. The risk was not properly managed. Often the risk was not identified but happened as an issue.

Project B was also carried out by a public organization. The aim of the project was to implement IS for supporting processes of service provision for citizens. As a result of the project the following types of IS were implemented: integration platform, business intelligence, enterprise portal, web based information portal and mobile terminal software. The project was undertaken as a consequence of the diagnosed problems arising from the lack of IT system integration. The lack of integration made it impossible to have quick access to information indispensable for effective functioning and monitoring of operations of public organizations and caused an ineffective flow of information between the public organizations and the cooperating institutions. The lack of system cooperation compounded the difficulties in monitoring funds allocation and expenditure, and the difficulties in monitoring the use of funds by individual public organizations.

Project B was managed using PRINCE2 methodology, where all necessary documents essential for effective project management were created. The project team was formally established. Particular people were permanently assigned to particular parts of the project. Their scope of responsibilities was explicitly defined. The project team consisted of an IT specialist group and a government group made up of specialists who were the main users of the system. Risk management was conducted concurrent with the project implementation. The end-users participated in a series of conferences, where a clearly defined project goal and

TABLE III.
PROJECT A AND PROJECT B – COMPARISON OF BASIC VARIABLES

| Features | Project A | Project B |
|---|---|---|
| Project type | Information system | Information system |
| Sector | Public organizations | Public organizations |
| Initial schedule | 12 months | 18 months |
| Budget | Realistic | Realistic |
| Success criteria | On time, within budget, successful installation of ERP system | On time, within budget, successful installation of web-based information system |
| IS software | Custom made | Custom made |
| Customers | Public organization employees | General public,<br>Public organizations employees |
| No of end users | 400 | 35 000 |
| Project management methodology | PRINCE2<br>(only few basic documents where created) | PRINCE2<br>(full documentation needed were created) |
| Risk management | No (no risk registry provided) | Yes (risk registry provided) |
| Project result after 12 months | Software was made but not fully used after 12 months | Software was made and fully used after 12 months |

successively accomplished tasks were presented. Moreover, they actively participated in analysis meetings where they defined the system requirements. The project had a coherently worked-out schedule that also included a business team meeting schedule. The business team was kept informed about the project progress and participated in the final IS testing.

### *B. Risk Management of IS Projects in Public Organizations*

In project A, the risk was not identified and managed. Whereas, RM was applied to project B in a methodologically correct manner.

Based on the examination of the case studies, the authors can draw the same observations (Table IV). Obviously, it can be stated that in case of project A, 27 risk factors did not occur, although 25 risk factors occurred and they were not managed. The lack of RM could have contributed to the failure of the IS project. Finally, IS was created and implemented with a significant delay. In case of project B, 38 risk factors did not occur, and 14 risk factors occurred and they were managed. Project B was completed on the schedule. It can be assumed that RM played a significant role in the IS project success.

However, there were several other risk factors which were not included in the risk factors framework for business organizations (Fig. 1). They were:

- changing government processes during project implementation;
- changing and inconsistent legal regulatory framework;
- challenges of procurement procedure;
- financial capability of project contractor; and
- managing contract.

**Changing government processes during project implementation**. Changes in government processes during the project always generate the need to change the IS requirements. The changes of IS requirements are one of the most frequent reasons of IS project failures. The change of requirements influences the scope of the project and its functionalities and can extend the project duration.

**Changing and inconsistent legal regulatory framework**. Changes to the rule of law which take place during the project can affect and often affect IS requirements. As it was mentioned above, the change to the requirements influence the scope of the project and its functionality and can extend the project duration. Unfortunately, the changes to Polish legal system are frequent. It is partially connected with the fact that recently the Polish economy has gone through the transition from a central planned economy to a market economy and it had to adjust and is still adjusting the legal system to the market economy.

**Challenges of procurement procedure**. There are several factors which must be met in a procurement procedure. One criterion of offer evaluation must be a price. Other criteria may be freely chosen depending on the object of the contract, e.g. quality, technical merit, functionality, usability. Typically, a tender is chosen using the price criterion. In Poland, the cheapest offer is often chosen. As a result the ratio of price to quality is not always maintained.

**Financial capability of project contractor.** The payment for the contractor for the works done within the IS project framework takes place after the final IS technical acceptance. In practice it may take from few to several months. During this time the contractor has to cover the running costs from own resources. This creates the risk of losing financial liquidity if the contractor does not have appropriate financial backing.

**Managing contract.** An effectively managed contract can impact on a timely completion of IS project. However, it is extremely difficult to predict all conditions that may occur during the contract realization process. There is a need for long term planning and considering, e.g. identifying all current and future systems that must be integrated. There is a high risk that some minor requirements might be omitted in the contract. The contract cannot be significantly changed

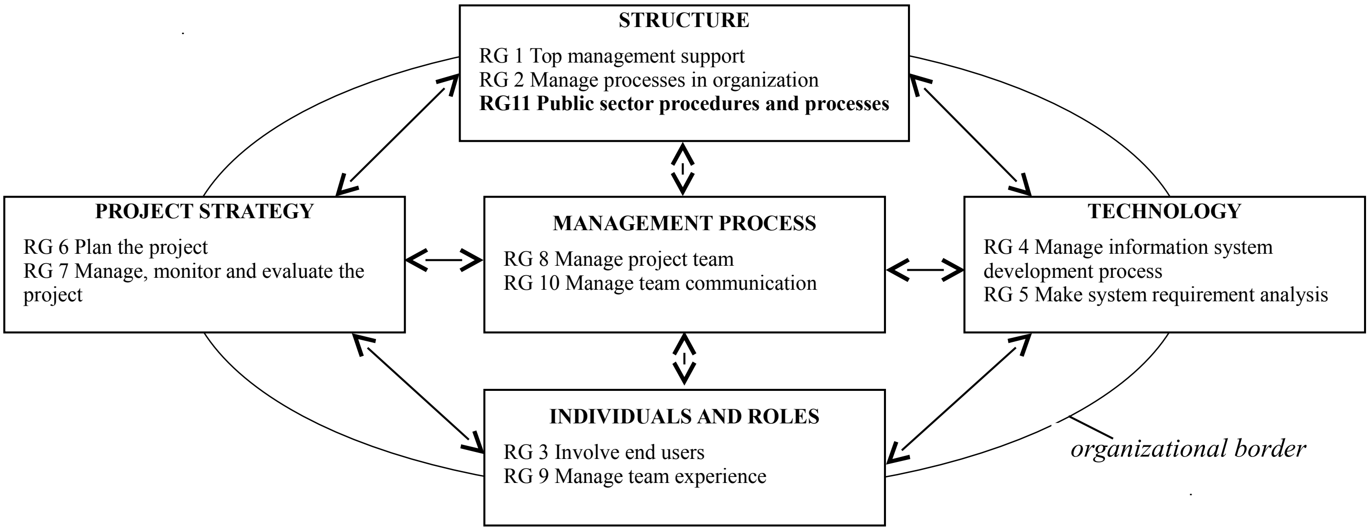

STRUCTURE
RG 1 Top management support
RG 2 Manage processes in organization
RG11 Public sector procedures and processes
PROJECT STRATEGY
RG 6 Plan the project
RG 7 Manage, monitor and evaluate the project
MANAGEMENT PROCESS
RG 8 Manage project team
RG 10 Manage team communication
TECHNOLOGY
RG 4 Manage information system development process
RG 5 Make system requirement analysis
INDIVIDUALS AND ROLES
RG 3 Involve end users
RG 9 Manage team experience
organizational border

Fig. 2. Risk factors framework for IS projects in public organizations

during the project, as it is one of the procurement procedures.

### C. Risk Factors Framework for IS Projects in Public Organizations

The above identified risk factors for IS projects in public organizations are creating the new risk factors group: public sector procedures and processes. This group is included in the risk factors framework proposed by Morton and presented in Fig. 1. The new group of risk factors was classified into the framework dimension named Structure. The factors are the part of this dimension, because they are related to processes and functions of IS projects in public organizations. Furthermore, they influence project strategy, management process and technology. The enhanced risk factors framework for IS projects in public organizations is presented in Fig. 2.

As mentioned above, public sector procedures and processes have an impact on project strategy. Especially, they affect project planning which has to account for such identified risk factors as a procurement procedure, changing government processes and financial capability of project contractor. Public sector procedures and processes also influence managing, monitoring and evaluating IS project as they are more complex and have to account for additional identified risk factors. Public sector procedures and processes have an impact on management process. Managing contracts, in particular, requires from managers specialist experience and the knowledge of the rule of law.

Public sector procedures and processes also influence technology. Changes to the requirements caused by changing government processes or a legal regulatory framework must be reflected in IS projects, especially in IS functionality. Moreover, these changes often result in delays in the IS project implementation. The potential changes to the IS project lead time are limited by the procurement procedure.

In conclusion, public organizations must take into account more risk factors in the risk management of IS projects than business organizations. Fig. 2 presents the framework of risk factors for IS projects in public organizations.

## V. Conclusion

Identifying and understanding risk factors is crucial for the success of IS projects in public organizations. The paper enhances the framework of risk factors identified in the literature and proposes a comprehensive risk factors framework for IS projects in public organizations.

This study contributes to the research on risk factors for IS projects in two ways. Firstly, the risk factors for IS projects in business organization are analyzed and presented. Secondly, the unique risk factors for IS projects in public organizations are identified based on the case studies approach. In summary, there are eleven groups of risk factors for IS projects in public organizations, namely (1) top management support; (2) manage processes in organization; (3) involve end users; (4) manage information system development process; (5) make system requirement analysis; (6) plan the project; (7) manage, monitor and evaluate the project; (8) manage project team; (9) manage team experience; (10) manage team communication (11) public sector procedures and processes. Moreover, the proposed risk factors framework for public organization is based on MIT90s framework. The framework indicates that risk factors are not standing alone, but they influence each other.

In this research, public organizations could find knowledge related to the risk factors impacting on successful IS projects. Especially, this research can be useful for the Central and Eastern European countries. This is because the countries are similar. Their similarity concerns their analogous geopolitical situation, their joint history, traditions, culture, and values. In addition, the similarity reflects in building democratic state structures and a free-market economy, participating in the European integration process, the levels of information systems implementation in public organizations. Moreover, they have to resolve the same problems and overcome the same political, economic, social, technological obstacles in their transition from traditional public organizations to organizations based on information systems.

As with many other studies, this study has its limitations. The main is that, it is only based on two case studies in Poland. Caution should be taken when generalizing our findings. The issues of risk factors for IS projects in public organizations, therefore, should be explored in greater depth. There is a need to examine other case studies, and verify and enhance the risk factors framework. This will be considered as a future work.

## Acknowledgment

This research has been supported by a grant entitled "Transformation of business and public administration by information technology and information systems" from the University of Economics in Katowice, Poland, 2014-2016.

## References

[1] M.O. Barros, C.M.L. Werner, and G.H. Travassos, "Supporting risks in software project management," *Journal of Systems and Software*, vol. 70(1–2), pp. 21–35, 2004.

[2] R.R. Nelson, "IT project management: Infamous failure, classic mistakes, and best practices," *MIS Quarterly Executive*, vol. 6(2), pp. 67–78, 2007.

[3] R.N. Charette, "Why software fails," *IEEE Spectrum,* vol 42(9), pp. 42–49, 2005.

[4] K. Bock and S. Trück, "Assessing uncertainty and risk in public sector investment projects," *Technology and Investment*, vol. 2, pp. 105–123, 2011.

[5] L. Zhou, A. Vasconcelos, and M. Nunes, "Supporting decision making in risk management through an evidence-based information systems project risk checklist," *Information Management & Computer Security*, vol. 16, no. 2, pp. 166–186, 2008.

[6] P.K. Dey, J. Kinch, and S.O. Ogunlana, "Managing risk in software development projects: A case study," *Industrial Management end Data Systems*, vol. 107(2), pp. 284–303, 2007.

[7] J.J. Jiang , G. Klein, and R. Discenza, "Information system success as impacted by risks and development strategies," *IEEE. Transactions on Engineering Management,* vol. 48 (1), pp. 46 –55, 2001.

[8] A. Jani, "Escalation of commitment in troubled IT projects: Influence of project risk factors and self-efficacy on the perception of risk and

TABLE IV
RISK FACTORS EVALUATION FOR PROJECT A AND PROJECT B

| | **Risk factor** | **Project A** | **Project B** |
|---|---|---|---|
| R01 | Lack of top management commitment to the project | occurred, not identified, not managed | not occurred |
| R02 | Resources shifted away from the project because of changes in organizational priorities | not occurred | not occurred |
| R03 | Major effect of project implementation on organizational structure | not occurred | not occurred |
| R04 | Top managers make important IT decisions without consulting the others | occurred, not identified, not managed | not occurred |
| R05 | Unrealistic projects outcomes | not occurred | not occurred |
| R06 | Change in ownership or senior management during the process of development | not occurred | not occurred |
| R07 | Excessive project size | not occurred | not occurred |
| R08 | Mismatch between organization culture and required business process changes needed for new system | not occurred | not occurred |
| R09 | Changes in organizational priorities | not occurred | not occurred |
| R10 | Continuous changes in the organizational environment | not occurred | not occurred |
| R11 | Time too short/too long | occurred, not identified, not managed | occurred, identified, managed |
| R12 | Unrealistic schedule | occurred, identified, not managed | occurred, identified, managed |
| R13 | Lack of user participation | occurred, identified, not managed | not occurred |
| R14 | Users resistant to change | occurred, identified, not managed | occurred, identified, managed |
| R15 | Target users are unfamiliar with the technology and require additional training | occurred, identified, not managed | occurred, identified, managed |
| R16 | Users with negative attitudes toward the project | occurred, not identified, not managed | not occurred |
| R17 | User is not committed to the project | occurred, identified, not managed | not occurred |
| R18 | Users constantly request further changes | occurred, identified, not managed | occurred, identified, managed |
| R19 | Conflicts between users departments | occurred, identified, not managed | occurred, identified, managed |
| R20 | High level of technical complexity | not occurred | occurred, identified, managed |
| R21 | Immature technology | not occurred | not occurred |
| R22 | New technology and use of technology that had not been used in prior projects | not occurred | not occurred |
| R23 | Lack of effective development methodology | not occurred | not occurred |
| R24 | Large number of links to other system required | occurred, identified, managed | occurred, identified, managed |
| R25 | Inadequate system documentation; incomplete or non-existent. | not occurred | not occurred |
| R26 | Lack of proper tests | not occurred | not occurred |
| R27 | Lack of integration between systems | occurred, identified, managed | occurred, identified, managed |
| R28 | Continually changing scope and system requirements | occurred, identified, not managed | occurred, identified, not managed |
| R29 | Unclear or incomplete system requirements | occurred, identified, not managed | not occurred |
| R30 | System requirements not adequately identified | occurred, identified, not managed | not occurred |
| R31 | Conflicting system requirements | not occurred | not occurred |
| R32 | Failure to manage end-user expectations | occurred, identified, managed | occurred, identified, managed |
| R33 | Lack of frozen requirements | not occurred | not occurred |
| R34 | Poor project planning | occurred, not identified, not managed | not occurred |
| R35 | Project progress not monitored closely enough | occurred, not identified, not managed | not occurred |
| R36 | Lack of an effective project management methodology | occurred, not identified, not managed | not occurred |
| R37 | Ineffective communication | not occurred | occurred, identified, managed |
| R38 | Inadequate estimation of required resources | not occurred | not occurred |
| R39 | Inexperienced project manager | not occurred | not occurred |
| R40 | Project manager lacks required skills | not occurred | not occurred |
| R41 | Critical activities are not identified | occurred, not identified, not managed | not occurred |
| R42 | Lack of knowledge management | not occurred | not occurred |
| R43 | Frequent turnover within the development team | not occurred | not occurred |
| R44 | Team members lack of specialized skills required by the project | occurred, identified, not managed | occurred, identified, managed |
| R45 | Inadequately trained development team members | not occurred | not occurred |
| R46 | Team members are unmotivated | occurred, identified, not managed | not occurred |
| R47 | Inadequate composition of project team | not occurred | not occurred |
| R48 | Team members are unfamiliar with the technology | occurred, not identified, not managed | not occurred |
| R49 | Conflict and no cooperation between the team members | occurred, not identified, not managed | occurred, identified, managed |
| R50 | Improper definition of roles and responsibilities | not occurred | not occurred |
| R51 | Team member are in many localizations | not occurred | not occurred |
| R52 | Inadequate team size | not occurred | not occurred |

the commitment to a failing project," *International Journal of Project Management*, vol. 29, pp. 934–945, 2011.

[9] M. Keil, L. Wallace, D. Turk, G. Dixon-Randall, and U. Nulden, "An investigation of risk perception and risk propensity on the decision to continue a software development project" *The Journal of Systems and Software*, vol. 53(2), pp. 145–157, 2000.

[10] D. Tesch, T.J. Kloppenborg, and M.N. Frolick, "IT project risk factors: The project management professional's perspective," *Journal of Computer Information Systems,* vol. 47(4), pp. 61–69, 2007.

[11] P. Patanakul, "Managing large-scale IS/IT projects in the public sector: Problems and causes leading to poor performance," *Journal of High Technology Management Research*, vol. 25, pp. 21–35, 2007.

[12] B. Aritua, N.J. Smith, and D. Bower, “What risks are common to or amplified in programmes: Evidence from UK public sector infrastructure schemes,” *International Journal of Project Management*, vol. 29, pp. 303–312, 2011.

[13] B. Whittaker, "What went wrong? Unsuccessful information technology projects," *Information Management & Computer Security*, vol. 7(1), pp. 23–30, 1999.

[14] D. Baccarini, G. Salm, and P.E.D. Love, "Management of risks in information technology projects," *Industrial Management & Data Systems*, vol. 104 iss: 4, pp. 286–295, 2004.

[15] J. Wateridge, “IT projects: A basis for success,” *International Journal of Project Management*, vol. 13, issue 3, pp. 169–172, June 1995.

[16] M. Yildiz, “E-government research: Reviewing the literature, limitations, and ways forward,” *Government Information Quarterly,* vol. 24, issue 3, pp. 646–665, July 2007.

[17] R. Kliem and I. Ludin, “Reducing project risk,” *Gower Publishing Limited,* Aldershot, 2000.

[18] J. Cadle and D. Yeate, *Project Management for Information Systems*, Financial Times/Prentice-Hall, Harlow, 2001.

[19] O. Zwikael, R.D. Pathak, G. Singh, and S. Ahmed, “The moderating effect of risk on the relationship between planning and success,” *International Journal of Project Management*, vol. 32, pp. 435–441, 2014.

[20] R. McGaughey Jr, Ch.A. Snyder., and H.H. Carr, “Implementing information technology for competitive advantage,” *Risk Management issues*, pp. 273-280, 1994.

[21] K. Jugdev, D. Perkins, J. Fortune, D. White, and D. Walker, “An exploratory study of project success with tools, software and methods,” *International Journal of Managing Projects in Business*, vol. 6, no. 3, pp. 534–551, 2013.

[22] R. Rabechini J. and M.M. de Carvalho, “Understanding the impact of project risk management on project performance: An empirical study”, *Journal of Technology Management & Innovation*, vol. 8, pp. 64–78, 2013.

[23] T. Raz and E. Michael, “Use and benefit of tools for project risk management,” *International Journal of Project Management*, vol. 19(1), pp. 9–17, 2001.

[24] A. Nalewaik, “Risk management for pharmaceutical project schedules,” *AACE International Transactions. Risk*, vol. 07, pp. 71–75, 2005.

[25] T.K. Das and B.S. Teng, “Managing risks in strategic alliances,” *The Academy of Management Executive*, vol. 13(4), pp. 50–62, 1999.

[26] P. Cook, “Formalized risk management: vital tool for project- and business-success,” *Cost Engineering*, vol. 47(8), pp. 12–13, 2005.

[27] V. Holzmann and I. Spiegler, “Developing risk breakdown structure for information technology organizations,” *International Journal of Project Management*, vol. 29, pp. 537–546, 2011.

[28] M. Keil, P. Cule, K. Lyytinen, and R. Schmidt, “A framework for identifying software project risks,” *Communication of the ACM*, vol. 41, no. 11, pp. 76–83, 1998.

[29] C. López, J. and L. Salmeron, “Risks response strategies for supporting practitioners decision-making in software projects,” *Procedia Technology*, vol. 5, pp. 437–444, 2012.

[30] S. Liu and L. Wang, “Understanding the impact of risks on performance in internal and outsourced information technology projects: The role of strategic importance”, *International Journal of Project Management,* vol. 32, pp. 1494–1510, 2014.

[31] L. Jun, W. Qiuzhen, and M. Qingguo, “The effects of project uncertainty and risk management on IS development project performance: A vendor perspective,” *International Journal of Project Management*, vol. 29, pp. 923–933, 2011.

[32] P.K. Dey, B.T. Clegg, and D.J. Bennett, “Managing enterprise resource planning projects,” *Business Process Management Journal,* vol. 16, no. 2, pp. 282–296, 2010.

[33] S. Sundararajan, M. Bhasi, and P.K. Vijayaraghavan, “Case study on risk management practice in large offshore-outsourced Agile software projects,” *IET Software*, doi: 10.1049/iet-sen.2013.0190.

[34] J.M. Denolf, J.H. Trienekens, P.M. Wognum, J.G. A. J. Vander Vorst, and S.W.F. (Onno), Omta, “Towards a framework of critical success factors for implementing supply chain information systems,” *Computers in Industry*, vol. 68, pp. 16–26, 2015.

[35] M.S. Scott Morton, *The corporation of the 1990s: Information technology and organizational transformation*, Oxford University Press, New York, 1991.

[36] M.L. Markus and C. Tanis, *The enterprise systems experience–from adoption to success*, in: R.W. Zmud (Ed.), Framing the domains of IT research: Glimpsing the future through the past, Pinnaflex Educational Resources, Inc., Cincinnatti, OH, pp. 173–207, 2000.

[37] C.P. Holland and B. Light, “A critical success factors model for ERP implementation,” *IEEE Software*, vol. 16(3), pp. 30–36, 1999.

[38] T. Bemelmans, *Administrative Information Systems and Automation,* Kluwer Bedrijfswetenschappen, Deventer, The Netherlands, 1998.

[39] W.J. Orlikowski, “The duality of technology: rethinking the concept of technology in organizations,” *Organization Science,* vol. 3(3) pp. 398–427, 1992.

[40] T. H. Davenport, *Mission critical: Realizing the promise of enterprise systems*, Harvard Business Press, Boston, MA, 2000.

[41] N.F. Doherty and M. King, “From technical to socio-technical change: tackling the human and organisational aspects of systems development projects,” *European Journal of Information Systems*, vol. 14(1) pp. 1–5, 2005.

[42] M.J. Verdecho, J.J. Alfaro-Saiz, R. Rodriguez-Rodriguez, and A. Ortiz-Bas, “A multi-criteria approach for managing inter-enterprise collaborative relationships,” *Omega,* vol. 40 (3), pp. 249–263, 2012.

[43] E. Ziemba and I. Obłąk, “Change management in information systems projects for public organizations in Poland,” *Interdisciplinary Journal of Information, Knowledge, and Management,* vol. 10, pp. 47–62, 2015.

DOI: 10.15439/2015F297
ACSIS, Vol. 5

# Integration of Domain Ontologies in the Repository of Website Evaluation Methods

Paweł Ziemba[1] Jarosław Jankowski[2,3] Jarosław Wątróbski[2] Waldemar Wolski[4] Jarosław Becker[1]

[1] The Jacob of Paradyż University of Applied Sciences in Gorzów Wielkopolski, Chopina 52, 66-400 Gorzów Wielkopolski, Poland
Email: {pziemba, jbecker}@pwsz.pl
[2] West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland
Email: {jjankowski, jwatrobski}@wi.zut.edu.pl
[3] Department of Computational Intelligence, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
[4] University of Szczecin, Mickiewicza 64, 71-101 Szczecin, Poland
Email: wwolski@uoo.univ.szczecin.pl

***Abstract*—Many methods can be used for the evaluation of website quality. While they can be used for different purposes and require different assessment approaches it is not easy to select proper method adequate to the needs. Presented research is focused on building a repository of knowledge about the methods for the assessing the quality of website. The repository in the form of ontologies covers various methods of quality assessment and makes possible their proper selection. Proposed approach was verified with main methods and the resulting ontology can act as a repository of domain knowledge.**

## I. Introduction

In the sectors related to e-commerce and online advertising, the number of users visiting corporate websites, blogs, portals and social platforms directly determines revenue. More users increase the potential of advertising, which has a direct impact on the number of transactions, the amount of revenue, and the ability to engage new customers [1]. In the United States, recent revenues from online advertising amounted to 42.78 billion dollars [2], while in Europe, it amounted to 27.3 billion euros in the advertising sector [3] and 363,1 billion euros in e-commerce [4]. It is worth noting that for businesses using a website to generate transactions, the website's quality can have a major impact on sales [5]. Poor quality of service and user experience can cause existing Internet customers [5], potential sales, and repeated visits to be lost [6]. Therefore, in order to maximize profits from e-commerce or online advertising, website owners should only offer the highest quality experience and services.

The quality of a website can be understood as the qualification of how well it meets the needs of users [7]. It should be noted that quality is defined by a model composed of characteristics and features/criteria describing its various components [8]. In the literature, there are many methods used to assess the quality of Internet websites, with the most formalized of those including: eQual [9], Ahn [10] SiteQual [11], Web Portal Site Quality [12] and Website Evaluation Questionnaire [13]. They have been widely used in both academic work [14] and business practice [15]. By examining the different methods was observed similarity between them. Characteristics and criteria for quality assessment based on a Likert scale were used and for the getting knowledge about quality of services surveys were explored. The reason for these similarities is that different methods are often based on the same scientific theories and source references.

Analysis of the literature and areas of practical use of methods and models for the assessment of website quality indicates a gap in the area of knowledge repository construction. Namely, there is no repository of knowledge covering some important methods for assessing the website quality. The possible construction of such a repository in the form of ontology allows formal specification and analysis of the various methods of assessment and factors affecting the quality of website [24], as well as consequent sharing and reusing that domain knowledge [17]. The ontological form provides access to the knowledge gained from individual assessment methods and enables the processing of this knowledge. It is also important to enable the integration of heterogeneous data [18] from a variety of assessment methods. As a result, it will also be possible to assess websites through a variety of methods defined in the ontology and compare individual results of that assessment in a single terminology and a reference plane. In this approach, the problem of building such a repository is divided into two subproblems. They successively concern (1) the construction of the source ontologies, reflecting the

978-83-60810-66-8/$25.00©2015, IEEE

specific methods for quality assessment; (2) the integration of source ontologies in the target ontology, which is a complete repository. This approach is justified by the above-mentioned similarities between the different methods, so that individual source ontologies are close to each other in terms of structure and concepts. Proposed approach of building a repository of knowledge is graphically represented in Figure 1.

This article presents guidelines on how to design an ontology reflecting the various quality assessment methods and their integration in the repository of knowledge. Then, using these assumptions new algorithm was proposed based on the integration of the five methods in the repository. The article concludes with a presentation of research findings and possible future directions for work.

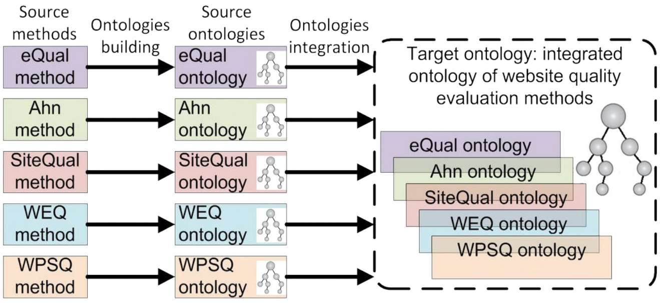


Fig 1. Proposed approach to building repository of knowledge by integration of ontologies

## II. Literature review

The possibility of constructing a repository of knowledge on the methods and models of assessing the quality of Internet services in the form of ontologies confirms its definition, as the term "ontology" in computer science is defined as the "formal specification of conceptualization" [19]. Essentially, it allows concepts and domain knowledge to be captured. A similar definition states that ontology is treated as a form of data structure and a tool for data representation, allowing knowledge to be shared and reused in artificial intelligence systems that use a common vocabulary [20]. Therefore, ontology seems to be a natural form of representation for the repository of knowledge concerning methods of quality assessment. This is due to the fact that the use of ontologies will create conceptual models that explain the structure of the different methods of evaluation criteria. The use of ontologies will also be shared, making it possible to repeat the use of such structures and facilitate better management.

The possibility of using ontologies as a repository of knowledge is confirmed by various studies where among others a biomedical knowledge base was created using ontologies [21]. Ontologies are also implemented in the knowledge bases of various systems, e.g. an expert system for the study of company financial ratios [22] and a decision support system for the construction of railway infrastructure [23]. Analysis of the literature shows that ontologies are also used in the systems and methods of quality assessment. For example, [24] presents an ontology quality that formalizes the knowledge necessary to evaluate the quality of e-government. In contrast, [25] proposes the use of ontologies in the quality assessment of tourism services websites. In this case, the ontology serves as a repository of knowledge on issues related to the tourism sector. However, analysis of the literature showed that there is no repository of knowledge covering some important methods for assessing the quality of website. Meanwhile, the reasons discussed above and the use of ontologies both demonstrate that the construction of a repository of knowledge about the methods of assessing website quality based on ontologies is justified.

The ontology construction methodology is frequently used, although it differs in the degree of formalization, destination and detail [26]-[30]. The most formal and detailed methodology includes Methontology [31] and NeOn [32]. Methontology proposed an "ontology life cycle" based on the so-called "evolving prototypes" [29]. This approach assumes that the individual steps of this methodology, namely I specification, II conceptualization, III formalization, IV implementation, and V maintenance, are performed continuously, and each cycle is replaced by a new version of the ontology. The NeOn methodology is focused on developing an ontology network instead of a single ontology. An ontology network is understood here as a set of related ontologies through a variety of compounds, e.g. mapping and versioning [26]. Methontology defines the process of conceptualization in detail, while NeOn largely formalizes the problem of ontology specification.

Concerning the integration of ontologies, it should be noted that the concept of integration is not clearly defined in different works. Reference [33] distinguishes three types of ontology integration: integration of ontologies into applications, integration and reusing, and integration by merging. Integration and reusing is defined as the construction of a new ontology by using existing ontologies. Merging is the unification of multiple ontologies from a given field into a new ontology. Merged ontologies may vary in appearance and function, including taxonomy of concepts, method of implementation, etc. Conflicts arising from these differences should be resolved when merging ontologies [33]. Reference [34] stated that ontology created by such a merger should capture all the knowledge contained in the original ontologies. Two approaches may be used: a new ontology can be created that reflects the source ontologies, or what is known as bridge ontology can be created. Bridge ontology contains the original ontologies and the relationship between them. In [35] integrating and merging operations are defined as equal. They include the creation of a new ontology from two or more existing ontologies with overlapping parts. In addition, reference [34] defines the terms being similar to integration, i.e. alignment and mapping. Alignment is a preliminary process by which it is possible to integrate ontology in a widely understandable way. It is a process of discovering similarities between ontologies. These similarities may occur between the concepts, their instances, or a similarity in the ontology structure. Mapping, on the other hand, refers to representing the relationships that exist between ontologies. As result a specification of the semantics of one ontology coverage by another is obtained which can be represented as a map describing the mutual relations between the elements of mapped ontology). Relationships between ontologies are stored separately from the ontology and are not part of them. Therefore, the mapping does not change in any of the ontology. Mapping allows to get a result somewhat similar to bridge ontology. However, in bridge ontology, as opposed to mapping, source ontologies and connections between them are stored in one entity.

## III. Construction of the source ontologies

In order to successfully complete integration, it is necessary to construct an appropriate source ontologies, reflecting the various website quality assessment methods. Consequently, the design of the source ontology is affected by the structure of the quality models included in the assessment methods and aspects relateted to their future integration. As indicated in Paragraph 1, each of the methods of evaluation is based on the quality model composed of characteristics and features/criteria describing its various components. Each of the source ontologies should reflect the quality model, which can be represented by a general structure, as illustrated in Figure 2.

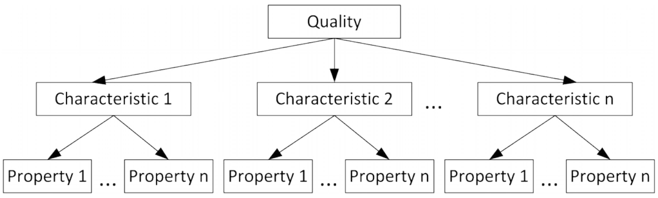


Fig 2. General quality model

With regard to the structure of the ontology generated by integration, it is assumed that the target (integrated) ontology should contain a unified model, including all the characteristics and evaluation criteria, and simultaneously contain therein quality models derived from different methods. The purpose of storing individual quality assessment models (source ontologies) in the target ontology is to determine which method and characteristics derives a given criterion. Ability to determine the source method of a given criterion allows the integration of heterogeneous data from a variety of assessment methods into knowledge base. The inclusion in the ontology of a unified model allows to use it as a knowledge repository on evaluation criteria. Therefore, the ancestor of each criterion should be a model from which the criterion is derived together with the characteristics. In addition, if certain criteria are found in many different models of assessment contained in the ontology, then it will take more than two direct ancestors. This problem is illustrated in Figure 3.

Figure 3 shows part of the quality characteristics contained in Ahn and eQual methods. Moreover Figure 3 also includes a hypothetical model, which is a result of the unification of the ontologies of these methods. Three further examples of criteria are included, of which "availability" is derived from the Ahn method, "ease_of_learn_to_operate" comes from the eQual method, and "ease_of_navigation" occurs in both methods. Given that a unified model for the assessment will include the union of assessment models included in the methods eQual and Ahn, it should also include the criteria found in these models. Therefore, ancestors of the criterion of "availability" should be characteristics included in the Ahn and unified model. The criterion "ease_of_learn_to_operate" should be in the form of ancestral characteristics in eQual model and the unified model. In contrast, the ancestors of the criterion “ease_of_navigation” should be the characteristics of each of the three models. However, such a link between criteria with multiple characteristics using subsumption relations should not occur, as this would mean redundancy. A redundant

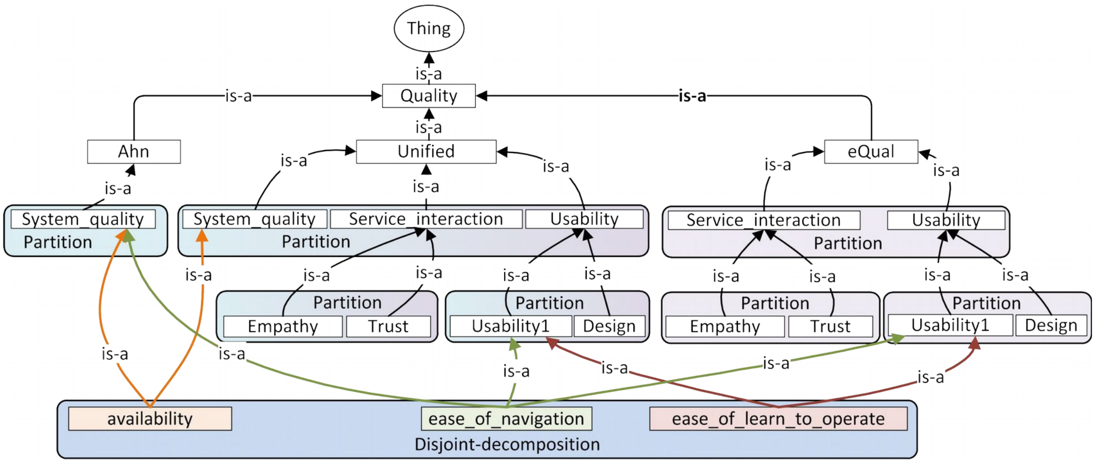


Fig 3. The problem of criteria membership for the evaluation models included in the target ontology

occurrence of a subsumption relationship is in turn treated as an error in the method of integration formulated in [35]-[38].

The problem presented above can be solved through the method of use adopted by the authors of the ontology. Namely, at the stage of the source ontology specification, it was found that the constructed ontology criteria would be separated from the characteristics. This approach allowed for applications with greater transparency of the ontology. This is also consistent with the representation of open and closed world assumptions in the ontology [39]. Namely, the each quality model contained in the various assessment models is a closed model. This means that it is complete, and a new one cannot be added to it. The quality evaluation criteria are the open portion of the world, which means that there may be additional criteria that are not yet included in the ontology. For these reasons, the part of ontology reflecting the quality model was used to partition taxonomic relationships between concepts occurring at the same level of hierarchy. In contrast, disjoint-decomposition taxonomic relationship was among the criteria used, as these criteria describe the different parts' quality, but there may be other criteria for describing the quality. This solution allows for independent criteria of quality models in certain ways. In addition, it allows for inclusion in the ontology of many quality models, including the unified model. Therefore, the hierarchy of concepts of a target ontology should be similar to structure shown in Figure 4, which is the result of the integration of ontologies eQual and Ahn. Figure 4 shows a hypothetical ontology with three quality models eQual, Ahn and model obtained as a result of their unification. Furthermore, in the illustrated ontology, regardless of the quality models the criteria belonging to these models were incporporated.

To get a similar structure as that shown in Figure 4, a compromise between a fully integrated ontology and so called bridge ontology was chosen [34]. Namely, the target ontology should include a fully unified model of quality, including the characteristics derived from the source ontologies, along with the methods of assessment, for which the source ontologies were built. Additional models, such as eQual and Ahn, as well as quality characteristics, should be included in the target ontology in the form of components, which can be simply described as external. This will avoid any inconsistencies, where the same quality characteristics will occur in the different models. An example may be the characteristics of "Information_quality" appearing in the eQual and Ahn models. As a result of the full integration, only one such characteristic appears in the resulting ontology, and it would be included in the unified model and the eQual and Ahn models. This would include criteria that belonged to both eQual and Ahn model. Consequently, in each of these models, union of information quality criteria has existed. These models would not be as consistent with their actual structure, as defined in the relevant methods of quality assessment. The solution of this problem, which allows for the incorporation of quality models (eg. eQual and Ahn) as "external", but simultaneously makes it possible to include them in the target ontology, is go find a use for them other than for IRI identifiers' unified ontology.

Separating the criteria of the quality characteristics of ontology concept hierarchies' source requires that the reasoner infers membership criteria for the relevant characteristics of the subsumption relation linking them. For this purpose, the criteria should be linked to the characteristics using the appropriate relationship, as shown in Figure 5.

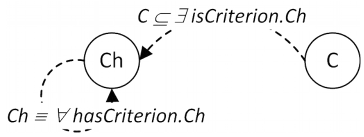


Fig 5. Proper configuration of relation “isCriterion” and “hasCriterion”

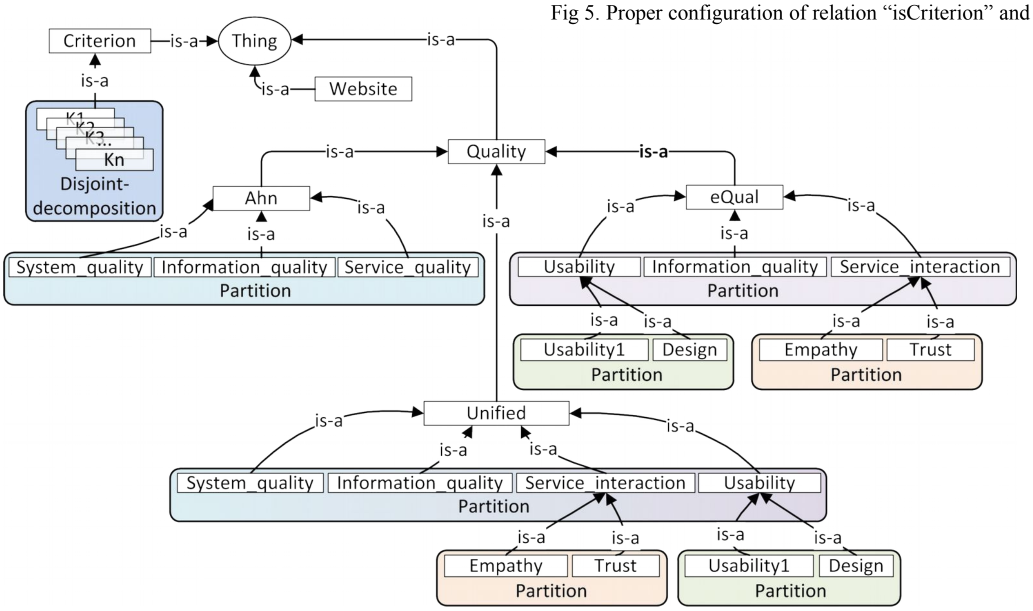


Fig 4. Hypothetical effect of eQual and Ahn integration in the target ontology

The diagram shown in Figure 5 contains concepts Ch and C, where Ch represents any quality characteristics, and C defines any criterion. By analyzing Figure 5, it can be seen that the relation "hasCriterion" with a universal quantifier is established as a necessary and sufficient condition (≡) Ch concept, which is also the domain and scope of this relationship. The relationship "isCriterion" of the existential quantifier is a necessary condition (⊆) criterion C, for which coverage is a quality characteristic Ch. This solution can be understood as follows: (a) there are certain criteria in the C group that belong to the characteristics of the Ch group ("C isCriterion some Ch"), (b) the characteristics of Ch are only those criteria that belong to the characteristic ("Ch hasCriterion only Ch"). This configuration relationship between "isCriterion" and "hasCriterion" allows for the exploration of membership criteria to the characteristics and describe this kind of membership as a comprising specific criteria in the relevant characteristics. Presented in the form indicated, the relationship is very important, as our study showed that a different configuration could lead to erroneous conclusions inferred by the reasoner. An example of such a configuration error in the indicated relationships is included in Figure 6.

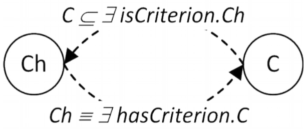


Fig 6. Wrong configuration of relation “isCriterion” and “hasCriterion”

Figure 6 shows a configuration in which the relationship "has Criterion” is marked as a necessary and sufficient condition (≡) with an existential quantifier, and the scope of this relationship is criterion C. This formulation relationship "hasCriterion" could cause inconsistent ontology at the level of the relationship. This will be explained by the example provided in Figure 7. In addition, will be used to clarify the description of the reasoning process, saved using descriptive logic expressions (1) - (12).

$$\text{Usability} \equiv \text{Design} \cup \text{Usability1} \tag{1}$$

$$\text{Design} \equiv \neg \text{Usability1} \tag{2}$$

$$\text{Design} \equiv \exists \text{hasCriterion.appropriateness\_design} \tag{3}$$

$$\text{Usability1} \equiv \exists \text{hasCriterion.ease\_of\_navigation} \tag{4}$$

$$\text{System\_quality} \equiv \exists \text{hasCriterion.ease\_of\_navigation} \tag{5}$$

$$\text{System\_quality} \equiv \exists\ \text{hasCriterion.appropriateness\_design\_style} \tag{6}$$

$$\text{appropriateness\_design} \equiv \text{appropriateness\_design\_style} \tag{7}$$

$$\text{from (4), (5)} \qquad \text{Usability1} \equiv \text{System\_quality} \tag{8}$$

$$\text{from (7)} \qquad \begin{array}{c} \exists\ \text{hasCriterion.appropriateness\_design} \ldots \\ \ldots \equiv \exists\ \text{hasCriterion.appropriateness\_design\_style} \end{array} \tag{9}$$

$$\text{from (9),(3),(6)} \qquad \text{Design} \equiv \text{System\_quality} \tag{10}$$

$$\text{from(8),(10)} \qquad \text{Design} \equiv \text{System\_quality} \equiv \text{Usability1} \tag{11}$$

$$\text{from(2),(11)} \quad (\text{Design} \equiv \text{Usability1}) \wedge (\text{Design} \equiv \neg\ \text{Usability1}) \equiv \emptyset \tag{12}$$

Figure 7 shows a portion of the target ontology, which is a hypothetical result of the integration of eQual and Ahn. The contents of Figure 7 include the quality characteristics eQual, i.e. "Design" and "Usability1", which make up the partition (i.e. they are disjoint and fully fill the space of eQual concept), which describes the axioms (1) and (2). In addition, Figure 7 contains the concepts of "ease_of_navigation", "appropriateness_design" and "appropriateness_design_style", which are some of the criteria. The criterion of "ease_of_navigation" is used in the Ahn method in the characteristics of "System_quality" and the method eQual characteristics "Usability1". The criterion of "appropriateness_design" occurs in the characteristics of the "Design" eQual method and an

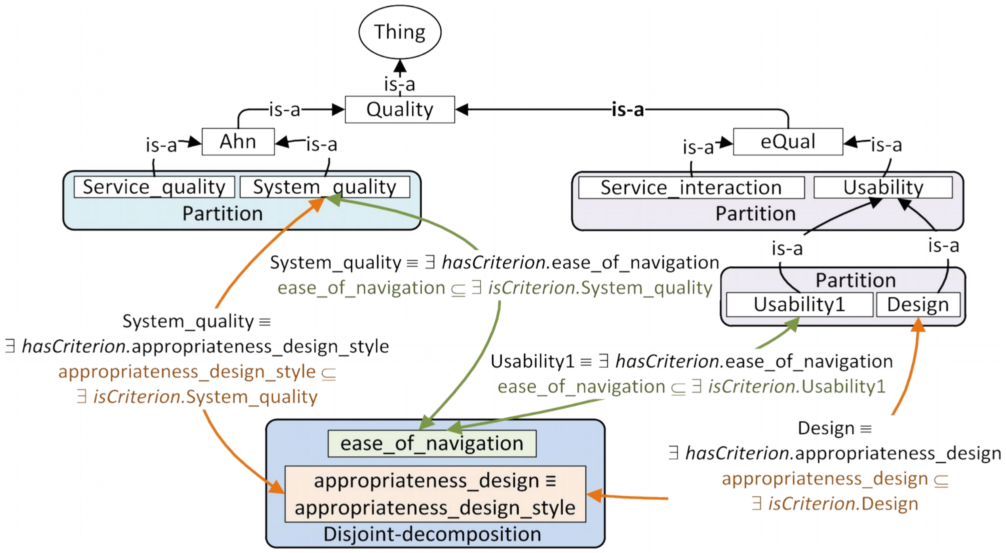

Thing
is-a
Quality
is-a
is-a
Ahn
is-a
is-a
Service_quality
System_quality
Partition
eQual
is-a
is-a
Service_interaction
Usability
Partition
is-a
is-a
Partition
Usability1
Design
System_quality ≡ ∃ hasCriterion.ease_of_navigation
ease_of_navigation ⊆ ∃ isCriterion.System_quality
System_quality ≡
∃ hasCriterion.appropriateness_design_style
appropriateness_design_style ⊆
∃ isCriterion.System_quality
Usability1 ≡ ∃ hasCriterion.ease_of_navigation
ease_of_navigation ⊆ ∃ isCriterion.Usability1
Design ≡
∃ hasCriterion.appropriateness_design
appropriateness_design ⊆
∃ isCriterion.Design
ease_of_navigation
appropriateness_design ≡
appropriateness_design_style
Disjoint-decomposition

Fig 7. Relationships between criterion and characteristics which raise the inconsistency inference

"appropriateness_design_style" functions in the characteristics of "System_quality" Ahn method. In addition, the concepts of "appropriateness_design" and "appropriateness_design_style" were considered to be equivalent to (7). Therefore, the concepts relating to design in an integrated ontology will be linked to the relationship "isCriterion" both with the concept of "System_quality" as well as "Design". The concept of "ease_of_navigation" in an integrated ontology will be linked to the relationship "isCriterion", the concepts "Usability1" and "System_quality". Assuming that:

- the concepts "Design" and "System_quality" are necessary and sufficient conditions include relations "hasCriterion some appropriateness_design" and "hasCriterion some appropriateness_design_style" (expressions (3) and (6)),
- then the concepts "Usability1" and "System_quality", as necessary and sufficient conditions, would cover the relationship "hasCriterion some ease_of_navigation" (expressions (4) and (5)),

during the integration of ontology, inconsistency would arise if the reasoner effect and "hasCriterion" relationships are defined as the necessary and sufficient conditions. If, for example the concept of "System_quality" has the necessary and sufficient condition "hasCriterion some ease_of_navigation", then it shall be deemed to be equivalent to the reasoner, the other concepts having the same condition. Therefore, it will be deemed to be equivalent to the concept of "Usability1" (8). On the other hand, based on a necessary and sufficient condition, "hasCriterion some appropriateness_design_style" shall be deemed equivalent to the concept of "Design", which has an equivalent condition "hasCriterion some appropriateness_design" (9), (10). The situation will appear when the concept of "System_quality" is equivalent to the concepts "Usability1" and "Design" (11). Accordingly, the reasoner also recognizes that these concepts are equivalent (12). The fact that the concepts of "Usability1" and "Design" create a partition is important, so that they are mutually exclusive and can not be considered equivalent to (12). This raises the inconsistency inference. The example given there is that there is no inconsistency in the target ontology due to the fact that the relationship "hasCriterion" and "isCriterion" between concepts of characteristics and quality criteria are defined as shown in Figure 5.

During integration there may be other inconsistencies that must be addressed on a regular basis. This will be discussed using the example of concepts called "Criterion" coming from the ontology eQual and SiteQual. In view of the fact that the method eQual grading scale covers a range of values from 1 to 7, a restriction should be imposed on the concept of "Criterion" in the ontology eQual " hasEvaluationValue only integer [> = 1, <= 7]". Meanwhile, in the SiteQual method the evaluation of criteria includes the values from 1 to 9, so that the concept of "Criterion" in the ontology describing this method will be limited to "hasEvaluationValue only integer [> = 1, <= 9]". While the integration of ontologies these concepts should be unified, the question arises: "What range of values should include restricting the concept of "Criterion" in a unified model of the target ontology?". In order to solve the conflicts mentioned, we decided to adopt a solution from the PROMPT algorithm. It occurs when one of ontology is considered to be the preferred one and conflicts of this type are resolved in its favor [40]. During the integration, target ontology that unifying source ontologies will be the preferred. Therefore, the order of the integrated source ontology will have a partial effect on the resulting form, a unified model of the target ontology.

Based on these observations, conclusions and guidelines, the source ontologies was built, reflecting the website evaluation quality methods. These ontologies, in the form of primary evidence, and inferred by the reasoner, are presented in: eQual [41], Ahn [42] SiteQual [43] Website Evaluation Questionnaire [44] and Web Portal Site Quality [45]. More of the construction process of source ontologies is described in [46]. Based on the literature analysis presented earlier, feedback and observations, we formulated the integration algorithm that was then applied in the process of ontology integration of website quality evaluation methods.

## IV. Source ontologies integration algorithm in the target ontology

The developed ontology integration algorithm is closely related to refactoring and ontology merging support tools offered by Protégé editor [47]. These tools make it much easier to carry out the integration process. The integration algorithm is divided into three parts. Part I is performed only once as part of the integration process. It concerns the preparation for the integration and the integration of the first source ontology in the target ontology. Part II includes the integration of the criteria and the placement in the target ontology the individual models of evaluation coming from the integrated ontology, as with the methods that are reflected by the different source ontologies. Part III deals with the construction of a unified quality model in the target ontology.

Part I: Preparing for the integration and integration of the first ontology. As a result of the first part of the algorithm, contained in the target ontology is the first source ontology. This part consists of four steps.

1. In order to prepare for integration, new empty ontology should be created in the Protégé editor, which are then integrated source ontologies. The created ontology IRI identifier must be given a value of "Integrated".
2. Then the first source ontology for integration should be selected; open the target ontology "Integrated" in editor Protégé and import the selected source ontology (e.g. "eQual").

3. The next step is to use the "Merge ontology" option in order to merge the source ontology (e.g. "eQual") and the destination ("Integrated"). In this way, the two ontologies will be included in the same file ontology.
4. Then, using refactoring tools, change the IRI identifier of the source ontology elements to "Integrated" value, which is the IRI identifier of the target ontology. This should be performed for all elements of the source ontology, with the exception of the concepts contained therein describing the quality model (for "eQual" the following concepts should be left unchanged: "eQual", "Service_interaction", "Empathy", "Trust", "Information_quality", "Usability", "Design" and "Usability1"). This action will preserve the quality model of the source ontology in the target ontology.

Part II: Integration of another source ontology in the section containing the criteria and quality model. As a result of this part of the algorithm, another source ontology will be placed into the target ontology (another website evaluation quality model and their criteria). It consists of two stages, having in total six steps.

Stage I: Selection of another source ontology to integration and alignment of parts of the source and target ontologies containing criteria.

1. Next source ontology for integration should be selected.
2. Then, to align the parts of the ontologies containing criteria, identify the relationship between the quality criteria present in the target ontology (i.e. "Integrated" in accordance with its current content) and in the source ontology to be integrated into the next stage. To do this, use the dictionaries, thesaurus, toolsof the Protégé editor and source literature concerning approaches represented by ontologies.

Stage II: Integration of another source ontology.

1. With Protégé editor open the target ontology "Integrated", perform another source ontology imports (e.g. "Ahn").
2. Using the tool "Merge ontologies", connect the source ontology (e.g. "Ahn") and the target (e.g. "Integrated", which already contains concepts from "eQual").
3. The next step is to change the IRI of the previously assigned values (e.g. "Ahn") to "Integrated" for all elements from the source ontology except those that describe the quality model (for Ahn they are concepts, "Ahn", "Information_quality", "System_quality" and "Service_quality"). This action will fit together just called concepts of criteria of source and target ontologies. The relations and concepts that are rooted in the ontology, i.e. "Criterion", "Quality" and "Service", will also be matched in this way.
4. The next action is to introduce the relationships of equivalence and subsumption between concepts of criteria derived from the source and the target ontologies. These relationships should be identified at the stage of aligning criteria (Part II, Phase I, Step 2).

Part III: Building a unified model of quality. This part consists of three stages, having in total 4 steps that should be done at the end of Part II of the algorithm.

Stage I: Aligning quality models. You need to specify the link between quality characteristics included in the unified model. As in the alignment step occurring in the second part of the integration algorithm, a number of tools must also be used here to facilitate the identification of links.

Stage II: Construction of a unified model of quality. This model should include the characteristics contained in the source and target ontology. After determining its structure, the previous unified model contained in the target ontology should be replaced. Quality characteristics of a unified model should contain relationship "hasCriterion". The criteria should be related to the characteristics of the model relationship "isCriterion".

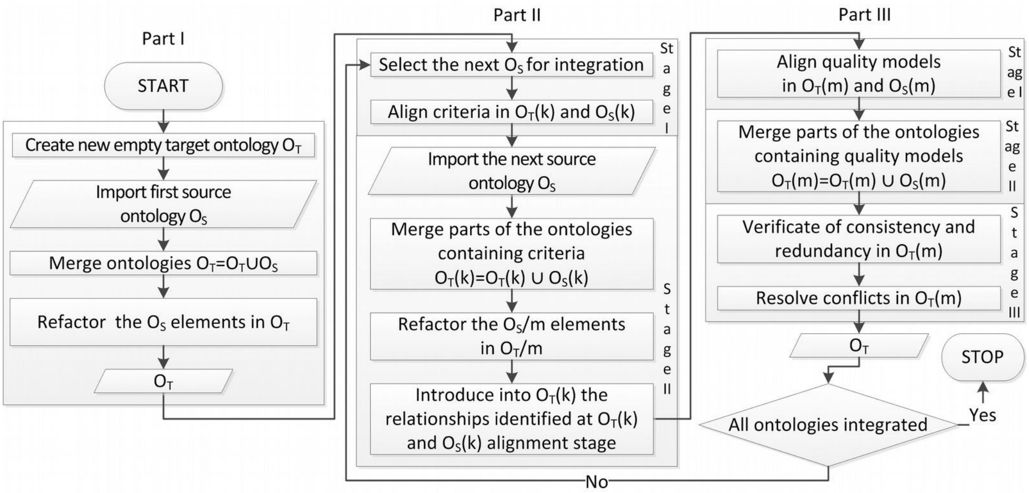

Part I
START
Create new empty target ontology $O_T$
Import first source ontology $O_S$
Merge ontologies $O_T$=$O_T$U$O_S$
Refactor the $O_S$ elements in $O_T$
$O_T$
Part II
Select the next $O_S$ for integration
Align criteria in $O_T(k)$ and $O_S(k)$
Stage I
Import the next source ontology $O_S$
Merge parts of the ontologies containing criteria $O_T(k)$=$O_T(k)$ ∪ $O_S(k)$
Refactor the $O_S$/m elements in $O_T$/m
Introduce into $O_T(k)$ the relationships identified at $O_T(k)$ and $O_S(k)$ alignment stage
Stage II
Part III
Align quality models in $O_T(m)$ and $O_S(m)$
Stage I
Merge parts of the ontologies containing quality models $O_T(m)$=$O_T(m)$ ∪ $O_S(m)$
Stage II
Verificate of consistency and redundancy in $O_T(m)$
Resolve conflicts in $O_T(m)$
Stage III
$O_T$
All ontologies integrated
Yes
STOP
No

Fig 8. Algorithm for ontology integration

Stage III: Verification of consistency and absence of redundancy and resolving conflicts.

1. This should focus on the consistency of the target ontology obtained using reasoner.
2. Conflicts involving two equivalent characteristics having different names can also occur here. Then you need to decide what the characteristics that will be included in the model should be called. Furthermore, one criterion may be associated by the relationship "isCriterion" with two characteristics included in the unified model. Such conflicts must be resolved by selecting the preferred characteristics and relationships derived from a previous version of the unified model.

A block diagram of the integration algorithm is shown in Figure 8.

## V. Results

In order to integrate the source ontology built earlier, according to the algorithm presented in point IV, the first part of the process of integration was completed, i.e. the creation of an empty target ontology, and integrated in there the eQual source ontology. Then a second ontology was selected for integration, which was Ahn ontology, and the alignment of concepts representing criteria was carried out. In this step links were identified between the target ontology, which already contains eQual, and the source ontology Ahn. The links detected are shown in Table 1. It should be noted that the alignment of the ontologies created some problems. Namely, certain relationships between concepts may not be obvious. Therefore, in order to determine the relationships between concepts of criteria, dictionaries, thesauruses and source literature concerning methods of integration represented by ontologies were used. The second problem results from the the fact that the integrated ontologies are creations of secondary structures, developed on the basis of quality assessment methods. The criteria contained in the individual methods are typically in the form of sentences, so that their names are in the long form. Therefore, in the analysis of the links there may appear to be some confusion regarding the full wording of the concept.

Based on the identified relationship between the criteria set out in Table 1, the next stage of integration was carried out, i.e. a merging ontologies in Protégé editor. In this framework the following things were achieved: import Ahn source ontology to target ontology, ontology merging with the use of the tool "Merge ontologies", the change values of IRI identifiers for concepts / criteria in Ahn ontology, linking respective pairs of concepts with equivalent relationships. Therefore, the target ontology, which is a unified ontology eQual and Ahn at the level of criteria, was achieved.

Then the concepts representing the characteristics of quality models Ahn and eQual were aligned, as shown in Table 2. The relationship between the characteristics set out in Table 2 is presented in that way due to the fact that the criteria included in the particular characteristics were symlinked in it at the same level of hierarchy. For example, if the "Service_quality" has been recognized as being equivalent to the characteristics of "Service_interaction", then e.g. the criterion of "instills_confidence " (belonging to the ”Service_quality”) would occur on the 3rd level of the proposed hierarchy along with the characteristics of the "Trust” (including the ”Service_interaction"). The criterion of "reputation" (belonging to the "Trust") would be at level 4, resulting thus in a hierarchy in which on one level in the context of the characteristics of "Service_interaction" are both criteria and subcharacteristics. This could cause problems with inconsistency during the inference by the reasoner, so in this case these characteristics associated with subsumption relationship.

The last step was to check the consistency and lack of redundancy in an integrated ontology. Such redundancy appeared as a result of the reasoner, in the case of the criteria from the target and source ontology having the same name or having been recognized as equivalent, and belonging to two different characteristics included in the unified model. In each of these cases, there was a redundancy of subsumption relationship between the criteria and characteristics of the

Table I. Aligning the concepts of criteria in the target ontology containing the eQual model and source ontology Ahn

| No. | Relation type | Target ontology (containing eQual) | | Source ontology Ahn | |
|---|---|---|---|---|---|
| | | Characteristic | Criterion / Concept | Characteristic | Criterion / Concept |
| 1. | equivalence | Usability1 | ease of navigation | System quality | ease of navigation |
| 2. | equivalence | Design | appropriateness design | System quality | appropriateness design style |
| 3. | equivalence | Design | sense of competency | Service quality | professionalism and competence |
| 4. | equivalence | Design | positive experience | System quality | audio-visual experience |
| 5. | equivalence | Information quality | information accuracy | Information quality | information accuracy |
| 6. | equivalence | Information quality | information believability | Information quality | information reliability |
| 7. | equivalence | Information quality | information timeliness | Information quality | information timeliness |
| 8. | equivalence | Information quality | appropriate format of information | Information quality | appropriate format of information |
| 9. | equivalence | Trust | security of personal information | System quality | security of personal information |
| 10. | equivalence | Empathy | personalization | Service quality | adaptation to the user's needs |
| 11. | equivalence | Trust | confident about delivery goods and services | Service quality | providing whatever promised |

TABLE II. ALIGNING THE CONCEPTS OF QUALITY MODEL IN THE TARGET ONTOLOGY CONTAINING THE EQUAL MODEL AND SOURCE ONTOLOGY AHN

| No. | Relation type | Target ontology (containing eQual) | | Source ontology Ahn | |
|---|---|---|---|---|---|
| | | Parent concept | Characteristic | Parent concept | Characteristic |
| 1. | equivalence | Quality | Information quality | Quality | Information quality |
| 2. | reverse subsumption | Quality | Service interaction | Quality | Service quality |

unified model. Selected examples of such redundancy are presented in Figure 9, which shows both the redundant relationship within one sheet, as well as within different characteristics.

For example, a couple of equivalent concepts "personalisation" and "adaptation_to_the_user's_needs" belongs to the characteristics of the "Service_quality" and "Empathy". On the other hand, the subcharacteristics are included in the single characteristic "Service_interaction". So here there is a redundancy in a single sheet. This redundancy resulted from the fact that the criterion of "personalisation" in the ontology eQual being part of the category of "Empathy" and the criterion of "adaptation_to_the_user's_needs" in Ahn ontology being one of the characteristics of "Service_quality". As a result of the recognition of these criteria as being equal to each other, they take the relationship between them and the characteristics to which they belong. A similar situation occurs, for example, in the case of matched pairs of concepts "sense_of_competency" and "professionalism_and _competence", which are among the characteristics of "Service_quality" and "Design" as a result of the equivalence that exists between them. This type of redundancy relationship is resolved in accordance with the assumptions described in Section III of the article, i.e. for the benefit of the target ontology, keeping derived from the relationship, and removing accounts from the source ontology (in this case from the ontology Ahn). Integrating ontology models eQual and Ahn and including the quality model unifies these two methods, presented in the elementary form [48], in the form deduced by the reasoner [49].

Other iterations of the process of integration of the various source ontology and of the target ontology integration proceeded analogously to those presented above. Previously built ontologies accompanied the target ontology in turn systematically: SiteQual, Website Evaluation Questionnaire and Web Portal Site Quality [50]. Therefore, the ontology [50] contains five source ontologies. For the unified model of quality, consisting of five basic models as deduced by the reasoner, each model is presented separately in [51].

## VI. CONCLUSION

Even a cursory analysis of [50] and [51] shows a very high level of complexity in this built ontology. At the same time, it can be concluded that the construction of such an extensive ontology, as a result of the integration of the source ontologies, is much less complex than its construction process from scratch, even with the use of formal ontology construction methodology. Thanks to the presented approach, the problem of ontology construction unifying quality assessment methods was decomposed for a few minor problems involving construction of the source ontologies were integrated. This resulted in an ontology containing quality models used in the various methods and a model that unifies the different methods. It is worth noting that in all of the integrated methods a total of 115 criteria were used. In the integrated ontology, 94 criteria are listed, due to the fact that part of the criteria are repeated in the various methods. However, when one takes into account the fact that some criteria are equivalent to those of others, one can consider there to be 70 different quality criteria in an integrated ontology. Therefore, we managed to limit the number of applied ontology criteria by almost 40%. The resulting

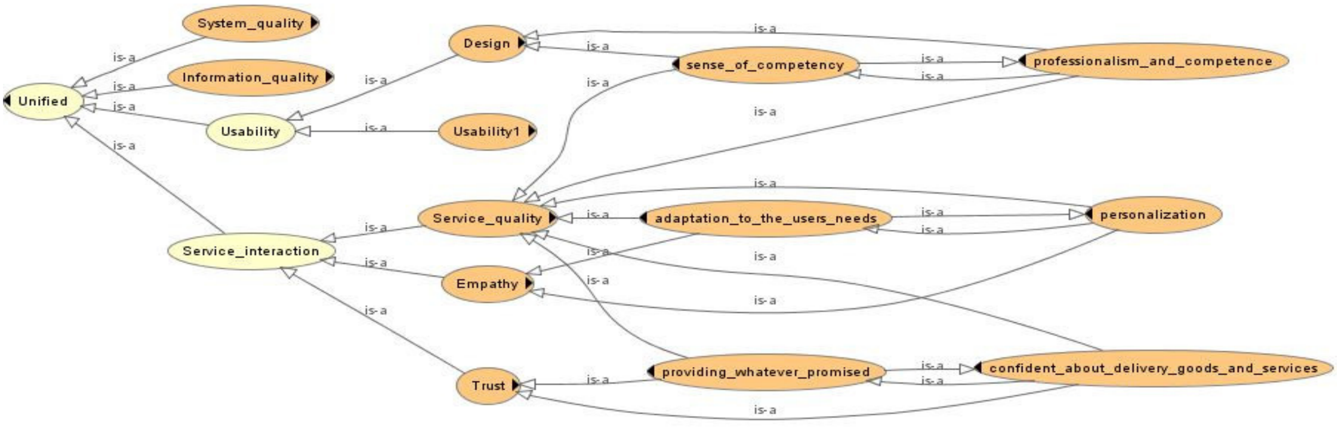


Fig 9. Redundancy of relationships in the target ontology containing the source ontologies eQual and Ahn

ontology presented in [51] can act as a repository of domain knowledge, due to the fact that it includes a number of methods and models that evaluate the quality of websites. It may also allow for the integration of heterogeneous data from a variety of assessment methods, and thus assessment websites through a variety of methods defined in the ontology, so that the individual results of the assessment can be compared in a the same terminology and a reference plane. Constructed repository of knowledge, together with presented in [16] website assessment criteria selection process could be at the core of the expert system of website quality assessment, which should be the direction of further research. In addition, further work should include the development of ontology about the possibility of environmental data records about ontology users.

## Acknowledgment

The work was partially supported by Fellowship co-Financed by European Union within European Social Fund, by European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE].

## References

[1] W.C. Chiou, C.C. Lin, C. Perng, "A strategic framework for website evaluation based on a review of the literature from 1995–2006," Information & Management, vol. 47, no. 5-6, pp. 282-290, 2010, http://dx.doi.org/10.1016/j.im.2010.06.002

[2] IAB, IAB internet advertising revenue report. 2013 full year results. PricewaterhouseCoopers LLP, 2014.

[3] IAB Europe, Adex Benchmark 2013. European online advertising expenditure.IAB Europe, 2014.

[4] ECOMMERCE Europe, European B2C E-commerce Report 2014. www.ecommerce-europe.eu, 2014.

[5] S. Kim, L. Stoel, "Dimensional hierarchy of retail website quality," Information & Management, vol. 41, no. 5, pp. 619-633, 2004.

[6] J. Hwang, Y.S. Yoon, N.H. Park, "Structural effects of cognitive and affective reponses to web advertisements, website and brand attitudes, and purchase intentions: The case of casual-dining restaurants," International Journal of Hospitality Management, vol. 30, no. 4, pp. 897-907, 2011, http://dx.doi.org/10.1016/j.ijhm.2011.01.011

[7] W.C. Chou, Y. Cheng, "A hybrid fuzzy MCDM approach for evaluating website quality of professional accounting firms," Expert Systems with Applications, vol. 39, no. 3, pp. 2783-2793, 2012, http://dx.doi.org/10.1016/j.eswa.2011.08.138

[8] ISO/IEC 25010:2010(E), Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models.

[9] S.J. Barnes, R. Vidgen, "Data triangulation and web quality metrics: A case study in e-government," Information & Management, vol. 43, no. 6, pp. 767-777, 2006.

[10] T. Ahn, S. Ryu, I. Han, "The impact of Web quality and playfulness on user acceptance of online retailing," Information & Management, vol. 44, no. 3, pp. 263-275, 2007.

[11] H.W. Webb, L.A. Webb, "SiteQual: an integrated measure of Web site quality," Journal of Enterprise Information Management, vol. 17, no. 6, pp. 430-440, 2004.

[12] Z. Yang, S. Cai, Z. Zhou, N. Zhou, "Development and validation of an instrument to measure user perceived service quality of information presenting Web Portals," Information & Management, vol. 42, no. 4, pp. 575-589, 2005.

[13] S. Elling, L. Lentz, M. de Jong, H. van den Bergh, "Measuring the quality of governmental websites in a controlled versus an online setting with the 'Website Evaluation Questionnaire'," Government Information Quarterly, vol. 29, no. 3, pp. 383-393, 2012.

[14] H. Sorum, K.N. Andersen, T. Clemmensen, "Website quality in government: Exploring the webmaster's perception and explanation of website quality," Transforming Government: People, Process and Policy, vol. 7, no. 3, pp. 322-341, 2013, http://dx.doi.org/10.1108/TG-10-2012-0012

[15] T. Kaya, "Multi-attribute Evaluation of Website Quality in E-business Using an Integrated Fuzzy AHPTOPSIS Methodology," International Journal of Computational Intelligence Systems, vol. 3, no. 3, pp. 301-314, 2010, http://dx.doi.org/10.1080/18756891.2010.9727701

[16] P. Ziemba, M. Piwowarski, J. Jankowski, J. Wątróbski, "Method of Criteria Selection and Weights Calculation in the Process of Web Projects Evaluation," Lecture Notes in Artificial Intelligence, vol. 8733, pp. 684-693, 2014, http://dx.doi.org/10.1007/978-3-319-11289-3_69

[17] M. Hepp, "Ontologies: state of the art, business potential, and grand challenges," in Ontology Management. Semantic Web, Semantic Web Services, and Business Applications, M. Hepp, P. de Leenheer, A. de Moor, Y. Sure, Ed. Heidenberg: Springer, 2008, pp. 3-23.

[18] I.F. Cruz, H. Xiao, "Ontology Driven Data Integration in Heterogeneous Networks," Studies in Computational Intelligence, vol. 168, pp. 75-98, 2009.

[19] T.R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, vol. 5, no. 2, pp. 199–220, 1993.

[20] A. Guzman-Arenas, A.D. Cuevas, "Knowledge accumulation through automatic merging of ontologies," Expert Systems with Applications, vol. 37, no. 3, pp. 1991-2005, 2010, http://dx.doi.org/10.1016/j.eswa.2009.06.078

[21] N. Villanueva-Rosales, M. Dumontier, "yOWL: An ontology-driven knowledge base for yeast biologists," Journal of Biomedical Informatics, vol. 41, no. 5, pp. 779-789, 2008.

[22] L.Y. Shue, C.W. Chen, W. Shiue, "The development of an ontology-based expert system for corporate financial rating," Expert Systems with Applications, vol. 36, no. 2, pp. 2130-2142, 2009, http://dx.doi.org/10.1016/j.eswa.2007.12.044

[23] R. Saa, A. Garcia, C. Gomez, J. Carretero, F. Garcia-Carballeira, "An ontology-driven decision support system for high-performance and cost-optimized design of complex railway portal frames," Expert Systems with Applications, vol. 39, no. 10, pp. 8784-8792, 2012, http://dx.doi.org/10.1016/j.eswa.2012.02.002

[24] B. Magoutas, C. Halaris, G. Mentzas, "An Ontology for the Multi-perspective Evaluation of Quality in E-Government Services," Lecture Notes in Computer Science, vol. 4656, pp. 318-329, 2007.

[25] L. Mich, M. Franch, "Instantiating Web Sites Quality Models: an Ontologies driven Approach," in Proceedings of the CAISE'05 Workshop on Web Oriented Software Technologies, 2005.

[26] N. Casellas, Legal Ontology Engineering. Methodologies, Modelling Trends, and the Ontology of Professional Judicial Knowledge. Dordrecht: Springer, 2011, http://dx.doi.org/10.1007/978-94-007-1497-7

[27] O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez, "Ontological Engineering: Principles, Methods, Tools and Languages," in Ontologies for Software Engineering and Software Technology, C. Calero, F. Ruiz, M. Piattini, Ed. Berlin: Springer, 2006, pp. 1-48.

[28] Y. Sure, C. Tempich, D. Vrandecic, "Ontology Engineering Methodologies. in Semantic Web Technologies. Trends and Research," in Ontology-based Systems, J. Davies, R. Studer, P. Warren, Ed. Chichester: Wiley, 2006, pp. 171-190.

[29] M. Fernandez-Lopez, A. Gomez-Perez, "Overview and analysis of methodologies for building ontologies," The Knowledge Engineering Review, vol. 17, no. 2, pp. 129-156, 2002.

[30] A. Gomez-Perez, M. Fernandez-Lopez, O. Corcho, "Methodologies and Methods for Building Ontologies," in Ontological Engineering, With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web, A. Gomez-Perez, M. Fernandez-Lopez, O. Corcho, Ed. London: Springer, 2004, pp. 107-197.

[31] O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez, A. Lopez-Cima, "Building Legal Ontologies with METHONTOLOGY and WebODE," Lecture Notes in Computer Science, vol. 3369, pp. 142-157, 2005.

[32] B. Villazon-Terrazas, J. Ramirez, M.C. Suarez-Figueroa, A. Gomez-Perez, "A network of ontology networks for building e-employment advanced systems," Expert Systems with Applications, vol. 38, no. 11, pp. 13612-13624, 2011, http://dx.doi.org/10.1016/j.eswa.2011.04.125

[33] H. Pinto, A. Gomez-Perez, J. Martins, "Some issues on ontology integration," in Proceedings of IJCAI99's Workshop on Ontologies and Problem Solving Methods, 1999.

[34] J. de Bruijn, M. Ehrig, C. Feier, F. Martin-Recuerda, F. Scharffe, M. Weiten, "Ontology Mediation, Merging, and Aligning," in Semantic Web Technologies. Trends and Research in Ontology-based Systems, J. Davies, R. Studer, P. Warren, Ed. Chichester: Wiley, 2006, pp. 171-190.

[35] M. Klein, "Combining and relating ontologies: an analysis of problems and solutions," in Proceedings of the IJCAI Workshop on Ontologies and Information Sharing, 2001, pp. 53-62.

[36] J. Zhang, Y. Lv, "An Approach of Refining the Merged Ontology," in Proceedings of the 9th IEEE International Conference on Fuzzy Systems and Knowledge Discovery, 2012, pp. 802-806, http://dx.doi.org/10.1109/FSKD.2012.6233973

[37] M. Taboada, D. Martinez, J. Mira, "Experiences in reusing knowledge sources using Protege and PROMPT," International Journal Human-Computer Studies, vol. 62, no. 5, pp. 597-618, 2005.

[38] M. Gaeta, F. Orciuoli, P. Ritrovato, "Advanced ontology management system for personalised e-Learning," Knowledge-Based Systems, vol. 22, no. 4, pp. 292-301, 2009, http://dx.doi.org/10.1016/j.knosys.2009.01.006

[39] M. Knorr, J.J. Alferes, P. Hitzler, "Local closed world reasoning with description logics under the well-founded semantics," Artificial Intelligence, vol. 175, no. 9-10, pp. 1528-1554, 2011, http://dx.doi.org/10.1016/j.artint.2011.01.007

[40] N.F. Noy, M.A. Musen, "The PROMPT suite: interactive tools for ontology merging and mapping," International Journal of Human-Computer Studies, vol. 59, no. 6, pp. 983-1024, 2003.

[41] http://tinyurl.com/WQM-eQual-example-inst; http://tinyurl.com/WQM-eQual-inferred-example-ins

[42] http://tinyurl.com/WQM-Ahn; http://tinyurl.com/WQM-Ahn-inferred

[43] http://tinyurl.com/WQM-SiteQual; http://tinyurl.com/WQM-SiteQual-inferred

[44] http://tinyurl.com/WQM-WEQ; http://tinyurl.com/WQM-WEQ-inferred

[45] http://tinyurl.com/WQM-WPSQ; http://tinyurl.com/WQM-WPSQ-inferred

[46] P. Ziemba, J. Jankowski, J. Wątróbski, J. Becker, "Knowledge Management in Website Quality Evaluation Domain," Lecture Notes in Artificial Intelligence, vol. 9330, pp. 75-85, 2015, http://dx.doi.org/10.1007/978-3-319-24306-1_8

[47] http://protege.stanford.edu/

[48] http://tinyurl.com/ont-integration1

[49] http://tinyurl.com/ont-integration1-inferred

[50] http://tinyurl.com/ont-integration

[51] http://tinyurl.com/ont-integration-inferred

DOI: 10.15439/2015F58
ACSIS, Vol. 5

# Multicriteria support of choosing a group decision

Andrzej Łodziński
Warsaw University of Life
Sciences ul. Nowoursynowska 159,
02 - 776 Warszawa, Poland
Email: andrzej_lodzinski@sggw.pl

*Abstract*—**The paper presents the method of a group decision making in a competitive environment. We deal with a group decision when the group of people with different preferences are to make one single decision. The group decision selection process is modeled with the use of multi-criteria optimization task. It is solved with the use of reference point method. This method is an interactive method in which every person specifies its requirements in the form of a reference point, expressing the desired values for its evaluation function. On the basis of the provided reference point, a scalar achievement function is built. Maximization of this function generates a solution of the multi-criteria task. This solution is presented to every person for acceptance or as a basis for the modification of the reference point. The paper contains the example of application of the proposed method to support a group decision by three people with different preferences.**

## I. Introduction

The paper presents the method of a group decision making in a competitive environment. A group decision means that several people, whose interests are conflicting, are supposed to make one decision. One should conjoin divergent interest of all people, in order to arrive to a compromise solution for all. The aim of the selection decision is the best solution for the group, and not for individual members of that group. No solution is selected for a single person, it is looking for all persons in a group.

The selection process of a group decision can be modeled with the use of game theory {5], [6], [11], [14].

The process of a group decision making is modelled with the use of multi-criteria optimization with a vector evaluation function. Each coordinate of this vector is the value of decision evaluation function for each person. The decision selection is performed with the use of an interactive computer system. Each person provides his proposition of the decision result for his/her evaluation function. These propositions constitute parameters of the multi-criteria optimization task and that is then solved. Then, each person evaluates the solution. Each of them may agree to the obtained result or not. In the second case the person or persons provide a new value of the parameter - their new propositions and the problem is solved again for the new parameters. The selection process is not a one-time process, but an iterative process of learning about the decision making. The process of a group decision making is to support the members of a group to obtain as much as possible.

## II. Modeling of a group decision making

Our aim is to find an adequate group decision in a competitive case. The process of making a group decision is modeled by introducing a respective decision variable. Moreover, there are the s. c. decision evaluation functions, which constitute criteria evaluating the solution from the point of view of each person. Each person has its own evaluation criterion - its evaluation function. These functions are a measure of satisfaction of every person by a given solution; they evaluate a degree of achieving a goal by every person. The bigger value of the function means a bigger satisfaction, so every function is maximized. The basis for evaluation and selection of a group decision are all evaluation functions – the criteria for all persons.

The group decision selection problem is modeled as a multi-criteria optimization task:

$$\max_x\{(f_1(x), f_2(x),\ldots,f_k(x)): \quad x \in X_0\} \quad (1)$$

where: $1,2,\ldots,k$ – particular persons,

$X_0 \subset R^n$ - the feasible set,

$x = (x_1, x_2,..., x_n) \in X_0$ - a group decision,

$f_i : X_0 \to R$ – the decision evaluation function by a person $i$, $i = 1,2,\ldots,k$.

Task (1) relies on finding such a feasible decision $x \in X_0$ for which $k$ evaluations attain the best possible values. There is a common restraining of decision which constitutes a solution.

The vector functions $f = (f_1, f_2,..., f_k)$ defines the correspondence of any decision variable vector $x \in X_0$ and

978-83-60810-66-8/$25.00©2015, IEEE

the respective evaluation vector $y = (y_1,..., y_k)$. They measure the decision quality from the point of view of decision evaluation. Particular coordinates $y_i = f_i(x), \quad i = 1,..., k$ are scalar functions of decision evaluation for $i - th$ person, $i = 1,2,...,k$. The image of the feasible set $X_0$ by the function $f$ constitutes a collection of achievable evaluation vectors $Y_0$.

Task (1) is formulated in the domain of evaluations, i.e. the following task is considered:

$$\max_x\{(y_1,...,y_k): \quad y \in Y_0\} \tag{2}$$

where : $x \in X$ – a vector of decision variables,

$y = (y_1,..., y_k)$ – a vector of evaluations, particular coordinates $y_i$ representing the results of the decision $x$ for the person $i$, $i = 1,2,...,k$,

$Y_0 = f(X_0)$ – the set of achievable evaluation vectors.

The set of achievable evaluation vectors $Y_0$ is provided in a non-explicit way – through the set of feasible decisions $X_0$ and the model $f = (f_1,..., f_k)$. In order to calculate the value $y$, a simulation of the respective model is necessary: $y = f(x), x \in X_0$.

The aim of task (1) is the aid in finding a decision that would be the most compromising for all persons.

## III. Equitably efficient solution

The solution in the selection decision process should satisfy certain properties that persons accept as reasonable. Namely, such a solution should be:

- an optimal solution in the sense of Pareto – i.e. such that you can not improve the solution for one person without worsening the solution for the other persons,
- symmetric solution – i.e. that it should not depend on the way the persons are numbered; as no one is more important that the others. Persons are treated in the same way in the sense that the solution does not depend on the name of person or on other factors specific to a given person,
- equalizing solution - that is, a vector that has less variation of coordinates of evaluation is preferred in comparison to a vector with the same sum of coordinates, but with a greater diversity of coordinates.

Any decision, that satisfies the above conditions is an equitably efficient decision. Hence, this Pareto-optimal decision satisfies additional conditions – anonymity and the axiom of equalizing solution.

The non-dominated results ( Pareto - optimal) are defined as follows:

$$\hat{Y}_0 = \{\hat{y} \in Y_0 : (\hat{y} + \tilde{D}) \cap Y_0 = \varnothing\} \tag{3}$$

where: $\tilde{D} = D \setminus \{0\}$ – a positive cone without the top. As a positive cone, it can be adopted $\tilde{D} = R^k_+$. Appropriate acceptable decisions are specified in the decision space. The decision $\hat{x} \in X_0$ is called efficient decision (Pareto - optimal), if the corresponding vector of evaluations $\hat{y} = f(\hat{x})$ is a non-dominated vector [7], 16], [17].

Finally, in the multi-criteria problem (1), which is used to select a group decision, the relation of preferences should satisfy additional properties: anonymity property and the property of equalizing solution.

The relation is called an anonymous relation if, for every vector $y = (y_1, y_2,..., y_k) \in R^k$ and for any permutation $P$ of the set $\{1,...,k\}$, the following property holds:

$$(y_{P(1)}, y_{P(2)},..., y_{P(k)}) \approx (y_1, y_2,..., y_k) \tag{4}$$

No distinction is made between the results that differ in the arrangement of coordinates. Evaluation vectors having the same coordinates, but in a different order are identified and that is the anonymity property.

Moreover, the relation of preferences satisfies the axiom of equalizing transfer, if and only if the following condition is satisfied:

for the evaluation vector $y = (y_1, y_2,..., y_k) \in R^k$:

$$\begin{gathered} y_{i'} > y_{i''} \Rightarrow y - \varepsilon \cdot e_{i'} + \varepsilon \cdot e_{i''} \succ y \\ \text{for } 0 < y_{i'} - y_{i''} < \varepsilon \end{gathered} \tag{5}$$

Equalizing transfer is a slight deterioration of a better coordinate of evaluation vector and simultaneously improvement of a poorer coordinate. The resulting evaluation vector is strictly preferred in comparison to the initial evaluation vector. This is a structure of equalizing – the evaluation vector with less diversity of coordinates is preferred in relation to the vector with the same sum of coordinates, but with their greater diversity.

Non-dominated vector satisfying the anonymity property and the axiom of equalizing transfer, is called equitably non-dominated vector. The set of equitably non-dominated vectors is denoted by $\hat{Y}_{0W}$. In the decision space, the equitably efficient decisions are specified. The decision $\hat{x} \in X_0$ is called equitably efficient decision, if the corresponding evaluation vector $\hat{y} = f(\hat{x})$ is an equitably non-dominated vector. The set of equitably efficient decisions is denoted by $\hat{X}_{0W}$ [2], [9], [13].

The relation of equalizing domination can be expressed as the relation of inequality for cumulative, ordered evaluation vectors. This relation can be determined with the use of mapping $\overline{T}: R^k \to R^k$ that cumulates nondecreasing coordinates of evaluation vector.

The transformation $\overline{T}: R^k \to R^k$ is defined as follows:

$$\overline{T}_i(y) = \sum_{l=1}^{i} T_i(y) \quad \text{for } i = 1,2,...,k \tag{6}$$

Define namely by $T(y)$ the vector with decreasing ordered coordinates of the vector $y$, i.e.. $T(y) = (T_1(y), T_2(y),..., T_k(y))$, where $T_1(y) \le T_2(y) \le ... \le T_k(y)$ and there is a permutation $P$ of the set $\{1,...,k\}$, such that $T_i(y) = y_{P(i)}$ for $i = 1,..,k$.

The relation of equalizing domination $\succ_e$ is a simple vector domination for evaluation vectors with nondecreasing coordinates of evaluation vector [2], [9], [13].

The evaluation vector $y^1$ equitably dominates the vector $y^2$ if the following condition is satisfied:

$$y^1 \succ_e y^2 \Leftrightarrow \overline{T}(y^1) \ge \overline{T}(y^2) \tag{7}$$

Solving the problem of decision selection in the group decision process consists in determination of the equitably efficient decision which satisfies the preferences of every persons.

## IV. Scalaring the problem

For determination of equitably efficient solutions of multi-criteria task (1), a specific multi-criteria task is solved. It is the task with the vector function of the cumulative, ordered evaluation vectors, i.e. the following task:

$$\max_y \{(\overline{T}_1(y), \overline{T}_2(y),..., \overline{T}_k(y)): \quad y \in Y_0\} \tag{8}$$

where: $y = (y_1, y_2,..., y_k)$ – an evaluation vector,

$\overline{T}(y) = (\overline{T}_1(y), \overline{T}_2(y),..., \overline{T}_k(y))$ a cumulative, ordered evaluation vector,

$Y_0$ – the set of achievable evaluation vectors.

Effective solution of multi-criteria optimization task (8) is an equitably efficient solution of the multi-criteria task (1).

To determine the solution of a multi-criteria task (8), the scalaring of this task with the scalaring function $s: Y_0 \times \Omega \to R^1$ is solved:

$$\max_x \{s(y, \overline{y}): x \in X_o\} \tag{9}$$

where: $y = (y_1, y_2,..., y_k)$ – an evaluation vector,

$\overline{y} = (\overline{y}_1, \overline{y}_2,..., \overline{y}_k)$ – a control parameters for individual evaluations.

It is the task of single objective optimization with specially created scalaring function of two variables - the evaluation vector $y \in Y$ and control parameter $\overline{y} \in \Omega \subset R^k$; we have thus $s: Y_0 \times \Omega \to R^1$. The parameter $\overline{y} = (\overline{y}_1, \overline{y}_2,..., \overline{y}_k)$ is available to each person. That allows any person is capable to review the set of equitably efficient solutions.

The optimal solution of task (9) should be a solution of the multiple criteria task (8). Scalaring function should satisfy certain properties - the property of completeness and that of sufficiency. The property of sufficiency means that for each control parameter $\overline{y}$ the solution of the scalaring task is the equitably efficient solution, i.e. $\hat{y} \in \hat{Y}_{0W}$. The property of completeness means, that by appropriate changes of parameter $\overline{y}$ any solution $\hat{y} \in \hat{Y}_{0W}$ can be achieved. Such a function completely characterizes equitably efficient solutions. Inversely, each maximum of such a function is an equitably efficient solution. Each equitably efficient solutions can be obtained with appropriate value of control parameter $\overline{y}$.

Complete and sufficient parameterization of the set of equitably efficient solutions $\hat{Y}_{0W}$ can be achieved, using the method of the reference point for the task (8). In this method the aspiration levels are applied as control parameters. Aspiration level is such value of the evaluation function that satisfies a given person.

The scalaring function defined in the method of reference point is as follows:

$$s(y, \overline{y}) = \min_{1 \le i \le k} (\overline{T}_i(y) - \overline{T}_i(\overline{y})_i) + \varepsilon \cdot \sum_{i=1}^{k} (\overline{T}_i(y) - \overline{T}_i(\overline{y})_i) \tag{10}$$

where: $y = (y_1, y_2,..., y_k)$ – an evaluation vector,

$\overline{T}(y) = (\overline{T}_1(y), \overline{T}_2(y),..., \overline{T}_k(y))$ - a cumulative, ordered evaluation vector,

$\bar{y} = (\bar{y}_1, \bar{y}_2, ..., \bar{y}_k)$ – a vector of aspiration levels,

$T(\bar{y}) = (T_1(\bar{y}), T_2(\bar{y}), ..., T_k(\bar{y}))$ - a cumulative, ordered vector of aspiration levels,

$\varepsilon$ – an arbitrary small, positive adjustment parameter.

Such scalaring function is called a function of achievement. The aim is to find a solution that approaches as close as possible the specific requirements – the aspiration levels [2], [7], [13].

Maximizing this function w. r. to $y$ determines the equitably efficient solution $\hat{y}$ and the equitably efficient decision $\hat{x}$. Note, the equitably efficient solution $\hat{x}$ depends on the aspiration level $\bar{y}$. A solution of the multicriteria optimization problem makes correspond of solution proposals of particular members of the group to the respective levels of aspiration.

## V. Method of supporting the group decision

The solution of the multi-criteria task (8) is a set of equitably efficient solutions. In order to solve a given problem it is necessary to pick one solution which will be evaluated by all persons. Due to the fact that the equitably efficient solution is a whole set of solutions, the persons perform the selection with the help of an interactive computer system. Such a system makes possible to have a guided overview of a whole set of solutions. The tool used to view this set of solutions is function (10). Maximum of this function depends on the parameters $\bar{y}_i, i = 1,2,,...,k$, which are applied by all persons. In the reference point method each person expresses its preferences by specification, with the aid of his/her evaluation function, of such a value that would be fully satisfactory. That is the value of the aspiration level for his/her evaluation function. For any stage of the selection process the persons may provide different aspiration levels. Such levels of aspiration constitute steering parameters of the scalarization function. On this basis the task is solved and the system proposes the solution corresponding to the current values of those parameters - for further analysis.

The method of supporting the group decision is the following:

1. Iterative algorithm - propositions of particular decision.
   1.1. Interaction with the system - each person provides his/her own proposition of the decision for its evaluation function as his/her level of aspiration $\bar{y}_i, i = 1,2,...,k$.
   1.2. Calculations – computing particular values from the equitably efficient solution $\hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_k) \in \hat{Y}_{0W}$ and the equitably efficient decision $\hat{x} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_k) \in \hat{X}_{0W}$.
   1.3. Evaluation of the obtained solution - each person may accept the solution or not. In the second case - each person provides his/her new proposition and provides a constant value of his/her level of aspiration $\bar{y}_i, i = 1,2,..,k$ and another equitably efficient solution is set out. (Return to sub-point 1.2).
2. Establishing the decision, when the decision fulfills the requirements of all persons.

This is not a single optimization act but a dynamic process of looking for solutions, during which the persons learn and may change their preferences. Comparing the result of the decision $\hat{y}_i, i = 1,2,...,k$ with the aspiration point $\bar{y}_i, i = 1,2,...,k$, each person finds what is not achievable and how his/her proposition $\bar{y}_i, i = 1,2,...,k$ is far from a possible solution $\hat{y}_i, i = 1,2,...,k$. This allows for a proper modification of their own propositions – with regard to their own levels of aspiration. These levels of aspirations are specified adaptively in the process of teaching. This process finishes when such decisions are found, which allow to fulfill the aspirations of persons in a maximum possible degree.

Method of supporting the group decision is presented at diagram 1.

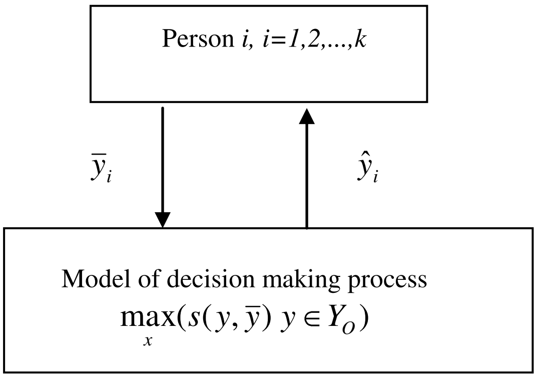

Person *i, i=1,2,...,k*
$\bar{y}_i$
$\hat{y}_i$
Model of decision making process
$\max_x(s(y,\bar{y})\ y \in Y_O)$

Fig. 1. Method of supporting the group decision making

Such a manner of making decisions does not impose any strict scenario and allows for the possibility of modifying the preferences for every person in the decision making process. Persons learn during the selection process about the decision making problem. The persons may check the results of every allowed proposition. All members of a group have an equal part in the decision making process. They all have an equal possibility for eventual changes of their preferences. The computer will not replace people in the decision making process; the whole process of selecting a decision is guided by all persons.

## VI. Example

To illustrate the support of the group decision making the following example is presented - selection of a group decision by three persons [3].

The problem of selecting the decision is the following:

1,2,3 –persons,

$X_0 = \{x \in R^4 : 2\cdot x_1 + x_2 + 4\cdot x_3 + 3\cdot x_4 \le 60,$

$3\cdot x_1 + 4\cdot x_2 + x_3 + 2\cdot x_4 \le 60,$ -

$x_1 \ge 0,\ x_2 \ge 0,\ x_3 \ge 0,\ x_4 \ge 0\}$

the feasible set,

$x = (x_1, x_2, x_3, x_4) \in X_0$ - a group decision,

$f_1(x) = 3\cdot x_1 + x_2 + 2\cdot x_3 + x_4$ – the decision evaluation function by person 1,

$f_2(x) = x_1 - x_2 + 2\cdot x_3 + 4\cdot x_4$ – the decision evaluation function by person 2,

$f_3(x) = -x_1 + 5\cdot x_2 + x_3 + 2\cdot x_4$ – the decision evaluation function by person 3,

The problem of selection of a group decision is expressed in the form of multi-criteria optimization task with three evaluation functions:

$$\max_x\{(3\cdot x_1 + x_2 + 2\cdot x_3 + x_4,\\ x_1 - x_2 + 2\cdot x_3 + 4\cdot x_4,\\ -x_1 + 5\cdot x_2 + x_3 + 2\cdot x_4): \quad x \in X_0\} \tag{11}$$

where: $X_0$ - the feasible set,

$x = (x_1, x_2, x_3, x_4) \in X_0$ - a group decision

To select the solutions of (11), the reference point method is used for the task with cumulated coordinates of the evaluation vector ordered in a non decreasing manner.

The first step of the vector analysis is to use the one-criteria optimization for evaluation function of every person separately. As a result there is the so-called matrix of goal realization including the values of each criterion, received by solving one of the three one-criteria problems. This matrix allows for evaluation of the scope of changes of particular evaluation function on the allowed set; it provides a certain information about the conflict of the evaluation functions. Matrix of goal realizations generates the utopia vector that represents the best values of each separate criterion.

Table 1. Matrix of goal realization with the utopia vector.

| Optimization criterion | | Solution<br>$\hat{y}1$ | <br>$\hat{y}2$ | <br>$\hat{y}3$ |
|---|---|---|---|---|
| Person's Evaluation 1 | $y1$ | 66 | 30 | -12 |
| Person's Evaluation 2 | $y2$ | 20 | 80 | 40 |
| Person's Evaluation 3 | $y3$ | 15 | -15 | 75 |
| Utopia vector | | 66 | 80 | 75 |

When analyzing the table 1 it might be observed that the biggest selection possibilities has person 2, lower - person 3 and the lowest one - person 1.

People in the group do control the process by means of aspiration levels. The multi-criteria analysis is presented in table 2.

Table 2. Interactive analysis of looking for solutions.

| Iteration | | Pers. 1<br>$\hat{y}1$ | Pers2<br>$\hat{y}2$ | Pers.3<br>$\hat{y}3$ |
|---|---|---|---|---|
| 1 | Aspiration point $\bar{y}$ | 66 | 80 | 75 |
| | Solution $\hat{y}$ | 24 | 66 | 66 |
| 2 | Aspiration point $\bar{y}$ | 55 | 65 | 60 |
| | Solution $\hat{y}$ | 26,76 | 65 | 60 |
| 3 | Aspiration point $\bar{y}$ | 50 | 60 | 55 |
| | Solution $\hat{y}$ | 30,17 | 60 | 55 |
| 4 | Aspiration point $\bar{y}$ | 48 | 58 | 53 |
| | Solution $\hat{y}$ | 31,54 | 60 | 55 |
| 5 | Aspiration point $\bar{y}$ | 45 | 55 | 50 |
| | Solution $\hat{y}$ | 33,59 | 55 | 50 |
| 6 | Aspiration point $\bar{y}$ | 43 | 53 | 48 |
| | Solution $\hat{y}$ | 34,5 | 53 | 48 |

At the beginning of the analysis every person specifies its preferences as the aspiration point equal to the utopia vector coordinate. The obtained solution prefers by person 2 and person 3 and is too small for person 1. The group wants to improve the solution. Therefore, all the people in the group decrease their requirements in the next iteration. One obtains a slight improvement for person 1 and deterioration for person 2 and person 3, but the group wants to improve the solution for person 1. In subsequent iterations all individuals reduce their requirements and improve the value obtained for the assessment of persons 1, at the cost of two other person. For iterations 5 and 6 the following decisions are found: $\hat{x}^5 = (3{,}78;\ 5{,}12;\ 2{,}02;\ 13{,}07)$ and $\hat{x}^6 = (4{,}27;\ 5{,}07;\ 2{,}41;\ 12{,}24)$. The analysis reveals that there is deep influence of person 2 and 3 on the solution; however, for person 1 it is far less significant.

The final selection of the specific solution depends on the preferences of all persons. The presented example shows that the method allows the persons to learn about their decision-making possibilities. The search for compromise for everyone is continued in this method.

## VII. Conclusion

The paper presents the method of supporting the group decision making. The selection of decision is performed by solving the multi-objective task according to the optimization criteria. This method is characterized by:

- the use of information about everyone's preferences in the form of aspiration points - values of goal function that are fully satisfactory to them and the optimal option of the scalar achievement function in order to organize the interactions with all persons,
- the assumption that the preferences of persons are not completely fixed and they may change during the decision making process.

Reference point method applied to the of multi-criteria problem indicates a solution which would be suited to the preferences of all individuals in the group.

The participation of any members of the group in the decision making implies acceptation of a final choice. In such a course of action one does not replace people in decision making. The whole process of decision making is guided by all persons.

## References

[1] U. Chevaleyre,. J. Endriss, M. Lang, and N. Maudet "A Short Introduction to Computational Social Choice". Proc. SOFSEM-2007, Springer-Verlag, 2007.

[2] M. Kostreva., W. Ogryczak. and A., P., Wierzbicki "Equitable Aggregation and Multiple Criteria Analysis". European Journal of Operational Research, vol. 122., 2004.

[3] S. Krawczyl "Mathematical analysis of the situation of decision-making". Warsaw ( in polish), 1990.

[4] L. Kruś "On some Procedures Supporting Multicriteria Cooperative Decisions". Foundations of Computing and Decision 33, (3), 2008.

[5] L. Kruś "Multi-criteria decision cooperative. Computer-aided methods", Warsaw, 2011.

[6] D. Luce., H. Raiffa. "Games and decisions". (in polish) PWN, Warsaw, 1996.

[7] A. Lewandowski A. and A., P. Wierzbicki eds. "Aspiration Based Decision Support Systems". Lecture Notes in Economics and Mathematical Systems. Vol. 331, Springer-Verlag, Berlin-Heidelberg, 1989.

[8] J. Lu, G Zhang and D. Ruan "Multi-objective group decision making: methods, software and applications with fuzzy set techniques" dl.acm.org. 2007.

[9] A. Łodziński A. „The reference point method applied to decision selection in the process of bilateral negotiations" Metody Ilościowe w Badaniach Ekonomicznych / Quantitative Methods in Economics, Warsaw, 2014.

[10] A. Łodziński A. "Interactive method of selection decisions under risk" Zarządzanie a Inżynieria Produkcji Management versus Production Engineering, (in polish) Kraków 2013.

DOI: 10.15439/2015F216
ACSIS, Vol. 5

# Modeling of software agents' societies in knowledge-based organizations. The results of the study.

Mariusz Żytniewski
University of Economics in
Katowice 1 Maja 50, 40-287
Katowice +48 32 2577277
zyto@ue.katowice.pl

Andrzej Sołtysik
University of Economics in
Katowice 1 Maja 50, 40-287
Katowice +48 32 2577277
soltys@ue.katowice.pl

Anna Sołtysik-Piorunkiewicz
University of Economics in
Katowice 1 Maja 50, 40-287
Katowice +48 32 2577277
apiorunkiewicz@ue.katowice.pl

Bartosz Kopka
University of Economics in
Katowice 1 Maja 50, 40-287
Katowice +48 32 2577277
bartosz.kopka@ue.katowice.pl

***Abstract*—Modern organizations, knowledge-based organizations in particular, seek new IT solutions supporting business processes and knowledge management realized by them. One of the solutions, postulated by the authors supporting the actions of such organizations may be computer software in the form of software agents, considered in our study in terms of software agents society. The purpose of this article is to analyze the results of authors' three-year study on the modeling aspect of the software agents society in knowledge-based organizations. The paper presents theoretical issues connected with the use of knowledge management systems in organisations, partial results of interviews with developers of agent solutions in Poland, a proposal of a methodology for designing agent societies, elements of a developed prototype of an agent solution and findings of qualitative research in the area of usability of software agents.**

## I. Introduction

DEVELOPMENT of organization's theory for modern forms of management of the company's action results in that solutions need to adapt their structure and functionality to meet the specific needs of the organization. One example is to look at the management-oriented approach to knowledge, which is seen as an important organization's resource. Such knowledge, perceived in the literature as an overt or covert knowledge, is an essential part of business processes that such organizations implement. Information systems, which will support organizations aware of the knowledge processing content, should support on one hand its codification, so that the knowledge partnership could be stored as a resource system used by the organization and its employees, and on the other hand the methodology to support the modeling of such systems should support its codification changing tacit knowledge into overt. Solutions, which according to the authors have the indicated characteristics, are agent systems.

The purpose of this article is to analyze the results of the authors' three-year study on the modeling aspect of software agents society in knowledge-based organizations. The study conducted by the authors, initiated in 2012, on the modeling of software agents' society in organizations based on knowledge, were focused on the search for forms of use of software agents in the context of their use in modern organizations and to determine whether these solutions, aided by semantic knowledge representation, contribute to the improvement of business processes. Due to its complexity, first an analysis of the literature in this area was conducted in order to arrange a society of agents in the registration system used in organizations [1] [2] [3]. Further the IT companies, which offer agent-based solutions, have been asked to indicate their main problems of modeling and implementation [4] [5]. As a result, studies on currently used multi-agent platforms [6] and agent systems design methodologies [7] allowed us to propose a modeling methodology of software agents' society [8] [9] [10], it's possible architecture [11][12] and to assess their impact on organizations [13][14][15]. As part of the qualitative research, indepth interviews and research experiments were used. The presented issues have been addressed in chapters.

Chapter 1 and 2 will present the introduction to the theory of knowledge-based organization and knowledge management. In Chapter 3 the results of the 2013 qualitative research in the form of in-depth interviews with companies forming agent-based solutions on the Polish market will be given. Later in Chapter 4, the methodology of modeling software agents society will be shown, focused on the organization's knowledge for agents specification. Chapter 5 will depict partial results of research, carried out in 2014, related to the software agents' usability analysis in promoting knowledge about the organization and the processes occurring in it.

## II. Knowledge Management in Organizations

The concept of knowledge management (KM) was developed to discover tools and methodology of management of knowledge, which was described as one of classical factor of production with land, labor and capital by Drucker [16] . The term of knowledge management is one of the most promotes an integrated approach to identifying, capturing, evaluating, retrieving, and sharing all of an enterprise's information assets. These assets may include databases, documents, policies, procedures, and previously uncaptured expertise and experience in individual workers [17].

Knowledge management system (KMS) is dedicated to help organization to meet its goals and to increase its effectiveness. The literature review shows that there are a multiple definitions of knowledge management system have been proposed in the literature, and debates about this concept have been expressed from a variety of perspectives and positions [18][19][20]. Also there are some models of life cycle knowledge management in organization [21][22].

978-83-60810-66-8/$25.00©2015, IEEE

Information and communication technologies (ICT) may play an important role in effectuating the knowledge-based view of the organization to manage the knowledge it possesses [21][17]. KMSs are technologies that support knowledge management in organizations, specifically, knowledge generation, codification, and transfer. Nowadays the impact of modern ICT (interactive communication channels, agent oriented technologies, etc.) in the company due to developing of Web 2.0/3.0, i.e., social media, blogs, micro blogs, forums, wikis, and others, make the impact in knowledge-based organization.

Knowledge-based organizations understand the importance of knowledge in the process of creating a competitive edge and focus on creating value added based on an effective use of knowledge [23]. ICT solutions focused on the aspect of supporting. Such organizations should support business processes that take place in them in the area of creating, processing and sharing a contextual knowledge about them. This results from the fact that knowledge-based organizations focus not only on business processes but also on knowledge management processes which should be treated in such organizations equivalently. Nowadays knowledge management system is facilitated by Web-based ICTs. It is worth underlining, that the majority of companies use well known ICTs, for example: e-mails, online surveys, social networks, Internet forums, business blogs, comments posted on a producer website, business (specialized) portals, online price comparison [24] [25].

The main tools of Web 2.0 allow users to share content with each other and collaborate online. Web 2.0 features are commonly known as postulates, namely the idea that describes the people-oriented approach [26]. The next generation of the Web is Web 3.0 [27], which is no longer focused on the people but on the machine. The idea of Web 3.0 requires not only data, but also the metadata, which is information about the data; it is reflected postulate transformation into a computer network database. This allows data binding on the Internet, taking into account the aspect of meaning, differentiating the data by the machine with the identical record, and also carried out on these data inference. It is an extension of the existing web, but has better connections between the blocks of information by defining the importance of accurate data available. As a result, programs that read the information, called agents, can analyze data, understand their meaning and find the relationships between them, and perform complicated tasks assigned by the users [28]. It is the main reason of implementation Web 3.0 into knowledge management systems in organization. One of the solutions, that can help such organizations, are software agents societies which can support the different stages of knowledge management systems life cycle.

The study shows the kind of models of agent supported organizations [1]:

- Information allocation model - an agent model refers to the way information flows between the organization and its environment, and additionally the influence of the information on the organization, using a software agent.
- The presence of authority's model - the participation of agents as the authorities in a decision making process relied on two features: modularity and decentralization.
- Organizational norms and culture model – an agent's behavior depends on the organization's historical factors which are contained in the organization's norms and culture,
- Motivating model – the human factor can be subjected to various influences which in the case of the use of agent-based solutions come down to a certain decision imperative of an agent.

There are different functionalities of knowledge management in such a kind of agent supported knowledge-based organization, e.g. in business context, health care context [29], etc. In the context of health care, multi-agent software (MAS) may play an important role in effectuating the knowledge-based view of the e-health organization by enhancing the capability to manage the knowledge it possesses [30] [31]. There is the method of evaluating MAS in e-health knowledge management system. The model of multi-agent knowledge management system is based on [32]:

- knowledge creation about the user,
- knowledge sharing of the presented problem,
- contextual knowledge about the course of the conversation during knowledge distribution,
- knowledge application in the organization.

There is a diversity of areas in medical industry and health care systems that could benefit from systems based on agent technology (especially MAS) [33] [34]: 1) systems diagnosing diseases, 2) systems that recommend treatment, 3) patient history examination systems, 4) the support of palliative care units.

## III. Program Agents' Societies And Problems With Their Implementation

One of the metaphors of embracing agent technologies is to look for purposeful solutions in terms of software agent society [35]. Agent societies used in knowledge-based organizations operate in an environment of ubiquitous communication and are usually aimed at supporting the organization by supporting the processing and distribution of information and knowledge using semantic knowledge representation mechanisms. In the process of developing agent solutions supporting the improvement of business processes in knowledge-based organizations a variety of approaches, depending on their architecture and construction, are used. The use of semantic knowledge representation mechanisms requires that, at the design stage of the system, methods for acquiring and processing knowledge necessary for the operation of agent solutions are set out.

Conducted in 2013 qualitative research have been realized in the form of a series of in-depth interviews with companies which are the creators of agent solutions on the Polish market [4][5]. A dozen of companies were invited for the study but only 5 of them decided to participate.

Studies have shown that the solutions used by respondents overwhelmingly support the human - computer interaction in the context of their possible use as an element of knowledge management systems, and thus can be considered in the context of the interface agents with their own codified knowledge base, associated with organizations' information systems and actively participating in the ongoing business processes. As part of the study different manufacturers' approach towards agent solutions supporting the improvement of business processes in organizations based on knowledge have been analyzed and compared. Among the solutions offered to the market in the use of agent technologies in the sphere of supporting business processes in organizations managed by knowledge, the vast majority are so-called "virtual advisors" showing, in accordance with the adopted typology, lowest level of socialization [30]. One of the important issues raised in the context of the study was to identify the workshop used by vendors of agent solutions on the stage of their development. In this context, the focus was mainly concentrated on methodologies of creating agent solutions. Methodologies used in the construction of agent systems, what was highlighted by the creators of solutions examined, are focused mainly on their architecture, but only to a small extent allow the modeling of system knowledge. There are no proven solutions that are using a methodical approach to the modeling of knowledge representation semantic mechanisms to the needs of agent structures. Respondents confirmed the heterogeneity of methodologies appropriate to the solutions they use, the design of which requires, on one hand the reference to the issues of software development methods, and on the second engineering theory of knowledge, which is required in the context of agents' knowledge base modeling. In addition to the problems of methodology, the respondents pointed to the need for finding methods to assess the impact of the solutions they create on the organization and its environment operation.

Despite differences in the details of the used methodologies design, or the tools used, companies generally agree on the stages of agent-based solutions implementation. In simple terms it can be assumed that this process is based on four main processes. The first one is to analyze and gather information from a user under which the knowledge an agent will have will be worded. Since the information comes from a variety of sources such as individual interviews, search for paper and electronic documents related to the user, and acquired knowledge is not always codified in clear and understandable manner, it is necessary to systematize this information. The next step is to design a model of this knowledge, systematization and arrangement in a certain and clear structure for the agent which will enable it to recognize the thread and giving the user the right answers. The next step is to implement the agent system. The last stage of testing the system is usually the user's input and feedback concerning its functioning. Feedback information from the customer allow us to evaluate whether the knowledge, which was introduced to an agent is correct, if something has been omitted, supplemented, whether the range or essence of knowledge was in some way changed. Then this knowledge should be updated. However, despite applying own methodologies the companies base on proven tools based on UML modeling structures to facilitate knowledge sets, allowing for structuring knowledge to describe some of the relations in the knowledge base, the design of systems architecture of agent.

Virtually all respondents covered by the conducted survey declared that the process of developing agent solutions in their company is more or less consistent with the outlined above methodological scheme. However, even if any particular methodology was not specified, it does not mean taking chaotic and random actions, but consistent common-sense approach, using necessary resources. Most differences in the used approaches are declared in the process of knowledge acquisition.

After implementation of the solution manufacturers attach great importance to feedback from the customer, confirming the correctness of topicality and completeness of the knowledge that was introduced to agent.

With the increasing number of implemented agent systems and their complexity merits, there is a natural need to systematize the knowledge concerning the activities undertaken and their standardization. The use of established methodologies in the implementation of solutions is not disputed by principle, but the very process of implementing the chosen methodology into force is not carried without complications. Among the fundamental problems that arise in the implementation of methodologies indicated by the respondents two groups were dominating:

The first group of problems of objective nature related to the need of increasing the workload associated with the introduction and updating adopted methodological solutions and methods and tools for their implementation. Even recognized methodologies, including a set of procedures, are subject to constant alteration forced by changing organizational and economic conditions resulting in assembly and aggregating new knowledge. According to the respondents implementation of the new system, as well as an agent system, often creates necessity to abandon the previously developed business process execution mode. According to the respondents, implementation and then maintaining particular solution require additional labor, which generally is not enthusiastically accepted by both the contractors of project tasks and team management

The second group of difficulties are the unpredictable problems subjectively linked to specific persons associated with the solutions' life cycle. Indicated by the respondents fear of change, loss of autonomy, wont to existing, not always fully formalized mode, the impression of redundancy of seemingly unnecessary documents required for the execution of implementation makes it difficult to maintain and care for the implemented solution.

A very important aspect in the context of the use of approved methodologies for projects involving complex, partly autonomous structure of software agents society, is their specificity and complexity that result in multiplication of methodological problems encountered in the implementation of individual solutions. At the same time manufacturers of commercially available software emphasized the lack of experience while working on more complex systems. None of the respondents did not declare directly of the implementation of solutions involving cooperation of more agents within a specific software agents' society.

Depicted here problems caused, that the authors' study have been directed to the aspect of modeling agents' society in the context of the methodological design aspect of such solutions and to develop a method to assess the impact of software agents on the environment in the context of the human - computer interaction. For this purpose, it became reasonable to conduct a comprehensive analysis of the available methodologies and propose the best possible methodology allowing for the implementation of software agents' society. The specificity, diversity and complexity of software agents' society designed to stimulate organizational processes in an organization managed by knowledge, forces the establishment of a transparent and universal model. Options include the use of ready methodologies developed by the companies with years of experience in developing agent solutions, or to create a proprietary methodology based on own experience, supplemented with elements taken from corporate methodologies. To conclude this aspect of research, three main aspects should be noted, which according to the authors should support the defined methodology, supporting the implementation of the agent system in organizations. First is the focus on supporting the business processes modeling in organizations. Its purpose is to better understand business conditions for which the organization's agent system is created by the designers and future users of the solution. Second aspect is the agents society architecture modeling which will be aiding this business process. Targeting organizations for the classic methods of semiformal defining the architecture of the system in the form of UML causes that in case of modeling agents' society it should also be used. The final element is the modeling of knowledge that must be considered in the context of the business process involving people and software agents. Methodology applied here should support the codification of knowledge in a way that allows its understanding and application by both: the users and agents. This aspect is crucial in the context of business process integration and knowledge management systems, where the task of developed methodology should be an indication of the use of codified knowledge in the framework of business processes in the organization.

## IV. Methodological Aspects of the Software Agents' Societies Designing

Conducted interviews indicated that currently created agent-based solutions offered by the companies relate to individual agent solutions. One of these aspects, that caused the lack of implementation in the area of multi-agent systems, was to identify the problems with the integration of knowledge abstracted within the agent system. Therefore, one of the research aspects was to develop proposals for taking up modeling methodology of software agents' society in the context of highest level of agents' socialization, which will focus on the aspect of system's semantic knowledge codification methods and business processes organization. In order to develop such methodology a number of multi-agent platforms [6] and methodologies [7] available in the literature and methods of agent design solutions [11] were tested. From the modern organizations' point of view, where organizational knowledge is a key aspect of system design, an indication of how multi-agent modeling environment with a strong focus on knowledge of the system becomes necessary, which the methodologies presented here do not show. None of analyzed methodologies [7] fully define organization's ontology, social relationships. Only three do this in limited way where designer is usually able to define only concepts of agent ontology. Also the mechanism of agent's interaction with the environment is not well realized by analyzed methodologies. Comparative analysis, at the stage of the assumptions related the to design of multi-agent system indicated, that the agent society, through the used methodology, will have limited functionality. The result of the study was to offer the methodology shown in chapter [8].

This methodology was created as a combination of software agents' society design good practice, ontologies design methods and BPMN notation used for the purpose of analyzing the requirements for the created agents' society in the context of the organization it's supposed to support.

The proposed methodology consists of 8 stages, which include:

1. Analysis and development of business process
   - 1.1. Specification of organizations involved in the process and the posts performing the tasks.
   - 1.2. Determination of relationships inside the organization. At this stage the relationship is defined within the organizational structure that supports the system. In case of an organization it is a structure linking the different departments and the process' participants positions.
   - 1.3. Defining the rules of starting and ending the process.
   - 1.4. Diagnosing the business process' tasks.
   - 1.5. Diagnosing the business process' events.
   - 1.6. Defining the conditions governing decision gates.
   - 1.7. Determining the extent of agent's support of a specific task (realization of tasks, assisting the task, none)

2. Resources' identification of in agents' society setting
   - 2.1 Identification of inputs and outputs of the main task
   - 2.2 Identification of resources in the form of services or external data

3. Analysis of the roles and responsibilities of agents' society
   - 3.1 Defining the tasks carried out in the agent society
   - 3.2 Defining the roles of agents in the system
   - 3.3 Diagnosing emergency situations (events)
   - 3.4 Defining the inputs and outputs based on events

4. Determining the hierarchical structure of the relationship inside the organization

4.1 Reference of the organizational structure with the main tasks carried out by agents
4.2 Determination of organization's internal relationships within the agent society

5. Determination of the extent of agent societies' knowledge
   5.1 Identification of knowledge range of agent society
   5.2 Identification of the resources provided by agent societies

6. Preliminary definition of agents' internal architecture
   6.1 Determining agent's classes
   6.2 Assigning agents' classes to roles
   6.3 Assigning agents' classes to resources

7. Essential definition of the agents' internal architecture
   7.1 Agent knowledge specification
   7.2 Defining agent's behavior

8. Designing the interaction between agents

The proposed approach for agents' society modeling is considered in terms of the heterogeneous construction agents' societies and determines the combination of best practices for agent solutions modeling to support business processes within the organization's information systems. In particular, this methodology is dedicated for knowledge-based organizations through its focus on modeling of the organization's knowledge using semantic mechanisms of representation.

The proposed methodology has been developed upon the experience regarding the developed prototype of agent-based solutions supporting the operation of the organization and developed in the context of building solutions supporting the interaction of users within the business processes in which they participate [12].

Depicted in the [12] prototype of agents' society was used to evaluate elements of the methodology and pointed to possible applications of agents' society to assist the interaction processes between users.

For the purpose of prototype solution realization BPMN process models were paid and extending the BPMN notation was proposed by a number of new artifacts used in the agents' society modeling process, based on the methodology presented previously. Table 1 present the main artifacts.

As a result, it became possible to develop elements of a modeler supporting the codification of knowledge about processes which involve users in accordance with BPMN notation and allowing to define the knowledge resources of the organization. Figure 1 presents such example.

Elements of the proposed methodology also served to develop a tool that aids software agents' usability analysis in the process of sharing knowledge in the organization's environment. The approach proposed herein offers the following advantages:

- Extending currently used standards for describing business processes to include sources of knowledge that supports the performance of users' tasks (in the context of the process, place and time).
- Enabling direct integration of organisational knowledge within any business processes taking place in an organisation within the scope of the process in which this knowledge should be used and the task that it supports.
- Automating processes of assessing the functioning of knowledge management systems in terms of their usefulness in supporting business processes.
- Generating new organisational knowledge at the interface of business processes and knowledge management.
- Using semantic mechanisms for knowledge description for easier integration of possessed knowledge with internal organisational knowledge.
- Independent operation from used IT solutions and enabling integration of any knowledge management systems and a process-oriented solution.

In order for agent systems to support business processes, it is necessary to develop tools facilitating the evaluation of an agent's usability in the process of performing users' tasks.

TABLE 1.
PROPOSED ARTIFACTS

| | |
|---|---|
| 👤 | Interface agent - designates possible applications of an interface agent as an element supporting the performance of a specific task by the user. It allows the user to go into the mode of evaluation of an agent's usefulness and use it to support the user's actions. |
| 👥 | A multi-agent system - designates a possible application of a multi-agent system to substitute the user or prepare a specific knowledge resource that will be necessary in the decision making process. |
| 🗄 | Knowledge resource - designates a specific knowledge resource in knowledge portal or the Internet that can be indicated to the user. It can be a document, web service, URL identifier. |
| 🧍 | Consultant - designates a specific person who has the relevant knowledge about the performance of this task. |

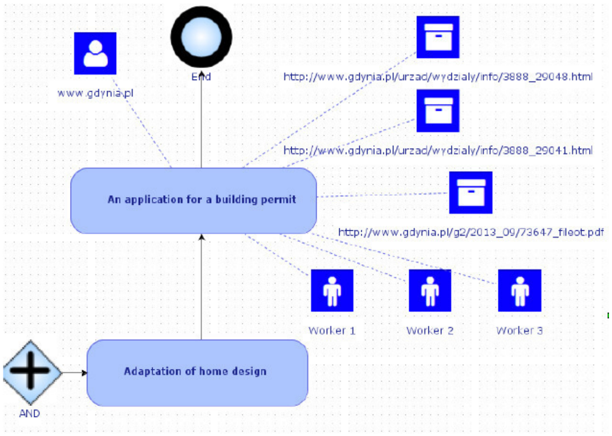

www.gdynia.pl
End
http://www.gdynia.pl/urzad/wydzialy/info/3888_29048.html
http://www.gdynia.pl/urzad/wydzialy/info/3888_29041.html
An application for a building permit
http://www.gdynia.pl/g2/2013_09/73647_fileot.pdf
Worker 1
Worker 2
Worker 3
Adaptation of home design
AND

Figure 1. Proposed extension of BPMN notation

## V. Software Agents' Usability

As part of the conducted qualitative research it was necessary to determine how software agents' societies improve business processes in organizations based on knowledge. On the basis of the study, creators of agent solutions in Poland were diagnosed and a test method allowing for evaluation of the usefulness of agents in the context of human - computer interaction was proposed - AUKP - Agent Usefulness and Knowledge Propagation analysis method (figure 2).

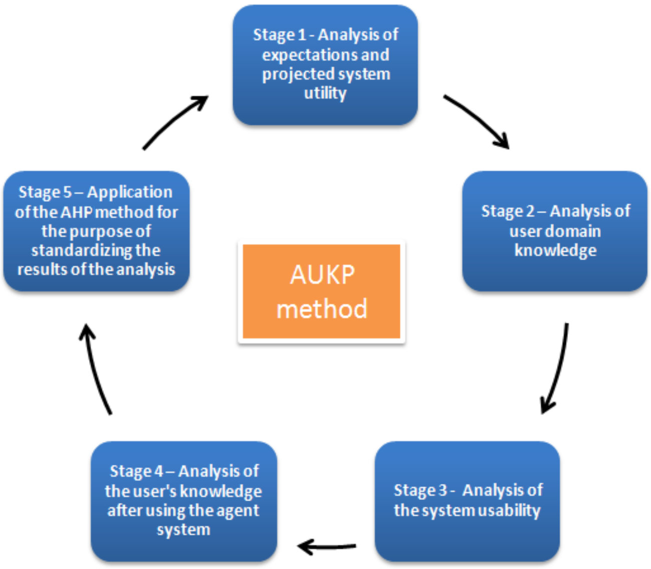

Stage 1 - Analysis of expectations and projected system utility
Stage 2 – Analysis of user domain knowledge
Stage 3 - Analysis of the system usability
Stage 4 – Analysis of the user's knowledge after using the agent system
Stage 5 – Application of the AHP method for the purpose of standardizing the results of the analysis
AUKP method

Figure 2. Agent Usefulness and Knowledge Propagation analysis method

During the tests of usability of software agents implemented in organizations, the analysis was conducted in the following stages [15]:

- Analysis of expectations and projected system usability. Aim: determine the expectations of the users in relation to the agent system and its functionality. Proposed method: research survey analyzing the significance of the basic indicators of the system usability.
- Analysis of user domain knowledge. Aim: determine the user's base knowledge in terms of the domain aided by the agent system. Proposed method: survey of knowledge which the user obtains as a result of working with the agent system.
- Analysis of the system usability. Aim: determine the values of the specific indicators of the assessment of the system usability for the user and the organization. Proposed method: direct analysis of the agent system's operation.
- Analysis of the user's knowledge after using the agent system. Aim: to determine the user's knowledge in the field supported by the agent system after using the system. Proposed method: survey of knowledge obtained by the user as a result of working with the agent system (as in stage 1).
- Application of the AHP method for standardizing the results of the analysis. Aim: standardization of the results with regard to users' expectations for a comparative analysis of agent systems. Proposed method: application of the AHP method based on the results obtained in stage 1 and stage 3.

First, an analysis of previously conducted research on the human - computer interaction in the context of agent usability testing environments was performed. These studies have shown objectivity of examining the usability in terms of the agents' impact on the users. The assumptions concerning the developed method and the results of the conducted experiments are discussed in more detail in the works [13][14][15]. In these studies researchers assumed, that in accordance with the concept of utility, it is necessary to refer to the analysis of effectiveness, efficiency, satisfaction and propagation/dissemination of agent system's knowledge. For this purpose, the process evaluation model for specified factors within the HCI theory was developed and 102 research experiments were carried out, which aim was to evaluate the operation of three software agents. The results confirmed, that in the group of users and agents can be indicated, that the agent-based solutions contribute to the improvement of business processes in which they participate by improving customer satisfaction and propagation/dissemination of knowledge among users regarding the organization and the processes, in which the organization and the business process recipient participate.

The final results was given: analysis of the general level of satisfaction shows that for users of software agents that were used for experiment satisfactory level was above 0,6; also all agents reached knowledge propagation indicator above 0,75.

Satisfaction and knowledge dissemination assume values from 0 to 1, where 1 is the highest indicator of satisfaction and the highest level of knowledge propagation/dissemination.

Next two parameters (main effectiveness and partial effectiveness) assume values from -1 to1. Values above 0 mean that the system is effective, whereas values below 0 – lack of the agent's effectiveness. All agents gained measures of effectiveness (main) >0 and effectiveness (partial) >0, which can be interpreted as the possibility of implementation of indicated main goals by users and shows that the answers generated by an agent were correct.

The last parameter indicates the number of agent objectives performed per a minute. The conducted research has also showed that agent C (with the result of 3.5 of an objective per minute) achieved the highest performance compared to agent B (3 objective per minute) and agent A (2.5 goal per minute).

All aggregate results shown here revealed, that in the case where it is possible to identify high efficiency of agents and their productivity, users indicated high levels of satisfaction and knowledge gain. What is characteristic, increased productivity of agents and their effectiveness, influenced the increase of user's satisfaction and the amount of knowledge they acquired.

## VI. Conclusion

As indicated in the article, the theory of modeling the software agents' society is a complex research problem, particularly in the context of their use as a solution supporting the selected type of organization. Developing the assessment methods of the agents' impact on the processes taking place in organizations and supporting their modeling methodologies, required to refer not only to the current state of research in this area, but also to focus on the essence of organizations for the purpose for which they will be developed. The developed methodology, from the organization's point of view, supports the codification of knowledge that is contained herein and contributes to a better understanding of business processes that take place within. Such an approach improves the process of organisational knowledge integration within the modelled society of software agents and allows the relationships of the company's organisational structure to be mapped within a specific society of agents. The next step in the development of this methodology may be extending it by elements of modelling trust and reputation within the agent society being developed.

The proposed assessment method allows for indication whether the influence of agents on the processes that are related to the operation of the organization is visible. The conducted experiments proved that the currently used agent systems in organizations improve ongoing business processes, allow for the improvement of customer satisfaction and help to propagate the knowledge about the organization and its processes. The proposed method, which is based on the assumptions of agents' interaction with users, can be used with any software agents that have an interface ensuring communication with the user. The proposed approach, in accordance with the accepted criterion of usability, indicates validity of an analysis of knowledge propagation and satisfaction indicators, by which the impact of agents on business process participants can be determined, as well as effectiveness and performance indicators, which allow the agent itself to be assessed. The experiment examined the interaction between the human being and computer, but the developed indicators of agents' effectiveness and performance can be also used in the process of interaction between agents in a given society.

## Acknowledgment

The project was financed from the funds of National Science Centre 2011/03/D/HS4/00782.

## References

[1] A. Sołtysik-Piorunkiewicz, M. Żytniewski "Software Agent Societies for Process Management in Knowledge-Based Organization" [in:] Proceedings of the 14th European Conference on Knowledge Management, Volume 2, ACPI, UK, 2013, pp. 661-669

[2] M. Żytniewski "Application of the software agents society in the knowledge management system life cycle " [in:] Cognition and Creativity Support Systems ed. M. Pańkowska, S. Stanek, H. Sroka, Publishing House of the University of Economics in Katowice, 2013 pp. 191-201.

[3] A. Sołtysik-Piorunkiewicz "The development of mobile Internet technology and ubiquitous communication in a knowledge-based organization" [in:] The Online Journal of Applied Knowledge Management (OJAKM), Volume 1, Issue 1, 2013, pp 29-41

[4] M Żytniewski, R. Kowal, A. Sołtysik "Creation of Software Agents' Society from the Perspective of Implementation Companies. The Advantages of Their Use, the Problems of Construction and Unique Features" [in:] Research Papers of the Wroclaw University of Economics in series "Business Informatics", Publishing House of Wrocław University of Economics 3(29), 2013, pp. 162-171

[5] M. Żytniewski, R. Kowal, A. Sołtysik "The outcomes of the research in areas of application and impact of software agents societies to organizations so far. Examples of implementation in Polish companies" [in:] "Annals of Computer Science and Information Systems, Volume 1" Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, pp 1181 – 1187

[6] M. Żytniewski, M. Klement "Analiza porównawcza wybranych platform wieloagentowych" [in:] Informatyka 2 Przyszłości 30 Lat Informatyki Na Wydziale Zarządzania UW, ed. J. Kisielnicki, W. Chmielarz, T. Parys, Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, 2015

[7] M. Żytniewski "Comparison of methodologies for agents' software society modeling processes in support for the needs of a knowledge-based organization" In Procedings of 11th International Conference "MULTIMEDIA IN BUSINESS AND MANAGEMENT", Institute of Management Information Systems Management Faculty Częstochowa University of Technology, 2015

[8] M. Żytniewski "Modelowanie systemów agentowych wspomagających organizacje oparte na wiedzy", [in:] "Technologie agentowe w organizacjach opartych na wiedzy" (praca zbiorowa pod redakcją M. Żytniewski), Wydawnictwo Naukowe Uniwersytetu Ekonomicznego w Katowicach, 2015 (after positive review)

[9] M. Żytniewski "Wprowadzenie do teorii społeczności agentów programowych oraz ich zastosowania w organizacjach opartych na wiedzy", [in:] "Technologie agentowe w organizacjach opartych na wiedzy" (praca zbiorowa pod redakcją M. Żytniewski), Wydawnictwo Naukowe Uniwersytetu Ekonomicznego w Katowicach, 2015 (after positive review)

[10] M. Żytniewski. M. Klement "Trust in software agent societies" [in:] The Online Journal of Applied Knowledge Management (OJAKM), 2015 (after positive review)

[11] M. Żytniewski, R. Kowal "Using Software Agents to Enhance the Functionality of Social Knowledge Portal" [in:] "Business Information Systems Workshops" ed. Witold Abramowicz, Springer Lecture Notes in Business Information Processing vol 160, pp 23-34

[12] M. Żytniewski "Integration of knowledge management systems and business processes using multi-agent systems" [in:] Proceedings of the Cooperative Online Organizations conference (presented at the AAMAS'15 Workshop), Turkey, 2015

[13] M. Żytniewski, B. Kopka "Indicators of software agents usability and ergonomic" In Procedings of I. MEDIAL INTERNATIONAL SCIENTIFIC CONFERENCE OF THE SERIES "Decisions in situations of endangerment", The Journal of Science of the Gen. Tadeusz Kosciuszko Military Academy of Land Forces, Wrocław 2015 (after positive review)

[14] B. Kopka, M. Żytniewski "The system ergonomics and usability as measurement of the software agent impact to the organization", in Proceedings of Advances in Ergonomics In Design, Usability & Special Populations ed. F. Rebelo, M.Soares, Published by AHFE Conference, 2014, pp. 21-34

[15] M. Żytniewski, B. Kopka "The proposition of agents' usability analysis method based on an analysis of Polish enterprises" [in:] Proceedings of the Human Agent Interaction Design and Models conference (presented at the AAMAS'15 Workshop), Turkey, 2015

[16] P. Drucker, Post-capitalist society. Harper Business, New York 1993

[17] E. Turban, D. Leidner, E. McLean, M. Wetherbe, Knowledge management, in Information technology for management: transforming organizations in the digital economy, John Wiley, New Jersey 2006, pp.: 365-405.

[18] I. Nonaka, H. Takeuchi, Knowledge-Creating Company. Oxford University Press, New York, 1995

[19] K.M. Wiig, Knowledge management foundations. Schema Press, Arlington 1993

[20] M.E. Jennex, Knowledge Management: Concepts, Methodologies, Tools and Applications, IGI Publishing Hershey, PA, USA, 2009

[21] T. Davenport, & L. Prusak, Working knowledge how organizations manage what they know, Boston: Harward Business School Press. 1998

[22] R. Van der Spek, & A. Spijkervet, Knowledge Management: Dealing intelligently withknowledge. In J. Liebowitz, & L. Wilcox (Eds.), Knowledge management and its intergrative elements. CRC Press, New York 1997

[23] W.M. Grudzewski, I. Hejduk, “Knowledge management in enterprises”, Difin, Warszawa 2004

[24] E. Ziemba, M. Eisenbardt, “Aktywności prosumenckie z wykorzystaniem technologii informacyjno-komunikacyjnych w świetle badań bezpośrednich” [Research on ICT application towards prosumers’ activities]. In A. Nowicki & D. Jelonek (Eds.), Wiedza i technologie informacyjne w kreowaniu przedsiębiorczości, 101-113. Częstochowa: Publishing House of Wydział Zarządzania Politechniki Częstochowskiej, 2013

[25] E. Ziemba, J. Wielki, The use of corporate portals in managing knowledge on entities operating in the electronic space. In S. Wrycza (Ed.), Proceedings of the Seventh International Conference on Perspectives in Business Informatics Research BIR’2008, 143-157. Gdansk: Gdansk University Press. 2008

[26] A. Sołtysik-Piorunkiewicz “The Telecom Business Strategies: a Comparative Study of Corporate Blogs”. MIDI Warszawa 2014

[27] J. Markoff, Entrepreneurs see a web guided by common sense, New York Times, 2006

[28] J. Gołuchowski J. „Wprowadzenie do inżynierii wiedzy”, Difin SA, Warszawa 2011

[29] M. Furmankiewicz, A. Sołtysik-Piorunkiewicz, P. Ziuziański “Artificial intelligence systems for knowledge management in e-health: the study of intelligent software agents” [in:] Latest trends on Systems: 18th International Conference on Systems: Santorini Island, Greece, July 17-21, 2014, str. 551-556

[30] A. Sołtysik-Piorunkiewicz “Knowledge management impact of information technology Web 2.0/3.0. The case study of agent software technology usability in knowledge management system”, AIP Conf. Proc. 1644, 219, 2015, http://dx.doi.org/10.1063/1.4907840.

[31] A. Sołtysik-Piorunkiewicz, M. Furmankiewicz, P. Ziuziański “The method of evaluation of multi-agent software for knowledge management in e-health”. [in:] L. Kiełtyka, R. Niedbał (Ed.) Wybrane zastosowania technologii informacyjnych zarządzania w organizacjach, Monografia Nr 296, Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2015.

[32] A. Sołtysik-Piorunkiewicz, M. Furmankiewicz, P. Ziuziański „Artificial intelligence and multi-agent software for e-health knowledge management system” [in:] Business Informatics (Informatyka Ekonomiczna), Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Nr 2 (32) / 2014, s. 51-63.

[33] A. Sołtysik-Piorunkiewicz „Technologie mobilne w zarządzaniu organizacją opartą na wiedzy”, [in:] R. Knosala (Ed.) Innowacje w zarządzaniu i inżynierii produkcji, PTZP, Opole 2015

[34] U. Cortés, R. Annicchiarico, C. Urdiales “Agents and Healthcare: Usability and Acceptance”, [in:] R. Annicchiarico, U.C. Garcia, C.Urdiales, Agent Technology and e-Health, Birhauser Verlag, Basel, pp. 1-3, 2008

[35] M. Żytniewski „Rozwój koncepcji społeczności agentów programowych” Europejska przestrzeń komunikacji elektronicznej red. J. Buko, Zeszyty Naukowe Uniwersytetu Szczecińskiego, 2013

# 4th Workshop on Information Technologies for Logistics

THE main purpose of the workshop is to provide a forum for researchers and practitioners to present and discuss current issues concerning use of ICT in logistic applications (hardware and software). There will be also an opportunity for hardware integrators, software developers and logistics companies to demonstrate their solutions, as well as achievements, in different logistic systems.

## Topics

The topics of interest include but are not limited to:

- Innovations in information systems supporting logistics and its management (WMS, SCM, TMS, LIS, VMI, CRP, PLM, and others)
- Innovative technologies in warehouse management: RFID, Voice Picking, Image Recognition, Pick Radar, etc.
- Logistics process modeling, including influence of warehouse automatic
- Optimization of logistics processes:
    - optimal vehicle routing and management, boundary conditions
    - optimal picking routing (global optimization, fast search, collision prediction and prevention)
    - shared mobility systems
    - day-to-day dynamic traffic assignment models
    - effective methods of picking (multi picking, batch picking ect.)
    - relationships between picking efficiency and products decomposition in warehouse area
- Environmental protection (for example carbon-aware transportation)
- Artificial intelligence systems and decision support systems in logistics
- BI, data mining and process mining in logistics
- Quality management algorithms and methods
- Material Flow Theory and applications

## Event Chairs

**Gontar, Beata,** Uniwersity of Lodz, Poland

**Gontar, Zbigniew,** University of Lodz, Poland

**Pamuła, Anna,** University of Łódź, Poland

## Program Committee

**Balicki, Jerzy,** Gdansk University of Technology, Poland

**Banaszak, Zbigniew,** Warsaw University of Technology, Poland

**Bobkowska, Anna,** Gdansk University of Technology, Poland

**Bruzda, Jaonna,** Nicolaus Copernicus University, Poland

**Duran-Grados,** Vanesa, University of Cadiz, Spain

**Fosner, Maja,** Faculty of Logistics, University of Maribor, Slovenia

**Franczyk, Bogdan,** University of Leipzig, Germany

**Gontar, Beata,** University of Łódź

**Hartványi, Tamás,** Széchenyi István University, Hungary

**Korczak, Jerzy,** Wrocław University of Economics, Poland

**Langviniene, Neringa,** Kaunas University of Technology, Lithuania

**Lim, Ming K.,** University of Derby, United Kingdom

**Malavasi, Gabriele,** University of Rome, Italy

**Matulewski, Marek,** Poznań School of Logistics, Poland

**Montemanni, Roberto,** University of Applied Sciences of Southern Switzerland, Switzerland

**Patasiene, Irena,** Kaunas University of Technology, Lithuania

**Rakovska, Eva,** University of Economics in Bratislava, Slovakia

**Ricci, Stefano,** Sapienza University of Rome, Italy

**Seong, Park Jong,** Korea National Open University

**Shinkevich, Aleksej Ivanovich,** Kazan National Research Technological University, Russia

**Sitek, Pawel,** Kielce University of Technology, Poland

**Speranza, Grazia,** University of Brescia, Italy

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1613–1621

DOI: 10.15439/2015F215
ACSIS, Vol. 5

# Information Logistics as a Paradigm

Anna E. Bobkowska
Faculty of Electronics, Telecommunications and Informatics,
Gdansk University of Technology
ul. Narutowicza 11/12, 80-233 Gdańsk, Poland
Email: annab@eti.pg.gda.pl

*Abstract*—**This paper presents a paradigm-based approach to information logistics. The approach deals with extracting concepts specific to information logistics in categories typical to paradigms. The resulting description should be easily connected to complementary components which are based on other paradigms, e.g. business process management and information systems development. Empirical part aims at checking and enhancing features typical to information logistics. It is conducted with action research of applying information logistics paradigm in coordination of Erasmus+ program.**

## I. Introduction

INFORMATION logistics has been a topic of research for about three decades with more than hundred contributions nowadays [1], [2]. It has several definitions with focus on a variety of issues in different contexts of use. A common motive for information logistics is transfer of the right information to the right place in the right time. The main active research directions are user-demand information supply [3], [4], efficiency of information flow [5], [6], cross-functional supply of analytical information [7], [8] and process improvement via information flow [9]. About two dozens of organizations offer their information logistics products for real-life solutions. One can observe that information logistics in real-life applications cannot be separated neither from defining chains of information processing and delivery nor from automation of information processing by software systems. Many information logistics products are based on Information and Communication Technologies (ICT) and support Business Process Management (BPM). Therefore, there is a need to explore relationship between information logistics, business process management and software engineering.

The goal of this paper is to analyze information logistics from the perspective of paradigm theory. It deals with extraction of concepts specific to information logistics and connection of resulting descriptions to components of solution which use other paradigms. Fig. 1 presents a schema of paradigm-based approach to integrating information logistics to other components using meta-modeling technology.

This approach is motivated by the following reasons. First, information logistics, business process management and software systems are complementary components of real-life solutions. They focus on different aspects and can be captured more precisely in terms of paradigms. The notion of paradigm is related to extraction of basic concepts specific to a given area and it allows to avoid uncontrolled overlapping with basic concepts from other paradigms. Second, the use of paradigm-based approach makes a difference in viewing information logistics. It makes a shift from entire information logistics solution (which informally defines enterprise processes and uses information systems) to a kind of specific component which can be connected to components which are based on other paradigms. The benefit of such approach is in solving each class of problems in the most effective technology and then connecting them in complete solution of real-life problems. Third, information logistics paradigm constitutes foundations for formalization of the paradigm with the meta-modeling technology provided by Object Management Group (OMG), especially in terms of UML profiles. OMG Unified Modeling Language (UML) [10] defines object-oriented software development perspective and OMG Business Process Model and Notation (BPMN) [11] defines business process management perspective. OMG offers UML profiles as extension mechanism and defines a profile diagram for the purpose of defining new UML profiles [12]. With use of this technology, formalisms which define all these complementary paradigms are expressed with the same meta-meta-models, which provides a strong common ground for integration and interoperability. This research shows that information technologies can offer to logistics not only methods and environments for computations or methods of information systems engineering, but also benefits from applying advanced meta-modeling technologies.

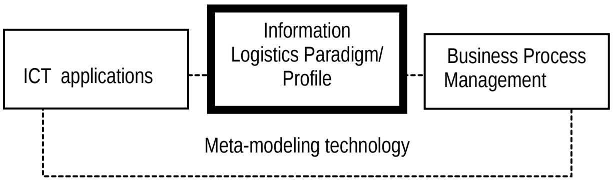

ICT applications
Information Logistics Paradigm/ Profile
Business Process Management
Meta-modeling technology

Fig. 1 A schema of integration of information logistics profile with other components of solution with use of meta-modeling technology

This paper is structured as follows. Section 2 presents and exploits paradigm-based approach. It explains the role of paradigms in computer science adhering to Kuhn's paradigms in philosophy of science, attempts to apply paradigms in computer science as well as multi-paradigm modeling approach and defines categories typical to paradigm description (paradigm template). Then, it describes analysis of information logistics definitions in literature made in order to capture core concepts and issues in information logistics. Section 3 contains description of action research in which features specific to information logistics are checked and en-

978-83-60810-66-8/$25.00©2015, IEEE

hanced when solving real-life problems in the area of coordination of Erasmus+ program. Section 4 makes an attempt to describe information logistics paradigm on the basis of both analytical and empirical parts. Section 5 presents conclusions and plans for further work.

## II. Information Logistics as a Paradigm

### A. Fundamentals of Paradigms

The popularity of the term paradigm, which has its roots in ancient Greak παραδειγμα meaning `pattern, examplar, example`, has been a contribution of a philosopher of science Thomas Kuhn, who explained incommensurability of theories in science by adhering to their diversified fundamentals. A paradigm in mature science stands for a set of fixed key theories, scientific language, values and metaphysical assumptions as well as procedures and instrumentation needed to solve practical problems [13].

The notion of paradigm corresponding to software has first appeared as paradigms of programming [14] with decreasing scope of paradigm as a side effect of moving the term of paradigm from science to technology. Then, it was elaborated in context of several reuse approaches to software development including design patterns, components, software architectures and frameworks [15]. Another use of this term can be found in multi-paradigm modeling approach [16], [17] which aims at dealing with heterogeneous solutions. The need for effective solutions has made a difference in viewing paradigm. It has replaced initial incommensurability of theories in science by the request for pragmatic integration of technologies which are based on different paradigms.

A paradigm-based approach is based on the following assumptions. Different classes of problems have effective solutions in technologies which use different paradigms. In order to develop software applications which deliver business value to customers there is a need to integrate technologies which are based on different paradigms.

When comparing paradigm-based approach with typical systems development the following differences can be identified. Typical development may be easier in simple cases whilst paradigm-based approach may require extra effort on interface definitions and collaboration between components based on different paradigms. On the other hand, solutions made with paradigm-based approach are easier to maintain especially when integral components need to be updated. It can deal more effectively with heterogeneous cases. Additionally, it is easy to connect to newly added components. It can provide more precision with traceability to sources. The challenge is in assessing global features of solutions. As this is a quite new approach, few experience reports from its applications are available.

### B. Questions and Methods

In order to capture information logistics as a paradigm there is a need to define knowledge gathered in this field of study in categories specific to paradigms. Thus, the following questions can be posed for information logistics: What are its values and (metaphysical) assumptions? What are its scientific language and key theories? What is the class of problems that can be solved effectively in this paradigm? What are instrumentation and procedures, i.e. methods and technologies, needed to solve these problems? How to integrate it with technologies which use other paradigms in context of paradigm-based approach?

We assume that information logistics paradigm should be based on its strong connotations to logistics. The intuitive understanding of information logistics is a kind of logistics which is related to information instead of material goods and services.

In order to capture the essentials of information logistics we have searched for common themes in information logistics definitions and notions provided by both researchers and practitioners collected in literature overview [1], [2]. Number of appearance of a given common theme against total number of definitions analyzed is presented in brackets.

Then, we have conducted a cross-analysis of these common themes with themes in logistics management definition provided by the Council of Supply Chain Management Professionals in order to check whether the common themes are strongly related to logistics and thus they have appropriate connotations to logistics.

A related field of study is information management. Therefore, we have compared our approach with the approach presented in Information Management Body of Knowledge.

### C. Common Themes of Information Logistics

Common themes of information logistics include:

- right information (20/33) with more precise specification of what is right information in terms of content or format (8/33), way or channel (3/33), size (1/33), quality (6/33) and requirements, needs or demands (6/33);
- right time (20/33) with adhering to "just in time" philosophy (3/33) or efficiency (5/33);
- right place or location (16/33) with additional distinction to recipient or person (7/33) or target group (3/33);
- information flow (9/33) or information value chains (5/33);
- right cost or optimization of costs (6/33);
- two cases: information distribution and information about logistics of goods and services;
- operational activities (8/33) of gathering (2/33), selecting (2/33), production, receiving, processing, organization, optimization, storage (3/33) and distribution (6/33); focusing on dissemination rather than production and classification;
- managerial activities of planning (2/33), control and implementation (1/33);
- infrastructure in terms of network of suppliers-requestors (1/33) or information logistics infrastructure with information systems support (13/33);
- relationship to other fields (2/33), such as information management, knowledge management, communication management and business networks.

The above classification of themes may be biased in several ways. Definitions and notions may not cover full range of issues addressed by research results or products. The analysis was made on the basis of selection of issues in cited literature overview. There are probably implicit relationships between research results, research bodies statement of notion and information logistics product providers. Therefore, the generalization of the common themes coverage to quantitative coverage of themes in the area of information logistics may be unjustified. However, they reflect quite well essentials and issues in this area.

### *D. Cross-Analysis with Logistics Management*

Council of Supply Chain Management Professionals (CSCMP) provides the following definition of logistics management [18]: "Logistics management is that part of supply chain management that plans, implements, and controls the efficient, effective  forward and reverse flow and storage of goods, services, and related information between the point of origin and the point of consumption in order to meet customers' requirements. Logistics management activities typically include inbound and outbound transportation management, fleet management, warehousing, materials handling, order fulfillment, logistics network design, inventory management, supply/demand planning, and management of third party logistics services providers. To varying degrees, the logistics function also includes sourcing and procurement, production planning and scheduling, packaging and assembly, and customer service. It is involved in all levels of planning and execution-strategic, operational, and tactical. Logistics management is an integrating function which coordinates and optimizes all logistics activities, as well as integrates logistics activities with other functions, including marketing, sales, manufacturing, finance, and information technology."

Table I presents a comparison of the common themes in information logistics with related logistics management themes. It shows that in almost all categories some equivalents can be found although they have different focus. The common themes of information logistics focus on provision and distribution of right information at the right time to the right place. Explosive speed of information available via global ICT applications, the need for gathering crucial business information and chaos of information in organizations are the main motivations for such statements of expected effects. Less attention is paid to the core term of logistics which is supply chain (similar to information flow and information value chain). In fact, logistics definition contains a suggestion that right information, time and place should be placed in a broader perspective of flows which contribute to a more complex enterprise.

For many logistics operations one can find equivalents in information logistics although they have their specifics or smaller importance. Equivalents of transportation are distribution, dissemination, gathering and receiving depending on context. Warehousing corresponds to storage. Logistics network design and fleet management are related to infrastructure management. Supply/demand planning finds its implementation in information flow design. Similar to auxiliary production planning and scheduling may be production, processing and optimization. Thus, on operational level, the analogy is quite strong. Logistics management can broader horizons of information logistics management. Although managerial activities of planning, control, implementation as well as relationships to other fields have been identified in scope of common themes for information logistics, logistics management is involved in all levels of planning and execution, i.e. strategic, operational, and tactical and it covers integrative activities of coordination and optimization of all logistics activities as well as activities in related fields.

TABLE I.
COMPARISON OF COMMON THEMES IN INFORMATION LOGISTICS WITH RELATED LOGISTICS MANAGEMENT THEMES

| COMMON THEMES IN INFORMATION LOGISTICS | LOGISTICS MANAGEMENT THEMES |
|---|---|
| information flow or information value chains | a part of supply chain management |
| managerial activities of planning, control and implementation | actions: plan, implementation and control |
| two cases: information distribution and information about logistics of goods and services | efficient, effective  forward and reverse flow and storage of goods, services, and related information |
| right place (location) with additional distinction to person  or target group | between the point of origin and the point of consumption |
| right information in terms of content or format, way or channel, size, quality or requirements, needs or demands | in order to meet customers' requirements |
| operational activities of gathering, selecting, production, receiving, processing, organization, optimization, storage and distribution; focusing on dissemination rather than production and classification;<br>infrastructure in terms of network of suppliers-requestors or information logistics infrastructure with information systems support | typical activities: inbound and outbound transportation, fleet management, warehousing, materials handling, order fulfillment, logistics network design, inventory management, supply/demand planning, and management of third party logistics services providers |
|  | all levels of planning and execution: strategic, operational, and tactical |
| right cost or optimization of costs; right time with adhering to "just in time" philosophy or efficiency | integrating function which coordinates and optimizes all logistics activities |
| relationship to other fields, such as information management, knowledge management, communication management and business networks. | integrates logistics activities with other functions, including marketing, sales, manufacturing, finance, and information technology. |

### *E. Comparison to Information Management*

One could ask questions: What is the relationship between information logistics and information management? Don't they deal with the same problems? (It seems similar to questions: What is the relationship between logistics and resource management? Don't they deal with the same problems?) In our opinion, information logistics and information management represent different traditions of research which address overlapping problems and deliver overlapping results. Additionally, both terms are broad and abstract, which makes difficulty in comparisons. One can expect that internal inconsistencies in both fields are encountered. In this situation, we compare only proposed paradigm-based approach to information logistics with a concrete representation of information management  in Information Management Body of Knowledge (IMBOK) [19].

IMBOK addresses problem of "huge complexities in the territory where information technology, business and society meet". It divides information management to six knowledge areas: information technology, information systems, business processes, business information, business benefits and business strategy. It describes four types of information management processes:

- projects related to information systems development with use of information technology,
- business change which makes changes in business process and business information areas with information systems,
- business operations which lead to accomplish business benefits,
- performance management where business strategy and business benefits meet.

Table II presents a comparison of paradigm-based approach to information logistics with IMBOK. Although the terms information logistics and information management seem to have a similar meaning, both have a broad scope and both include several approaches. The comparison which was made, shows that the differences might be on the level of fundamental assumptions, such as goals, focus, scope and values. (Information logistics solutions which use a more holistic approach could have more similarities to information management.) In fact, this is an example of incommensurability of theories in science, as Thomas Kuhn would call it. To continue his thought: established scientific paradigms, committed groups of researchers and consistent research results are the signs of mature science.

One of the common themes includes information management on the list of fields to which information logistics has relationships. This relationship may take form of appropriate application of relevant theories and techniques in concrete cases of information logistics solutions.

TABLE II.
COMPARISON OF PARADIGM-BASED APPROACH TO INFORMATION LOGISTICS WITH IMBOK

| PARADIGM-BASED APPROACH TO INFORMATION LOGISTICS | IMBOK |
|---|---|
| Information logistics profile as a component to be connected with others | Framework to cover complexities on the edge of IT, business and society |
| Focus on extracting and formalizing specific features of information logistics | Focus on a big picture, i.e. "everything that has anything to do with information in business" |
| Starts with transfers of the right information at the right time to the right place with interface to business process management (including IMBOK business processes, information, benefits and strategy) and interface to ICT application (including IMBOK information technology and information systems) | Covers:<br>• information technology,<br>• information systems,<br>• business processes,<br>• business information,<br>• business benefits,<br>• business strategy |
| Value of external consistency with standards and easy update when they evolve | Value of addressing entire field |

## III. ACTION RESEARCH BASED ON COORDINATION IN ERASMUS+ PROGRAM

### *A. Motivation*

Information Logistics paradigm seems to be helpful in coordination of Erasmus+ program [20]. There is a need to define rules and procedures, but business process perspective is not sufficient for capturing all aspects of the reality. Many parties are involved with their specific information needs. Changes are made outside of control of coordinators. There are several regulations of different kinds which need to be satisfied. Long documents which sometimes require interpretations and informal rules applied in order to unify action in similar exceptional cases cause difficulty to clearly understand the rules for those who just occasionally are involved in the program. The perspective which suggests to pose questions: What is the right information? What is the right time? What is the right channel to get to a given target? is helpful in systematic elaboration of the right solution in this complicated situation.

### *B. Action Research Description*

The goal of the action research is to check common themes of information logistics in a practical case and to enhance them by extracting new features specific to information logistics. The objective of the actions is to improve information logistics processes during coordination of Erasmus+ program in terms of analysis and explicit definition of the needs for information transfers, their content, channels and feedback. It should lead to a higher level of satisfaction of the parties involved in the processes. As the author fulfills duties of Faculty Erasmus+ coordinator, practical benefits resulting from process improvement this area are expected.

It is expected that the common themes of information logistics drive attention to several useful features which can be analyzed for process improvement. After validation and enhancements, they should be included in paradigm description. Managerial tasks should help in building framework for management. Operational activities are expected to help in extracting stereotypes of activities. Important issues of analysis are how to develop infrastructure and how to optimize effort related to its construction.

In the next section, an attempt to capture complexity of actions with traditional techniques and some limitations of their application are presented. In the further part, information logistics issues are analyzed. The following issues are discussed in details:

- goals of information transfer,
- management of group dynamics,
- flexibility of processes depending on diversity and changes in areas controlled by external participants,
- promotion via different channels,
- the right information at the right time and in the right place.

The descriptions are made according to the following template:

- description of the issue under consideration,
- comparison to business process management (BPM) perspective, especially to elements of BPMN,
- examples from the area of actions.

The comparison to BPM is made because business process and chain of information flow are similar concepts. Furthermore, analysis of relationships between information transfer and business process management is important from the perspective of making relationships between complementary paradigms which must work together in order to deliver solution.

### C. An Attempt to Capture Complexity of Actions

In order to deal with complexity of actions related to coordination and management of Erasmus+ program, several techniques available for project managers or business analysts can be applied with quite satisfactory results, e.g. stakeholder analysis, document analysis, organization modeling, process modeling, checklists, regulations and checking their compliance with framework regulations, metrics and key performance indicators. However, due to large diversity of cases, large dynamics and coordinative specifics, the application of these methods reveal some limitations. They are useful at higher level of abstraction, but it seems to be impossible to cover and trace all details of changes in a reasonable time. The resulting need for flexibility reduces the applicability of methods which tend to unification and formalization.

Fig. 2 presents stakeholders who require coordination during the Erasmus+ exchange actions. Erasmus+ agencies at European and national level as well as university authorities provide a variety of rules and regulations and then require documents and several kinds of reports. They influence mainly strategy and tactical levels. Other stakeholders are interact at operational level. The diversity of rules at partner university (including national regulations, university processes and on-line systems) has impact on all processes of exchange. It is reasonable that the work on information logistics is based on the results of stakeholder analysis.

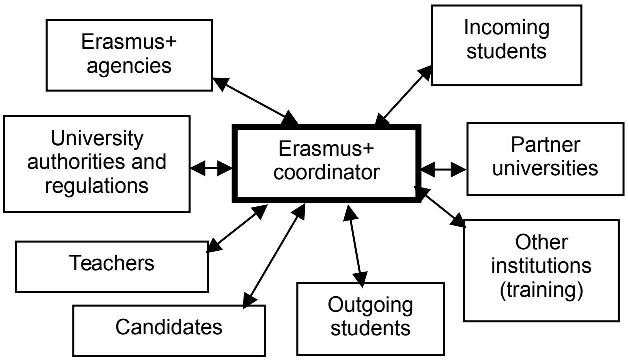


Fig. 2 Stakeholders who require coordination during the Erasmus+ exchange actions

High-level business process descriptions are useful for organizing and ordering the actions. However, due to coordinative specifics of these actions, low level business process descriptions does not reduce much complexity, i.e. it does not reduce neither the number of participants of a specific interaction nor helps in defining a single sequential course of actions. Furthermore, it seems that business process management paradigm has been invented to solve a slightly different problems than these appearing in coordinative actions.

### D. Goals of Information Transfer

*Description:* Goals of information transfers need to be defined because they are not default and they can be related to several actions with different effects. The goals organize related information transfers and they can be divided in sub-goals.

*Comparison to BPM:* A given business event usually starts business process which is strictly defined and formats of documents are fixed. Business goals and objectives are defined at higher level of business policy and strategy or they result from business process decomposition. Information transfers may be more flexible and thus more creativity can be involved in action. The goal acts as a frame of reference.

*Example:* Let's assume that the `goal` is to inform students about possibility to take part in Erasmus+ program, about the rules and recruitment procedure. It can be implemented via several actions which can be stated as `sub-goals` such as:

- provide short information via posters,
- organize a meeting,
- provide detailed information on the website,
- send e-mails to the students who expressed their interest,
- send short message via existing information systems.

Each of the sub-goals can be decomposed to several actions and information transfers which need proper management.

### E. Management of Group Dynamics

*Description:* Information logistics defines target in terms of both individuals and groups. Formation of groups and sub-groups allows for increasing efficiency comparing to individual treatments. It is worth to mention that these groups and sub-groups might be very dynamic. Therefore, group management and customization of messages is necessary. The benefits are visible when rapid reaction on changes is possible and when right information goes to the right target group. This action can be classified as an infrastructure management action.

*Comparison to BPM:* Participants and their responsibilities in business process can be specified in terms of swimlanes. BPMN allows to specification of multi-instance pools on conversation diagram, multi-instance participant on choreography diagram and multi-instance task on collaboration diagrams. However, no support for modeling dynamic change of participant's role is provided, i.e. phenomena that the same physical participant becomes a participant on a different pool.

*Example:* Let's consider a group of students interested in Erasmus+ exchange. Some of them can belong to a sub-group of students interested in bilateral agreements related to mobility. Another sub-group may be interested in offers of training. New students can be added to these groups as well as some students might want to withdraw. Another group of students are candidates for exchange, who can be divided with recruitment decisions for these nominated and those who are not. The sub-group of nominated students can be further divided for a sub-group of these who fulfill all requirements of application procedure and get accepted on one hand and a sub-group of those who do not achieve it on the other hand. All these changes happen dynamically in process and they cannot be predicted from the start.

### *F. Flexibility of Processes Depending on Diversity and Changes in Areas Controlled by External Participants*

*Description:* The difficulty in defining fixed processes with strictly defined information transfers might be caused by dependence on external participants who act according to the diversity of processes and they can make changes without notification. Thus, goals can be achieved via flexibility of action together with active processes of information acquisition.

*Comparison to BPM:* Business processes are usually fixed with limited number of routes connected by appropriate gateways. It is assumed that external data are inserted by participants within the defined processes. The difference in information logistics approach is in the need to customize the process and the content of information transfers to the requirements and specifics of external participants.

*Example:* The processes of sending nominations for exchange students and rules of application procedures are good examples of the diversity. Most universities accept nominations of exchange students by e-mail, but there are some exceptions, e.g. nominations via their on-line system or application by students themselves first with confirmation by coordinators later on. The number of combinations of online systems and kinds of required documents in paper in a variety of formats and ways of delivery is really large. Furthermore, these actions are spread over time adhering to different schedules and they are performed for a large numbers of students. These circumstances cause that they are really difficult to manage without systematic information logistics methods.

### *G. Promotion via Different Channels*

*Description:* Information logistics may require promotion management in order to reach target group. It needs to be done via different channels and insights from marketing studies are useful in order to make it effectively.

*Comparison to BPM:* This issue is specific to marketing and it is not present in BPM.

*Example:* Typical promotion is made in order to inform students about Erasmus+ program. In section III.D several sub-goals related to different channels were described. It is worth to mention that is should be a consistent campaign with use of creative methods as well as traditional ones.

### *H. The Right Information at the Right Time and in the Right Place*

*Description:* This statement could be a motto for information logistics as it covers the most popular common themes. It is good to understand that these are main dimensions of each information transfer and they should be assured all together.

*Comparison to BPM*: BPMN contains element called `data object` which represents information flowing through the business process. Timing and participants are described on collaboration diagram in context of the business process. The difference is in focus indicated by the word `right`, i.e. in information logistics one should analyze the needs for the right information at the right time and in the right place from the perspective of participant rather than optimization of request processing.

*Example:* In context of large amounts of rules and regulations the idea to provide the right information at the right time really promises to save participants time, increase confidence of action, and thus increase their satisfaction. Apart from general orientation, exchange students have different needs for detailed information on several stages of preparation to exchange. Before recruitment they need information about the rules and deadlines of recruitment. Then, they need details of other stages: application, actions after acceptance, actions during their stay and actions at departure. It is easier to prepare just one presentation for all, but the perspective of information logistics suggests that there should be separate information packages adhering to the paradigm of the right information at the right time to the right participants.

### *I. Discussion*

This action research was conceptually inspirational. It allowed to see the reality of Erasmus+ program coordination in the new perspective of information logistics. The connotations to logistics, transfers of right packages of information to right target just in time and the notion of infrastructure activate imagination to work on innovations in the coordination processes. Analysis according to the common themes in information logistics has driven attention to the aspects which were not analyzed in detail before and allowed to see them more clearly. The dimensions of the right information transfers, details of group dynamics and more clear understanding of changes controlled by external participants with related need for flexibility of processes should be taken into consideration when making changes in the process together with traditional techniques of business analysis and project management. To conclude, this action research (from the perspective of process improvement) has provided inspiration, useful terminology and solid bases for making decisions about improvements.

From the perspective of research on information logistics, it has appeared that the common themes are at very different level of abstraction. Some of them are very concrete and they can be transformed to an attribute of information transfer, e.g. content of information transfer. On the contrary others are extremely general without indication on operational action, e.g. relationship to other fields such as information management, knowledge management, communication management and business networks. These common themes which are concrete have led to more detailed results in analysis. Those more general require an extra effort to make them work. Regarding the enhancement to common themes, the examples of issues are hierarchy of goals of information transfers, management of group dynamics, and flexibility of processes depending on diversity. The question arise whether they are typical to several information logistics problems or they are specific just to the domain of this action research.

It seems that the common theme of `information flow or information value chain` is a good proposition for a common ground between information logistics and business process management. Actions related to sequences of information flows can be described by business processes. Information transfer details, such as channel, demands, content, time, size and quality constraints, can be provided by information logistics perspective. This action research has confirmed that business process management perspective is not sufficient to cover all aspects specific to information logistics. The issue of `the right information at the right time and in the right place` (III.H) shows that these approaches can be complementary. The goals of information transfers (III.D) sometimes correspond to business process but sometimes to a more general concepts in business strategy. Flexibility of processes and their changes (III.F) might be represented in their description, however this activity is profitable only when the cost of process update is lower than the cost of acquiring updated information when it is needed. The changes in assignments of individuals to group and dynamic change of roles (III.E) seem to be in opposition of business process fundamentals of having clearly defined participants of interaction. Promotion via different channels (III.G) is an example of issue which is outside of area of business process management. Therefore, this action research has revealed several shades of relationship between information logistics and business process management.

## IV. Towards Information Logistics Paradigm

The following attempt to define information logistics paradigm is based on both analytical results and action research insights with the categories of paradigms identified on the bases of paradigms in philosophy of science adjusted to technological context. Fig. 3 shows the contribution of the parts of presented research for the components of information logistics paradigm description.

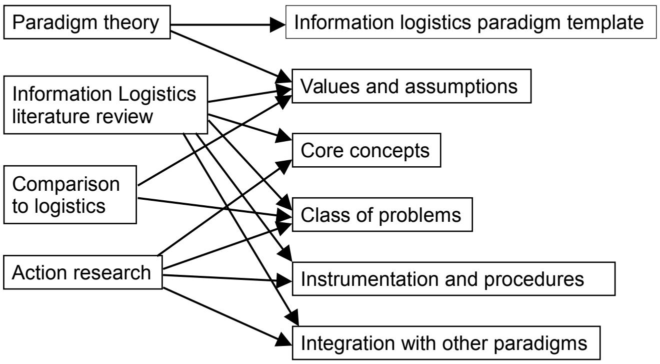

Paradigm theory
Information Logistics literature review
Comparison to logistics
Action research
Information logistics paradigm template
Values and assumptions
Core concepts
Class of problems
Instrumentation and procedures
Integration with other paradigms

Fig. 3 Relationship between the parts of research and the components of paradigm template

### A. Values and Assumptions

Values come from two kinds of sources. The first one is paradigm-based approach with its tendency to define typical solutions in terms of paradigms and allow for integration of different paradigms for real-life solutions. The second source is the area of information logistics which has identified several concepts and issues related to information logistics. In the diversity of the field, the value of strong connotations to logistics acts a point of reference when making decisions.

### B. Core Concepts

Mature paradigms have their scientific language and key theories. The first step to define them for Information Logistics is a common language expressed as core concepts presented in Fig. 4. It details concepts related to information transfer and envisages the need for further elaboration with two packages of Activities and Infrastructure.

### C. Class of Problems

Information logistics can effectively solve problems which appear as information chaos in organizations or during collaboration with other organizations. They usually happen in context of large amounts of information, large numbers of stakeholders with different information needs, changes of information outside of control of those who manage it, needs to access to crucial business information at the right time, needs for flexible and informed customizations to external parties as well as dynamic emergence of new information suppliers, requestors or channels.

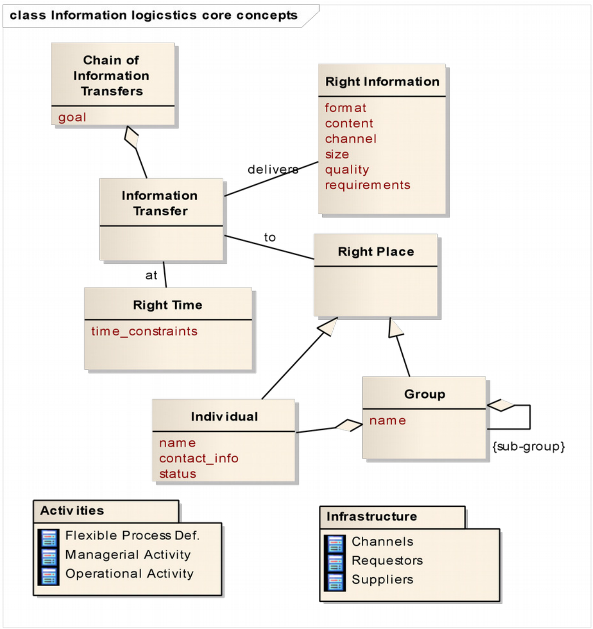


Fig. 4 Core concepts of Information Logistics

### D. Instrumentation and Procedures

Methods and technologies constitute technical equivalents for instrumentation and procedures in scientific context. They include ICT applications as well as techniques for visualization, case management, information transfer optimization, etc. In this approach a few kinds of methods can be distinguished. These can be methods specific to information logistics and methods for integration with other fields of study. Another classification can distinguish between methods for defining, elaborating and improving solution and methods for working with solution.

### E. Integration with Other Paradigms

The main connection point needs to be defined to business process management. The chains of information transfers are complementary to operational activities related with information acquisition and processing in context of circumstances which can be defined as stages of processing. The second connection point should direct to effective ICT systems which support users in performing their tasks. Other connection points can be made to results from related fields of study, such as marketing studies, information management, knowledge management, communication management and business networks. They address connection points related to solution. Other connection points can be made to techniques for elaborating solution, e.g. in the areas of business analysis and project management.

## V. Conclusions and Further Work

The paradigm-based approach is based on assumption that different classes of problems have effective partial solutions in technologies which use different paradigms and there is a need to integrate technologies which are based on different paradigms when developing real-life solutions. Information logistics can be viewed as a paradigm. This paper has addressed the first task on the way towards information logistics paradigm, which was the extraction of concepts specific to information logistics and the attempt of their description in categories specific to paradigms.

A frame of reference for categories specific to paradigms was elaborated on the basis of paradigms in philosophy of science and contemporary uses of the term of paradigm in computer science. When taking into account a large number of diversified definitions of information logistics in literature, the decision was made to build information logistics paradigm with the value of strong connotations to logistics, i.e. information logistics is a kind of logistics which is related to information instead of material goods and services.

Core concepts were extracted with analysis of common themes of information logistics definitions and notions provided by both researchers and practitioners. The essentials of information logistics appeared to be transfer of the right information at the right time to the right place. Common themes allowed to identify several issues, such as chains of information flow, cost optimization, distinction between two cases: information distribution and information about logistics of goods and services, types of operational activities with focus on dissemination rather than production and classification; managerial activities and infrastructure management, and relationships to other fields of study such as information management, knowledge management, communication management and business networks.

In order to keep compliance to the value of strong connotation to logistics, the cross-analysis was made in which the common themes were compared with themes of logistics management definition provided by the Council of Supply Chain Management Professionals. It has shown that in almost all categories some equivalents can be found although they have different motivation and focus. A brief comparison to Information Management Body of Knowledge was also made because information management and information logistics are related areas. The comparison has shown that, despite of similarities, diversity in approaches still exists.

The empirical part was made as action research in the area of coordination of Erasmus+ program. With the experience of facing real-life problems, the following issues have been discussed: need to set up a hierarchy of goals of information transfers, management of group dynamics, flexibility of processes depending on diversity and changes in areas controlled by external participants, promotion via different channels, and problems with the right information at the right time and in the right place. Each issue has been illustrated with the examples and comparison to business process management perspective has been discussed. A summary of issues on the edge of information logistics and business process management shows diversity of cases starting from clear complementary relationship, via issues which require consideration according to criteria from yet another perspective, to the situation where business process cannot help because they do not deal with these issues.

Both analytical research results and action research have contributed to an attempt to formulate information logistics paradigm. Within further work it is planned to supplement the model with more concepts specific to information logistics, formulate a proposal of UML profile for Information Logistics and provide connections to related technologies which are based on other paradigms. The concepts specific to information logistics can be gathered with more detailed review of related work in information logistics area. A proposal of UML profile for information logistics can be done with application of meta-modeling technology delivered by OMG. The most challenging task is providing connection to technologies which are based on other paradigms since there is no certainty whether it can lead to general results or these relationships are specific to a case at hand. Additionally, the research could benefit from more empirical research results conducted also in different domains.

## References

[1] D. Haftor, M. Kajtazi, A. Mirijamdotter, "A Review of Information Logistics Research Publications" *in Lecture Notes in Business Information Processing* LNBIP 97, 2011, DOI: 10.1007/978-3-642-25370-6_24.

[2] D. Haftor, M. Kajtazi, "What is information logistics? An explorative study of the Research Frontiers of Information Logistics", 2009, lnu.-diva-portal.org (access 7.05.2015)

[3] M. Apelkrans, and A. Håkansson, "Enterprise Systems Configuration as an Information Logistics Process - A Study" *in Proceedings of the 9th International Conference on Enterprise Information, Systems*, Portugal, 2007, DOI: 10.5220/0002370602120220

[4] S.Haseloff, "Context Awareness in Information Logistics", PhD dissertation, Technischen Universität Berlin, Germany, 2005.

[5] F. Markus, *Information Logistics in Supply Chain Networks. Concept, Empirical Analysis, and Design*. Ibidem Verlag, Hannover, Germany, 2004

[6] S. Grolik, *Information Logistics – Decentralization Approaches of Information Allocation in Information Exchange Networks*. Ibidem-Verlag, Germany, 2007

[7] B. Dinter, R. Winter, "Information Logistics Strategy Analysis of Current Practices and Proposal of a Framework" i*n Proceedings of the 42nd Hawaii International Conference on System Science*s, IEEE, Hawaii, 2009, DOI: 10.1109/HICSS.2009.253

[8] T. Bucher, and B. Dinter, "Process Orientation of Information Logistics – An Empirical Analysis to Assess Benefits, Design Factors, and Realization Approaches" *in Proceeding of the 41st Hawaii International Conference on System Sciences*, Hawaii, 2008, DOI: 10.1109/HICSS.2008.361

[9] W. Olthof, J. de Haan, J.Willems, Information Logistics, White Paper, NRG/Nashuatec Benelux, 2008

[10] Object Management Group (OMG), Unified Modeling Language v. 2.4.1, http://www.uml.org.

[11] Object Management Group (OMG), Business Process Model and Notation v.2.0, http://www.omg.org.

[12] L. Fuentes-Fernandez, A. Vallecillo-Moreno, "An introduction to UML profiles" *in UPGRADE* Vol. V, No. 2, April 2004.

[13] T.S. Kuhn, "The Structure of Scientific Revolutions", University of Chicago Press. Chicago, 1970.

[14] R. Floyd, "The paradigms of programming", *in Communications of the ACM*, 22 (8), 1979, DOI:10.1145/359138.359140

[15] S.H. Kaisler, "Software paradigms" Wiley & Sons. Inc. New Jersey, 2005.

[16] Giese H., Levendovszky T., Vanghekuve H., "Summary of the Workshop on Multi-Paradigm Modeling: Concepts and Tools" *in T. Kuhne (Ed.): MoDELS 2006 Workshops*, LNCS 4364, Springer-Verlag Berlin Heidelberg, 2007.

[17] C. Hardebolle, F. Boulanger, "Exploring Multi-Paradigm Modeling Techniques" *in SIMULATION*, Vol. 85, Issue 11/12, Nov./Dec. 2009

[18] Council of Supply Chain Management Professionals (CSCMP), Supply Chain Managament Terms and Glossary Updated 2013, http://cscmp.org (access: 10.04.2015)

[19] A. Bytheway, Investing in Information The Information Management Body of Knowledge, Springer International Publishing, Switzerland, 2014, DOI: 10.1007/978-3-319-11909-0_2

[20] European Commission, Erasmus+ Programme Guide, http://ec.europa.eu/, 2014 (access 18.06.2015)

DOI: 10.15439/2015F275
ACSIS, Vol. 5

# Supply Chain Coordination between Autonomous Agents – A Game Theory Approach

Gábor Kovács
Budapest University of Technology and
Economics, Műegyetem rkp 3, 1111
Budapest, Hungary
Email:
gabor.kovacs@logisztika.bme.hu

Katarzyna Grzybowska
Poznan University of Technology,
Strzelecka 11, 60-965 Poznan, Poland
Email:
katarzyna.grzybowska@put.poznan.pl

*Abstract*— **A supply chain is a network of suppliers, factories, warehouses, distribution centers and retailers, through which raw materials are acquired, transformed, produced and delivered to the customer. A supply chain management system (SCMS) manages the cooperation of these system components. In the computational world, roles of individual entities in a supply chain can be implemented as distinct agents [1]. In this paper we present supply chain coordination between Autonomous Agents. Moreover, we present a cooperative game theory approach to describe the SCM coordination. Numerical and theoretical game examples are detailed in this paper, which help to understand the usefulness of cooperative game theory in SCM.**

## I. Introduction

Coordination between agencies during multi-agency emergency responses, although a key issue, remains a neglected research area [2]. Coordination between the different agencies (enterprises) involved is a major challenge. Most of the components in Supply Chain Management (SCM) work in isolation and achieving coordination among Supply Chain Management partners turns out to be a difficult proposition. A supply chain typically extends across the multiple enterprises including suppliers, manufactures, transportation carriers, warehouses, retailers as well as customers and entails sharing forecast, order, inventory, and production information to better coordinate management decisions at multiple points throughout the extended enterprise [3].

The game theory approach is one of the best tools for modelling this complex system; and for modelling the cooperation between intelligent decision makers, the co-management and the autonomous agents. We know the players of the game, the information and actions are available to each player at each decision point, and the payoffs can be calculated for each outcome. In this paper, the cooperative game theory approach is detailed, through its main features.

The paper is organized as follows: Section (II) discusses the Supply Chain Coordination and co-management. Section (III) discusses the Multi-level Governance. Section (IV) introduces the construction of Supply Chain with the help of an agent. Section (V) presents the different types of games used in supply chain management. In section (VI) discusses the main features of cooperative games. And finally, section (VII) and section (VIII) detail numerical and theoretical SCM examples, in section (IX) can be read the conclusions.

The main goal of the paper is merging the up-to-date knowledge of supply chain coordination and the game theory. So, it contains a large number of reviews, but as an outlook, as an own contribution, it gives a usage example too. This paper integrates the SCM logistic and mathematical modelling knowledge.

## II. Supply chain coordination and Co-management

Supply chains (SC) are a system with "multiple actors". The supply chain is commonly seen as a collection of various types of companies (raw materials, production, trade, logistics, transport, etc.) working together to improve the flow of products, information and finance [4], [5]. Supply chains are complex systems, dynamic, dispersed and open. Those elements together with other factors (e.g. multiple subjects, independence of cooperating enterprises) determine difficulties in the field of management, or more broadly, of coordination of commonly take up and independently realized actions. The discussed systems are affected, as a whole, by a lack of internal rationality, unverified information and insufficient knowledge. The problem is also posed by uncertainty and a lack of precision [6]; [7], indispensable in the realized projects and complex undertakings.

Co-management is of growing interest among researchers. Centralized, top-down resource management is ill-suited to user participation. Centralized management are limited in their ability to respond to changing conditions, an anachronism in a world increasingly characterized by rapid transformations [8]; [9]. Changing ideas about the nature of resource management, ecosystems, and social-ecological

□ This work is supported by Poznan University of Technology and Budapest University of Technology and Economics

978-83-60810-66-8/$25.00©2015, IEEE

systems (integrated systems of people and environment) have been catalyzed by insights from complex adaptive systems thinking.

Selected features of adaptive co-management:

- Shared vision, goal, and/or problem definition to provide a common focus among actors and interests;
- A high degree of dialogue, interaction, and collaboration among multi-scaled actors;
- Distributed or joint control across multiple levels, with shared responsibility for action and decision making;
- A degree of autonomy for different actors at multiple levels;
- Commitment to the pluralistic generation and sharing of knowledge;
- A flexible and negotiated learning orientation with an inherent recognition of uncertainty [9].

Plummer and Fennell [10] build upon initial efforts to capture how adaptive co-management is being understood [11, 12, 13] to arrive at the following attributes.

- Pluralism and communication. Actors from diverse spheres of society (and at multiple levels) and who have varying principal interests enter into a process to generate shared understanding of an issue or problem. This process is grounded in communication and negotiation. Conflict is viewed as an opportunity.
- Shared decision-making and authority. Transactive decision-making is employed as a basis for achieving decisions. Multiple sources of knowledge are acknowledged. Authority (power) is shared in some configuration among the actors involved.
- Linkages, levels and autonomy. Actors are connected or linked both within levels and across scales. Despite shared interests and commitments, actor autonomy is appropriate at multiple levels. Institutional arrangements therefore encompass multiple levels as well as retain flexibility.
- Learning and adaptation. Actions and policies are considered experiments. Feedback provides opportunities for social learning in which outcomes are collectively reflected upon and modifications to future initiatives are based. Learning may concern routines, values and policies, and/or critical questions of the underlying governance systems; referred to as multiple-loop learning. Develops as trust and knowledge [14].

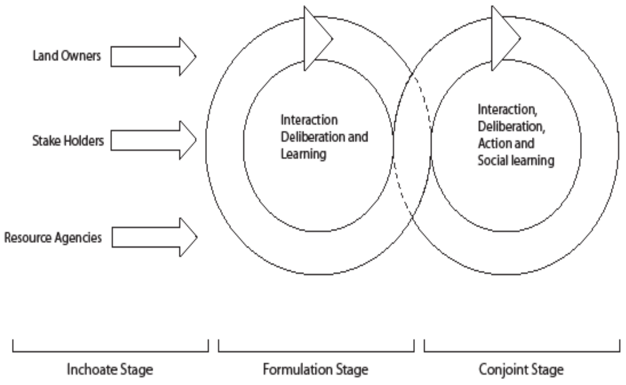

Land Owners
Stake Holders
Resource Agencies
Interaction Deliberation and Learning
Interaction, Deliberation, Action and Social learning
Inchoate Stage
Formulation Stage
Conjoint Stage

Fig. 1 A model of the adaptive co-management process [14]

Co-management is not a fixed unitary entity, rather it is a set of principles for institutional design that can assume various organizational forms depending on particular circumstances [15].

Coordination defined as the process of managing dependencies among activities. Starting with the individual activity it is easily recognized that the industrial reality contains a multitude of various activities. When focusing solely on individual activities, these might seem to have a generic value, for example considering a production or exchange activity [16].

## III. Multi-Level Governance

The chief benefit of multi-level governance (MLG) lies in its scale flexibility. Its chief cost lies in the transaction costs of coordinating multiple competence. The coordination dilemma confronting multi-level governance can be simply stated: To the extent that policies of one competence have spillovers (i.e. negative or positive externalities) for other jurisdictions, so coordination is necessary to avoid socially perverse outcomes. We conceive this as a second -order coordination problem because it involves coordination among institutions whose primary function is to coordinate activity [17]. Type (I.) multi-level governance describes jurisdictions at a limited number of levels. That is to say, they bundle together multiple functions. Type (II.) multi-level governance is distinctly different. It is composed of specialized competence. The number of such competence is potentially huge. They tend to be lean and flexible – they come and go as demands for change [17]. Multi-level governance is the domain of the European Union.

Multi-level governance characterizes the changing relationships between actors situated at different levels. MLG contributes to a growing awareness that many contemporary issues and challenges require analysis that transcends traditional disciplinary boundaries.

Multi-level governance:

- Decision-making competencies are shared by actors at different levels rather than monopolized by executives;
- Collective decision-making significant loss of control for individual executives.

## IV. Construction of Supply chain with the help of an agent

It should be assumed that this is one of the simplest coordination mechanisms. It assumes that the enterprises in the built structure possess a hierarchy, previously provided. In order for it to function effectively, the execution of the following tasks is necessary:

- Initiating the creation of a database of enterprises that will operate within the structure;
- Defining the scope of activities of the individual entities;
- Specifying the rights and obligations of the individual entities (regulations);
- Expanding the database of enterprises through own actions (sending information through the available communication channels, i.e. e-mail, press, internet...);
- Registration of structure participants;
- Approving the participants;
- Agreement;
- Establishing priorities and dependencies between the enterprises.

The agent should be understood and treated as coordinating activities of the organization. The agent should be:

- reactive – agent-coordinator identifies and responds to the tasks; It has current knowledge about the business,
- pro-active – agent-coordinator takes the initiative in order to carry out tasks,
- able to cooperate – agent-coordinator interacts with others in order to carry out the task.

The benefit of relations between enterprises defined in such a manner is the legible and explicit indication of the role that each enterprise is to play in the created structure. The building of structures with the help of an assistant most often assume the hierarchical master/slave structure. In such a case the agent master plans and sends out information on the orders to the individual subordinate agents (slave). And each of these agents transfers return information on the status of the completion of their order. The defect of such an approach is the small amount of autonomy for the slave agents. Coordination through the organization works ideally in the coordination of the tasks of agents connected by strong hierarchical relations [18].

Bennett and McCoshan (1993) [19] have suggested a typology of networks (Figure 1) which describes a range of relations between agents at the local or regional level. As they note, the networks A-D are derived from management science, and each has advantages and disadvantages in delivering economic development activities efficiently and sustainably. The introduction of the fifth form, E, is meant to be flexible and responsive to the needs of different agents [20].

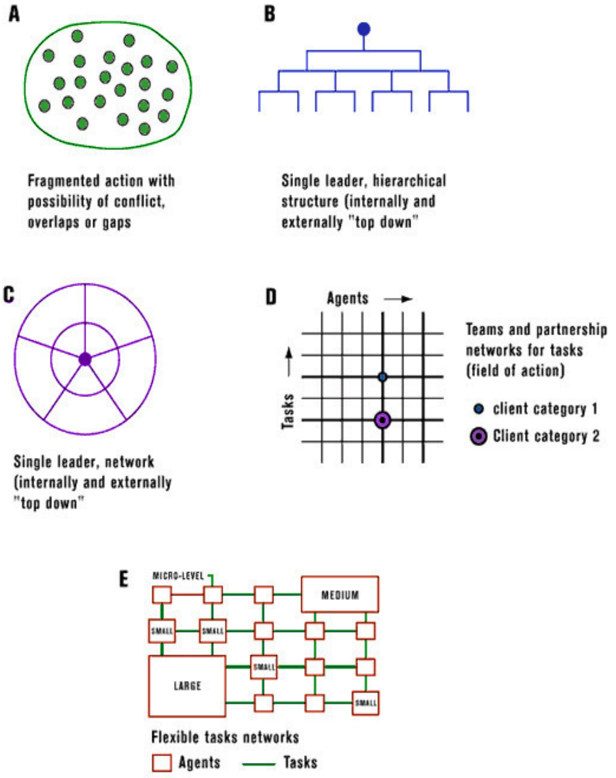

A
Fragmented action with possibility of conflict, overlaps or gaps
B
Single leader, hierarchical structure (internally and externally "top down"
C
Single leader, network (internally and externally "top down"
D
Agents
Tasks
Teams and partnership networks for tasks (field of action)
client category 1
Client category 2
E
MICRO-LEVEL
MEDIUM
SMALL
SMALL
SMALL
LARGE
SMALL
Flexible tasks networks
Agents
Tasks

Fig. 1 Network of relations between agents at a local level [20]

## V. The different types of games used in Supply Chain Management

Game theory is a powerful tool for analyzing situations in which the decisions of multiple agents affect each autonomous agent's payoff. The elements and rules mentioned in the previous section of this paper are the SCM conceptual basis: the decision-making, the coordination, the governance and the agents are the main subcomponents. The following game theory approach gives the opportunity of the mathematical modelling.

As such, game theory deals with interactive optimization problems. While many economists in the past few centuries have worked on what can be considered game-theoretic models, John von Neumann and Oskar Morgenstern (1944) [21] are formally credited as the fathers of modern game theory. Their classic book summarizes the basic concepts existing at that time. Game theory has since enjoyed an explosion of developments, including the concept of equilibrium by Nash (1950) [22], games with imperfect information by Kuhn (1953) [23], cooperative games by Aumann (1959) [24] and Shubik (1962) [25].

There are many game theory concepts, but this paper focuses on concepts that are particularly relevant to supply chain management (SCM) and, perhaps, already found their applications in the literature. The main state of the art: Myerson (1997) [29], Friedman (1986) [26], Fudenberg and Tirole (1991) [27], Topkis (1998) [30] and Vives (1999) [31], Moulin (1986) [28]. Some previous surveys of game theory models in management science include Lucas's (1971) survey of mathematical theory of games [32], Feichtinger and Jorgensen's (1983) [33] survey of differential games and Wang and Parlar's (1989) survey of static models [34], Porteus and Whang (1999) [38] survey of screening game. In addition, Fudenberg and Tirole (1991) [27] for more information on Bayesian games, Cachon and Lariviere (2001) [35] survey of signaling game, Brandenburger and Stuart (1996) [39] for more information of business process games, and [40], [41], [43], [44] about the core of the game, [44] about the Shapley value.

## VI. The main features of cooperative games

Cooperative game theory focuses on the outcome of the game, where the outcome is measured in terms of the value created through cooperation of a subset of players [35]. In what follows, we will cover transferable utility cooperative games (players can share utility via side payments) and three solution concepts:

- the core of the game;
- the Shapley value;
- and the nucleolus.

### A. Games in characteristic form and the core of the game

The cooperative game consists of the set of players $N$ with subsets or coalitions $S \subseteq N$ and a characteristic function $v(S)$ that specifies a (maximum) value (which we assume is a real number) created by any subset of players in $N$, i.e., the total pie that members of a coalition can create and divide. A frequently used solution concept in cooperative games is the core of the game. The utility vector $x_1, \ldots, x_N$ is in the core of the cooperative game if

$$\forall S \subseteq N,\ \sum_{i \in S} x_i \geq v(S) \text{ and } \sum_{i \in N} x_i = v(N)$$

A utility vector is in the core if the total utility of every possible coalition is at least as large as the coalition's value, i.e., there does not exist a coalition of players that could make all of its members at least as well off and one member strictly better off.

### B. Shapley value, nucleolus

The concept of the core, though intuitively appealing, also possesses some unsatisfying properties. Shapley (1953) offered an axiomatic approach to a solution concept that is based on axioms [45]. One of the most important is that: if $v_1$ and $v_2$ are characteristic functions in any two games, and if $\phi_1$ and $\phi_2$ are a player's Shapely value in these two games, then the player's Shapely value in the composite game, $v_1 + v_2$, must be $\phi_1 + \phi_2$.

An alternative equivalent formula for the Shapley value is:

$$\varphi_i(v) = \frac{1}{|N|!} \sum_{R} \left[ v\left(P_i^R \bigcup \{i\}\right) - v(P_i^R) \right]$$

where the sum ranges over all *INI!* orders $R$ of the players and $P_i^R$ is the set of players in $N$ which precede $i$ in the order $R$.

Another interesting value function for cooperative games may be found in the nucleolus, a concept introduced by Schmeidler (1969) [47]. The main idea: we look at a fixed characteristic function, $v$, and try to find an imputation $x = (x_1, \ldots, x_n)$ that minimizes the worst inequity. As a measure of the inequity of an imputation $x$ for a coalition $S$ is defined as the excess:

$$e(x, S) = v(S) - \sum_{i \in S} x_i$$

which measures the amount (the size of the inequity) by which coalition $S$ falls short of its potential $v(S)$ in the allocation $x$.

## VII. Numerical example

There are three enterprises (A, B, C; logistics providers – e.g. freight, storage, complex logistics processes, transhipment processes, with using roads and rails too -), the core of the game based on the following constraints (Fig. 2):

$$v(A)=v(B)=v(C)=0$$

$$v(AB)=3$$

$$v(AC)=5$$

$$v(BC)=4$$

$$v(ABC)=7$$

Here, individually, none of the players can receive any payoff. But if they cooperate, different coalitions result in a positive payoff for each coalition. If they all cooperate, then the grand coalition receives an amount $v(ABC)$ higher than any other coalition. Other words: they have to perform a multimodal logistics task.

The core of a game in characteristics form is defined as the set of all imputations $(x_1, x_2, \ldots, x_n)$ such that for all

$S \subseteq N$, $\sum_{i \in S} x_i \geq v(S)$. The core is the set of all $(x_A, x_B, x_C)$ satisfying:

$$x_A+x_B\geq v(AB)=3$$

$$x_A+x_C\geq v(AC)=5$$

$$x_B+x_C\geq v(BC)=4$$

$$x_A+x_B+x_C= v(ABC)=7$$

The set of imputations in this game can be represented by an equilateral triangle with high equal to $v(ABC)=7$. For any point $(x_A, x_B, x_C)$ in the triangle, $x_i$ is the distance to side of the opposite corner, $i=A, B, C$; as indicated in Figure 2.

Thus, player $i$ prefers imputations that are close to corner $i$. Since

$$x_i+x_j\geq v(ij) \leftrightarrow x_k\leq v(ABC)- v(ij)$$
$$for\ i\neq j\neq k$$

the latter inequalities can be drawn to obtain the core – provided that is nonempty.

The core in this game is obtain by drawing the regions

$$x_A\leq 3$$

$$x_B\leq 2$$

$$x_C\leq 4$$

These give rise to the area indicated by interrupted lines in Figure 2.

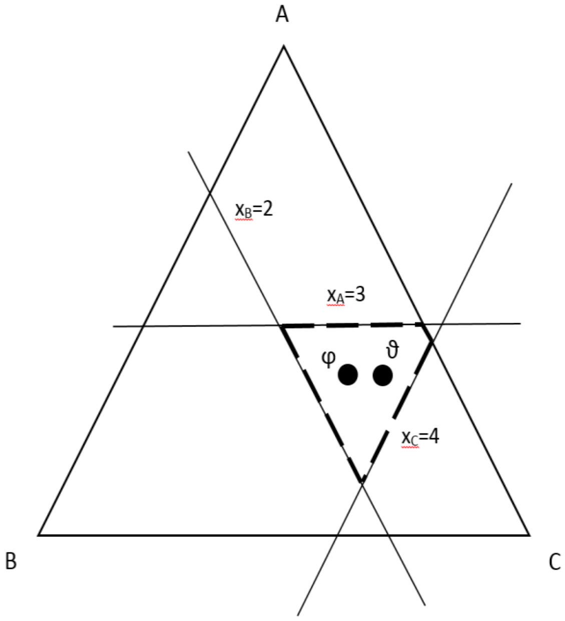


Fig. 2 The core of the game, the Shapley value and the nucleolus

Table 1. shows the marginal contributions of players, based on this, we can calculate the Shapley value.

TABLE I.
THE MARGINAL CONTRIBUTIONS OF PLAYERS

| Orders of the players | Marginal contributions of A | Marginal contributions of B | Marginal contributions of C |
|---|---|---|---|
| ABC | v(A)-v(0) | v(B)-v(0) | v(C)-v(0) |
| ACB | v(A)-v(0) | v(B)-v(0) | v(C)-v(0) |
| BAC | v(AB)-v(B) | v(AB)-v(A) | v(AC)-v(A) |
| BCA | v(ABC)-v(BC) | v(ABC)-v(AC) | v(ABC)-v(AB) |
| CAB | v(AC)-v(C) | v(BC)-v(C) | v(BC)-v(B) |
| CBA | v(ABC)-v(BC) | v(ABC)-v(AC) | v(ABC)-v(AB) |

The Shapley value for the three players are found as

$$\varphi_A(v) = \frac{14}{6} = 2{,}33$$

$$\varphi_B(v) = \frac{11}{6} = 1{,}83$$

$$\varphi_C(v) = \frac{17}{6} = 2{,}83$$

Based on Leng and Parlar (2010), we can use explicit formula to compute the nucleolus [46]:

$$xi = \frac{v(123) + v(ij) + v(ik) - 2v(jk)}{3}$$

$$for\ i,j,k = 1,2,3\ and\ i \neq j \neq k$$

The nucleolus $\vartheta(x_A, x_B, x_C)$ for the three players are found as (Table 2. shows $e(x,S)$):

$$x_A = \frac{7}{3} = \frac{14}{6} = 2{,}33$$

$$x_B = \frac{4}{3} = \frac{8}{6} = 1{,}33$$

$$x_C = \frac{10}{3} = \frac{20}{6} = 3{,}33$$

TABLE II.
THE E(X,S) IN THE NUCLEOLUS

| S | v(S) | e(x,S) | (14/6; 8/6; 20/6) |
|---|---|---|---|
| A | 0 | 0-$x_A$ | -2,33 |
| B | 0 | 0-$x_B$ | -1,33 |
| C | 0 | 0-$x_C$ | -3,33 |
| AB | 3 | 3- $x_A$- $x_B$ | -0,67 |
| AC | 5 | 5- $x_A$- $x_C$ | -0,67 |
| BC | 4 | 4- $x_B$- $x_C$ | -0,67 |

In this example, the Shapley value and the nucleolus is also in the core. They give solution alternatives of game, which are relatively close to each other.

Based on this numerical example, we can calculate the tangible benefits of a virtual logistics alliance. Moreover, there are three indicators (core of the game, Shapley value, nucleolus), to evaluate the benefit of this alliance, and the personal effects too. The great advantage of this solution is the quantifiability and the opportunity of the multi criteria decision making.

## VIII.THEORETICAL SCM EXAMPLE

The previous numerical example could be good for modelling cooperation in the freight and warehouse exchanges Kovács (2009) [49], Grzybowska and Kovács (2012) [50], Grzybowska and Kovács (2014) [51]. The simplified system model of the supply chain supported by electronic freight and warehouse exchanges is shown in Figure 3.

In this system, the electronic freight and warehouse exchanges perform the supply-demand (freight/storage capacities/tasks) harmonization; the decision supporting, the optimization and the whole software/hardware support. The logistics providers (storage providers, transportation providers, logistics centres) perform the physical freight/storage/transhipment tasks; whereas they have: suitable stock capacities, suitable freight capacities, equipment's, and logistics know-how. The wholesalers are responsible for the information processes; they manage the demands of retailers. This supply chain may be optimal, through using cooperative game theory, pollution or cost point of view. Consequently, green logistics systems, e.g. green city supply chains or combined transportation systems can be realized. In addition, this system is beneficial not only for the individual actors (e.g. retailers, wholesalers, logistics providers, manufacturers) but also for the national economy (reduce traffic flow, pollution, noise). The future plans include further development of algorithms and tests in real supply chains.

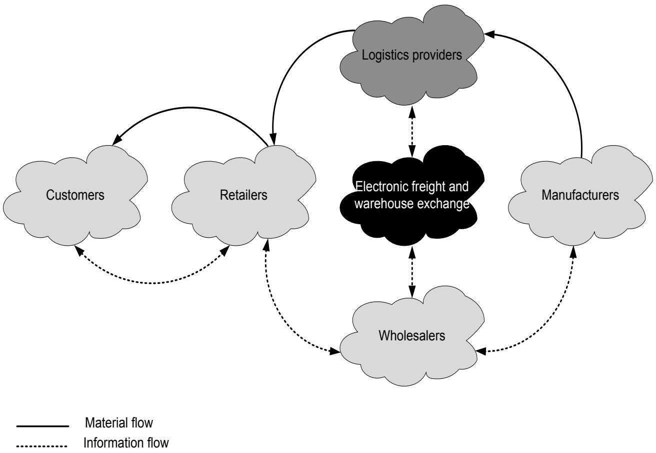


Fig. 3 The simplified system model of the supply chain supported by electronic freight and warehouse exchanges

As another example of potential SCM modelling, research at the Department of Material Handling and Logistics Systems in Budapest is aimed to help logistics processes at the construction industry. This work has been developed in the framework of the project "Development of construction processes from logistical and informatical aspects". This research is part of a project (KTIA-AIK-12-1-2013-0009) financed by the National Development Agency of Hungary. This project concentrates on the logistics aspects, where organization of the material flow is an important task. Based on this research, we can create flowcharts (for top-down modelling and for low-level modelling too), which help to analyse the real construction processes, and thereby we can build up realistic game models too.

## IX. CONCLUSIONS

The main result of this article is merging the supply chain coordination and the cooperative game theory approach. By the explanations and the numerical example, this logic modelling is reasonable. The occurring decision supporting problem can be modelled well, the branching points and a variety of outputs can be understood and managed. The main contribution is the combination of SCM and mathematical principal founds, by the addition of numerical and theoretical examples too.

One of most interesting application is the virtual alliances in the supply chain, such as freight exchanges, but other areas also may be promising. The next step in the research will be to make essential progress in the field of supply chains, e.g. a freight and warehouse exchange game model structure.

## REFERENCES

[1] Y. Chen, Y. Peng, T. Finin, Y. Labrou, and S. Cost, "Negotiating Agents for Supply Chain Management", *AAAI Workshop on Artificial Intelligence for Electronic Commerce*, Orlando, Florida, July, 1999.

[2] R. Chen, R. Sharman, H.R. Rao and S.J. Upadhyaya, "Coordination in emergency response management", *Communications of the ACM*, Vol. 51 No. 5, pp. 66-73, 2008.

[3] H.A. Rady, "Multi-Agent System for Negotiation in a Collaborative Supply Chain Management", *International Journal of Video & Image Processing and Network Security IJVIPNS-IJENS*, Vol. 11 No. 05, pp. 25-35, 2011.

[4] P. Sitek P., "A hybrid CP/MP approach to supply chain modelling, optimization and analysis", *Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 1345-1352, 2014. http://dx.doi.org/10.15439/2014F89

[5] Sitek, P., Wikarek, J.: A Hybrid Approach to the Optimization of Multiechelon Systems. Mathematical Problems in Engineering vol. 2015, Article ID 925675, 12 pages, 2015. doi:10.1155/2015/925675

[6] M. Relich, "Knowledge Acquisition for New Product Development with the Use of an ERP Database", *Federated Conference on Computer Science and Information Systems*, pp. 1285–1290, Krakow, Poland 2013.

[7] M. Relich, "A computational intelligence approach to predicting new product success", Proceedings of the 11th International Conference on Strategic Management and its Support by Information Systems, pp. 142–150, 2015.

[8] Gunderson L. H., Holling C. S., "Eds. Panarchy: understanding transformations in human and natural systems", Washington, DC: Island Press. 2002.

[9] D. Armitage, F. Berkes, N. Doubleday, "Adaptive Co-Management: Collaboration, Learning, and Multi-Level Governance", Toronto, 2007.

[10] R. Plummer, D. Fennell, Managing Protected Areas for Sustainable Tourism: Prospects for Adaptive Co-management. J. Sustain. Tour. 17, pp. 149–168, 2009.

[11] R. Plummer, D.R. Armitage, Charting the New Territory of Adaptive Co-management: A Delphi Study. Ecol. Soc. 2007.

[12] D. Armitage, F. Berkes, N. Doubleday, Introduction: Moving Beyond Co-management. In Adaptive Co-management: Collaboration, Learning and Multi-level Governance; Armitage, D., Berkes, F., Doubleday, N., Eds.; UBC Press: Vancouver, Canada, pp.1–18, 2007.

[13] D.R. Armitage, R. Plummer, F. Berkes, R.I. Arthur, I.J. Davidson-Hunt, A. Diduck, N.C. Doubleday, D.S.Johnson, M. Marschke, P. McConney, et al. Adaptive Co-management for Social-ecological Complexity. Front. Ecol. Environ. 6, pp. 95–102, 2009.

[14] R. Plummer, J. Baird, Adaptive Co-Management for Climate Change Adaptation: Considerations for the Barents Region, Sustainability 5, 629-642; doi:10.3390/su5020629, 2013.

[15] S. Jentoft, B. J. McCay, D. C. Wilson, "Social theory and fisheries co-management", *Marine Policy*, Vol. 22, No. 4-5, pp. 423-436, 1998.

[16] L. Bankvall, "Activity coordination from a firm perspective -towards a framework", *Proceedings IMP-conference in Uppsala*, Sweden 2008.

[17] L. Hooghe, G. Marks, "Types of Multi-Level Governance", *European Integration online Papers (EIoP)*, Vol. 5, 2001.

[18] K. Grzybowska, "Selected Activity Coordination Mechanisms in Complex Systems", J. Bajo et al. (Eds.), *PAAMS 2015 Workshops*, CCIS 524, Springer International Publishing Switzerland, pp. 1–11, 2015. http://dx.doi.org/10.1007/978-3-319-19033-4_6

[19] R.J. Bennett, A. McCoshan,1993. Enterprise and human resource development. Local capacity building. London: Paul Chapman, 1993.

[20] M. Danson, G. Whittam, Regional Governance, Institutions and Development, in European Research in Regional Science, volume 7: Regional governance and economic development, Pion Limited, London, 1996.

[21] J. von Neumann, O. Morgenstern, "Theory of games and economic behaviour", Princeton University Press, 1944.

[22] J. F. Nash, "Equilibrium Points in N-Person Games", *Proceedings of the National Academy of Sciences of the United States of America*, 36, pp. 48-49, 1950.

[23] H. W. Kuhn, "Extensive Games and the Problem of Information", *Contributions to the Theory of Games*, 2, pp. 193-216, 1953.

[24] R. J. Aumann, "Acceptable Points in General Cooperative N-Person Games", *Contributions to the Theory of Games*, pp. 287-324, Princeton University Press, 1959.

[25] M. Shubik, "Incentives, decentralized control, the assignment of joint costs and internal pricing", *Management Science*, 8, pp. 325-343, 1962.

[26] J. W. Friedman, "Game theory with applications to economics", Oxford University Press, 1986.

[27] D. Fudenberg, and J. Tirole, "Game theory", MIT Press, 1991.

[28] H. Moulin, "Game theory for the social sciences", New York University Press, 1986.

[29] R. B. Myerson, "Game theory", Harvard University Press, 1997.

[30] D. M. Topkis, "Supermodularity and complementarity", Princeton University Press, 1998.

[31] X. Vives, "Oligopoly pricing: old ideas and new tools", MIT Press, 1999.

[32] W. F. Lucas, "An overview of the mathematical theory of games", *Management Science*, 18, pp. 3-19, 1971.

[33] G. Feichtinger, and S. Jorgensen, "Differential game models in management science", *European Journal of Operational Research*, 14, pp. 137-155, 1983.

[34] Q Wang, and M. Parlar, "Static game theory models and their applications in management science", *European Journal of Operational Research*, 42, pp. 1-21, 1989.

[35] G. Cachon, and M. Lariviere, "Contracting to assure supply: how to share demand forecasts in a supply chain", *Management Science*, 47, pp. 629-646, 2001.

[36] E. Porteus, and S. Whang, "Supply chain contracting: non-recurring engineering charge, minimum order quantity, and boilerplate contracts", Stanford University, 1999.

[37] G. Cachon, and M. Lariviere, "Capacity choice and allocation: strategic behaviour and supply chain performance", *Management Science*, 45, pp. 1091-1108, 1999. http://dx.doi.org/10.1287/mnsc.45.8.1091

[38] M. Nagarajan, and G. Sošic, "Game-theoretic analysis of cooperation among supply chain agents: Review and extension", *European Journal of Operational Research*, 187, pp. 719-745, 2008. doi:10.1016/j.ejor.2006.05.045

[39] A. Brandenburger, and H. W. Stuart, "Value-based business strategy", *Journal of Economics and Management Strategy*, 5 (1), pp. 5-24, 2005. DOI: 10.1111/j.1430-9134.1996.00005.x

[40] H. W. Stuart, "Cooperative games and business strategy", *Game theory and business applications*, Kluwer Academic Publishers, 2001

[41] H. Moulin, "Cooperative microeconomics: a game-theoretic introduction", Princeton University Press, 1995.

[42] Q. Wang, and M. Parlar, "A three-person game theory model arising in stochastic inventory control theory", *European Journal of Operational Research*, 76, pp. 83-97, 1994.

[43] B. C. Hartman, M. Dror, and M. Shaked, "Cores of inventory centralization games", *Games and Economic Behavior*, 31, pp. 26-49, 2000. doi:10.1006/game.1999.0732

[44] A. Muller, M. Scarsini, and M. Shaked, "The newsvendor game has a nonempty core", *Games and Economic Behavior*, Vol. 38, 118-126. 2002. doi:10.1006/game.2001.0854

[45] L. Shapley, "A value for n-person game", *Contributions to the Theory of Games*, 2, pp. 307-317, Princeton University Press, 1953.

[46] G. Granot, and G. Sosic, "A three-stage model for a decentralized distribution system of retailers", *Operations Research*, 51 (5), pp. 771-784, 2003. http://dx.doi.org/10.1287/opre.51.5.771.16749

[47] D. Schmeidler, "The Nucleolus of a Characteristic Function Game", *SIAM J. Appl. Math.*, 17(6), pp. 1163-1170, 1969

[48] M. Leng, and M. Parlar, „Analytic solution for the nucleolus of a three-player cooperative game", *Naval Research Logistics*, 57 (7), pp. 667-672, 2010. DOI: 10.1002/nav.20429

[49] G. Kovács, “The structure, modules, services, and operational process of modern electronic freight and warehouse exchanges”, *Periodica Polytechnica Transportation Engineering*, 37(1-2), pp. 33-38, 2009. 10.3311/pp.tr.2009-1-2.06

[50] K. Grzybowska, and G. Kovács, “Developing Agile Supply Chains - system model, algorithms, applications”, *Agent and Multi-Agent Systems. Technologies and Applications, Lecture Notes in Computer Science*, pp. 576-585, 2012. http://dx.doi.org/10.1007/978-3-642-30947-2_62

[51] K. Grzybowska, and G. Kovács, “Sustainable Supply Chain - Supporting Tools”, *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Annals of Computer Science and Information Systems*, pp. 1321-1329, 2014. http://dx.doi.org/10.15439/2014F75

DOI: 10.15439/2015F83
ACSIS, Vol. 5

# A Hybrid Multi-Objective Programming Framework for Modeling and Optimization of Supply Chain Problems

Paweł Sitek
Institute of Management Control Systems, Kielce University of Technology Al. 1000-lecia PP 7, 25-314
Kielce, Poland,
Email: sitek@tu.kielce.pl

*Abstract*—**This paper presents a hybrid programming framework for solving multi-objective optimization problems in supply chain. The proposed approach consists of the integration and hybridization of two modeling and solving environments, i.e., constraint logic programming and mathematical programming, to obtain a programming framework that offers significant advantages over the classical approach derived from operational research. The strongest points of both components are combined in the hybrid framework, which by introducing transformation allows a significant reduction in size of a problem and the optimal solution is found a lot faster. This is particularly important in the multi-objective optimization where problems have to be solved over and over again to find a set of Pareto-optimal solutions. An over two thousand-fold reduction in size was obtained for the illustrative examples together with a few hundred-fold reduction in the speed of finding the solution in relation to the mathematical programming method. In addition, the proposed framework allows the introduction of logical constraints that are difficult or impossible to model in operational research environments.**

## I. Introduction

SUPPLY chain (SC) is an integrated process in which a group of several organizations and/or companies, such as suppliers, producers, distributors and retailers, work together to acquire raw materials with a view to converting them into end products which they distribute to retailers [1]. Decision and optimization problems occurring in the real-world supply chain are characterized by multiple objectives, constraints and many different decision variables. The presence of multiple conflicting objectives and constraints is natural and results from the complexity and interrelated character of problems and different interests of individual supply chain participants. Environmental aspects such as CO2 emissions, noise, etc., which are a new type of constrains in SC, are emerging to become an important factor in the design of supply chains. Operational research models, with mathematical programming (MP) in particular, are most often used. They include MILP (mixed integer linear programming), MIP (mixed integer programming), IP (integer programming), etc. and MOOP (multi-objective optimization problem) [2]. The vast majority of these models have a very large number of constraints, and two major difficulties appear in their application. First, discrete optimization problems, both single and multi-objective, contain many discrete decision variables. This increases their computational complexity and the finding of the optimal solution is long and costly. Second, mathematical programming models have linear constraints, which is insufficient for the description of many of the SC problems. This paper deals with a problem of supply chain modeling, multi-objective optimization and solving. An important contribution of the presented approach is to propose a programming framework that supports the hybrid modeling, hybrid multi-objective optimization and analysis of decision problems in the supply chain. In this programming framework two environments are hybridized, constraint logic programming (CLP) and mathematical programming (MP), in which constraints are treated in different ways and different methods are implemented to use the strengths of both for solving complex and constrained problems. The hybrid approach offers a lot more possibilities and higher efficiency in both the modeling and multi-objective optimization. The concept of hybridization is complemented by the interaction algorithm and a complete transformation of the problem and together creating an application programming framework. The rest of the paper is organized as follows: Section 2 describes literature review. Next Section presents Methodology. Section 4 is about our motivation. Section 5 gives the concept of the novel constraint logic programming approach with MP-based solver and implementation platform. The optimization models as the illustrative examples are described in Section 6. Computational examples and tests of the implementation platform are presented in Section 7. The discussion on possible extensions of the proposed approach and conclusions is included in Section 8.

## II. Literature review of supply chains models and multi-Objective optimization

The build-to-order supply chain (BOSC) model, a key operation model for providing services/products, has received more attention in recent years, car manufacturers

This work was not supported by any organization

including [3]. In the BOSC model, production activities are not executed until orders from customers have been received, which can effectively reduce the costs of demand prediction and inventory and credibly reflect market demands. When BOSC is started, the selection of suppliers becomes the priority. Product assembly begins after this selection. Based on the findings reported in the literature [4],[5]. BOSC is a successful supply chain model that is currently widely in use. There are two most widely encountered objectives of the objective function in multi-objective programming models for SC. The first objective is the cost of activity in the supply chain, including the cost of particular chain link, transport, work and even product design, etc. The second objective is associated with the costs or volumes of CO2 emission and other environmental aspects [6]. The second objective may also comprise delivery time [7] or, calculated in many different ways, the level of customer satisfaction [8],[9]. In [10] the multi-objective optimization mathematical model of BOSC has been presented. There are three objective functions. The first is cost minimizing including order cost, purchase cost and transport cost. The second is minimizing the maximum time of transporting the purchased parts to customers (delivery). The last is the part quality. Many researchers have recently reported the results of their multi-objective optimization studies. For example, a multi-objective programming model is proposed in [2],[11] to analyze solid waste management. The model for simultaneously optimizing the operations of both integrated logistics and its corresponding used-product reverse logistics in a close-looped supply chain has been presented in [12]. The common feature of these problems is the number of decision variables resulting from the allocation of resources, choice of location, route selection, choice of factory and distribution center, choise o mode of transport etc. These are usually binary and/or integer decision variables. Besides, all the problems are characterized by a large number of constraints binding decision variables. The overview of the models and algorithms of these problems is shown in [2],[27].

### A. Multi-objective optimization

The multi-objective optimization problem (MOOP) can be defined as the problem of finding a vector of decision variables $\hat{x}$, which optimizes a vector of $N$ objective functions $f_i$ $(\hat{x})$ where $i = 1, 2, .. ,N$; subject to inequality constraints $g_j$ $(\hat{x}) \geq 0$ and equality constraints $h_k$ $(\hat{x}) = 0$ where $j = 1, 2, .. , J$ and $k = 1, 2, .. ,K$.

A set of objective function is a multi-dimensional space, in addition to typically the decision space. This additional space is called the objective space Z. For each solution $\hat{x}$ in the decision variable space, there exists a point in the objective space:

$$\hat{f}\ (\hat{x}\ )=Z\,(z_1\ , z_2\ ,..., z_N\ )^T$$

In a MOOP, we want to find a set of values for the decision variables that optimizes a set of objective functions. A decision vector $\hat{x}$ is said to dominate a decision vector $\hat{y}$ (i.e. $\hat{x}{>}\hat{y}$) if:

$$f_i\ (\hat{x}\ )\ \leq\ f_i\ (\hat{y})\ \forall\ i \in \{1,2,..,N\}$$

and

$$\exists i \in\ \{1,2,..,N\}\ |\ f_i\ \ (\hat{x}\ )\ \leq\ f_i\ (\hat{y})$$

All decision vectors that are not dominated by any other decision vector are called non-dominated or Pareto-optimal and constitute the Pareto-optimal front/set. There are several methods for find the Pareto-optimal set of these optimization problems. Among the most widely techniques are: ε-constraint method, weighting method, goal programming, sequential optimization etc. [13].

## III. Methods and Methodology

The key problem in the modeling and optimization of problems in the supply chain are multiple constraints of different types and character-linear, integer, non-linear, logical etc. Constraints are logical relations between variables, each variable taking a value from a specific domain. Thus a constraint restricts the possible values that a variable can take, i.e. it represents some partial information about the variables of interest. Constraints are:

- declarative, they specify a relationship between entities (decision variables) without determining a specific computational or programming procedure;
- additive, we are interested in the conjunction of constraints and not in the order in which they are imposed;
- rarely independent, normally constraints share decision variables.

Thus constraints are a natural medium and form to express problems in many fields, especially in logistic, transport, manufacturing, scheduling, distribution, supply chain etc. by all (researchers, practitioners, professionals, end-users etc.). In the above problems, there are resource, financial, capacity, time, transportation, environmental, multimodal, sale etc. constraints. Based on numerous studies and our own experience, the constraint-based environment [14], [15], [16], [17], [29] is believed to offer a very good framework for representing the knowledge, information and methods needed for the decision support and optimization. The central issue for a constraint-based environment is a constraint satisfaction problem (CSP) [14]. Constraint satisfaction problem is the mathematical problem defined as a set of elements whose state must satisfy a number of constraints. Constraint satisfaction problems (CSPs) on finite domains are typically solved using a form of search. The most widely used techniques include variants of backtracking, constraint propagation, and local search. Constraint propagation embeds any reasoning that consists in explicitly forbidding values or combinations of values for some variables of a problem because a given subset of its constraints cannot be satisfied otherwise [15]. CSPs are frequently used in constraint programming. Constraint programming is the use of constraints as a programming language to encode and solve problems. Constraint logic programming (CLP) is a form of constraint programming (CP), in which logic programming is extended to include concepts from constraint satisfaction. A constraint logic program is a logic program that contains constraints in the body of clauses (predicates). In CLP the declarative approach and the use of logic programming provide incomparably greater possibilities for decision problems modeling than the pervasive approach based on mathematical programming. Unfortunately, discrete optimization is not a strong suit of CP-based environments. This weakness is more pronounced in multi-objective optimization problems, where a parameterized problem of a single-objective optimization problem has to be solved multiple times depending on the size of Pareto set. Based on [14],[15] and previous work on hybridization [16],[17],[18] some advantages and disadvantages of these environments have been observed. The hybrid approach of constraint logic programming and mathematical programming can help to solve optimization problems that are intractable with either of the two methods alone [19],[20],[21]. In both MP and CLP, there is a group of constraints that can be solved with ease and a group of constraints that are difficult to solve. Both MP and finite domain CP/CLP involve variables and constraints. However, the types of the variables and constraints that are used, and the way the constraints are solved, are different in the two approaches [21]. MP relies completely on linear equations and inequalities in integer variables, i.e., there are only two types of constraints: linear arithmetic (linear equations or inequalities) and integrity (stating that the variables have to take their values in the integer numbers). In finite domain CP/CLP, the constraint language is richer. In addition to linear equations and inequalities, there are various other constraints: disequalities, nonlinear, symbolic (alldifferent, disjunctive, cumulative etc.) [14]. Integrity constraints are difficult to solve using mathematical programming methods and often the real problems of MP make them NP-hard. In CP/CLP, domain constraints with integers are easy to solve. The system of such constraints can be solved over integer variables in polynomial time. The inequalities between many variables, general linear constraints, and symbolic constraints are difficult to solve, which makes real problems in CP/CLP NP-hard [21]. This type of constraints reduces the strength of constraint propagation. As a result, CP/CLP is incapable of finding even the first feasible solution [16].

## IV. Motivation and Contribution

The motivation and contribution behind this work was to apply a hybrid approach as a hybrid multi-objective programming framework for supply chain problems. The hybrid multi-objective programming framework is a concept that combines hybrid approach with iterative algorithm (Appendix A) in the context of multi-criteria optimization. The hybrid approach proved to be very effective when applied to a single objective optimization problems [17],[18]. This hybrid approach is an original concept whose elements and outline are presented in [16],[17],[18]. Application of this approach to multi-objective optimization has not been presented before. The best structure for the implementation of the above approach is a declarative CLP environment with operation research MP as a hybrid system. Furthermore, such a hybrid approach allows the use of all layers of the problem (data, structure, methods) to solve it. Finally, it allows the transformation of the problem (Section VIC) to such a form that can fully exploit the strengths of the constraint propagation and data instances. It is well-known that there exist multiple non-dominated solutions for a multi-objective optimization problem. Those solutions are called "Pareto-optimal" solutions. In this paper, our objective is to obtain a "Pareto-optimal" set which provides evenly distributed Pareto solutions and it is convenient for the decision maker to select a suitable costs between production, distribution and environmental (F1 and F2) or between all costs and total distributor capacity (F1' and F2') etc. (Section VIA and Section VID). This hybrid programming framework is not just a blind attempt to integrate two environments, CLP/MP. The proposed approach is reinforced with the transformation, different representation of the problem (Section VIC) and using the algorithm for finding a "Pareto-optimal" set. In addition, hybridization refers to the class of decision problems which has certain property (This property is characterized by the constraints of many discrete decision variables and their summation).

## V. The concept and implementation aspects of the hybrid multi-objective programming framework

In this approach to the modeling and multi-objective optimization of supply chain problems, the hybrid multi-objective programming framework has been proposed, where:

- the decision and optimization models solved using the proposed framework can be formulated as a pure model of MOOP/MOLP or a hybrid model (with logical and non-linear constraints);
- knowledge related to the problem can be expressed as facts and constraints (linear, non-linear, logical and symbolic etc.);
- the problem is modeled in the constraint logic programming environment by CLP predicates, which is far more flexible than the MP environment;
- transforming the optimization model to explore its structure and data has been introduced by CLP predicates;
- optimization is performed by MP-based environments.
- the effective algorithm for finding a "Pareto-optimal" set has been introduced.

The schematic diagram of the implementation framework for the hybrid approach is presented in Figure 1. The names and descriptions of the predicates and procedures are shown in Table 1. From a variety of tools for the implementation of the CP-based environment, ECL$^i$PS$^e$ software [22] was selected. ECL$^i$PS$^e$ is an open-source software system for the cost-effective development and deployment of constraint programming applications. MP-based environment in implementation platform was LINGO by LINDO Systems [23]. LINGO Optimization Modeling Software is a powerful tool for building and solving mathematical optimization models. ECL$^i$PS$^e$ was used to implement the following predicates of the framework: CLP1, CLP2, CLP3 and CLP4 (Fig. 1, Table 1.). CLP predicates significantly restrict the space of feasible solutions (Fig. 1). The transformed files of the model were transferred from ECLiPSe to LINGO where they were merged (MP1). Then the complete model was solved using LINGO efficient solvers (MP2). After that the model was modified for next step in procedure MP1 and solved again and again by MP2 in order to find a set of Pareto-optimal solutions (algorithm see appendix A).

TABLE I.
THE NAMES AND DESCRIPTIONS OF THE PREDICATES AND PROCEDURES

| Symbol | Description |
|---|---|
| CLP1 | The implementation of the model in CLP, the term representation of the problem in the form of predicates. |
| CLP2 | Constraint propagation for the model or transformed model. Constraint propagation is one of the basic methods of CLP. As a result, the variable domains are narrowed, and in some cases, the values of variables are set, or even the solution can be found. |
| CLP3 | The transformation of the original problem aimed at extending the scope of constraint propagation. The transformation uses the structure of the problem and data. The most common effect is a change in the representation of the problem by reducing the number of decision variables, and the introduction of additional constraints and variables, changing the nature of the variables, etc. |
| CLP4 | Generation the model for mathematical programming-MOOP; |
| MP1 | Merging files generated by CLP4 into one file. It is a model file format in LINGO system. |
| MP2 | The solution of the model from the MP1 by MP solver. Sending the new solution to MP1 as a new constraint. Modification of the model. The process is repeated until the whole Pareto set is found (algorithm Appendix A). |

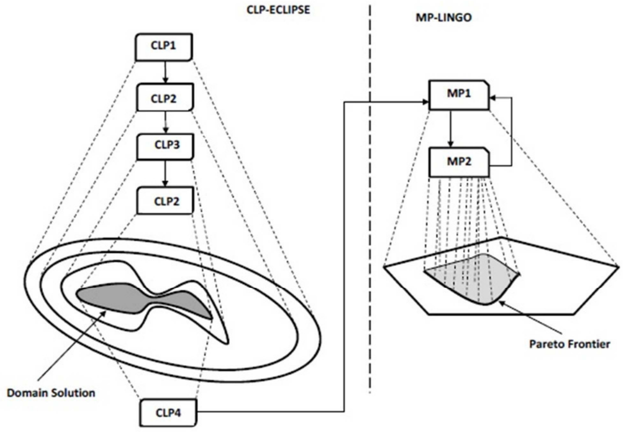


Fig. 1 The scheme of the hybrid multi-objective programming framework

## VI. ILLUSTRATIVE EXAMPLES OF SUPPLY CHAIN MODELING AND MULTI-OBJECTIVE OPTIMIZATION - THE HYBRID APPROACH

The proposed approach was used and tested on two illustrative supply chain multi-objective optimization models. The model was formulated as a multi-objective mixed-integer optimization problem (MOOP) problem based on under constraints (2) .. (23) in order to test the proposed approach (Fig. 1) against the classical operation research approach. The models differ from one another objective functions (Section 6.A). Indices, parameters and decision variables used in the models together with their descriptions are summarized in Table 2 and [16],[17]. The simplified structure of the supply chain network for this model, composed of producers, distributors and customers is presented in Figure 2. Both models are the cost multi-objective models that take into account other types of parameters, i.e., the spatial parameters (area/volume occupied by the product, distributor capacity and capacity of transport unit), time (duration of delivery and service by distributor, etc.), the transport mode and environmental aspects.

These models are based on optimization models taken from [16],[17].

TABLE II.
INDICES, PARAMETERS AND DECISION VARIABLES

| Symbol | Description |
|---|---|
| | *Indices* |
| k | product type (k=1..K) |
| m | delivery point/customer/city (m=1..M) |
| n | manufacturer/factory (n=1..N) |
| e | distributor /distribution center (e=1..E) |
| d | mode of transport (d=1..D) |
| N | number of manufacturers/factories |
| M | number of delivery points/customers |
| E | number of distributors |
| K | number of product types |
| D | number of mode of transport |
| | *Input parameters* |
| $F_e$ | the fixed cost of distributor/distribution center *e* |
| $C_{n,k}$ | the cost of product *k* at factory *n* |
| $K1_{n,e,k,d}$ | the variable cost of delivery of product *k* from manufacturer *n* to distributor *e* using mode of transport *d* |
| $A_{n,e,d}$ | the fixed cost of delivery from manufacturer *i* to distributor *s* using mode of transport *d* |
| $K2_{e,m,k,d}$ | the variable cost of delivery of product *k* from distributor *e* to customer *m* using mode of transport *d* |
| $G_{e,m,d}$ | the fixed cost of delivery from distributor *e* to customer *m* using mode of transport *d* |
| $Od_d$ | the environmental cost of using mode of transport *d* |
| | *Decision variables* |
| $X_{n,e,k,d,m}$ | delivery quantity of product *k* from manufacturer *n* to distributor *s* using mode of transport *d* to customer *m* |
| $Xa_{n,e,k,d,m}$ | if delivery is from manufacturer *i* to distributor *e* product *k* using mode of transport *d* to customer <u>*m*</u> then $Xa_{n,e,k,d,m}$ =1, otherwise $Xa_{n,e,k,d,m}$ =0 |
| $Xb_{n,e,d}$ | the number of courses from manufacturer *n* to distributor *m* using mode of transport *d* |
| $Y_{e,m,k,d}$ | delivery quantity of product *k* from distributor *e* to customer *m* using mode of transport *d* |
| $Yb_{e,m,d}$ | the number of courses from distributor *e* to customer *m* using mode of transport *d* |
| $Tc_e$ | if distributor *e* participates in deliveries, then $Tc_e$=1, otherwise $Tc_e$=0 |

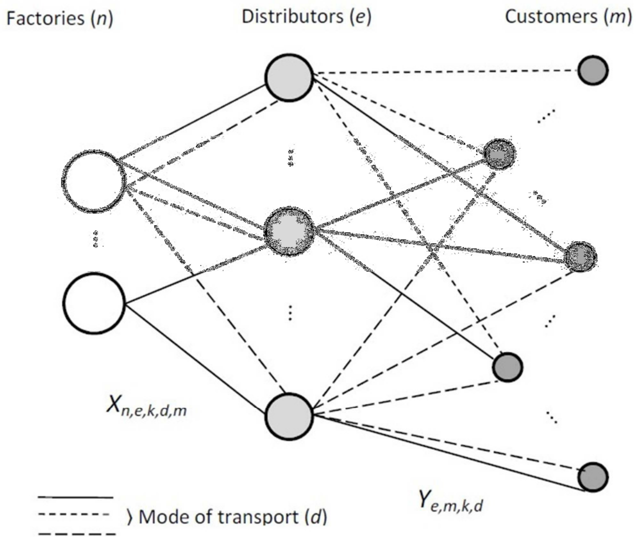


Fig. 2 Simplified supply chain network

### A. *Objective function*

Two typical objective functions, F1 and F2, are commonly used in optimization issues. Objective function F1 (1a) defines the aggregate costs of the entire chain and consists of four elements. The first element comprises the fixed costs associated with the operation of the distributor involved in the delivery (e.g. distribution centre, warehouse, etc.). The second component determines the cost of the delivery from the manufacturer to the distributor. Another component is responsible for the costs of the delivery from the distributor to the end user (the store, the individual client, etc.). The last component of the objective function F1 determines the cost of manufacturing the product by the given manufacturer. The second objective function F2 (1b) corresponds to environmental costs of using various means of transport. Those costs are dependent on the number of courses of the given means of transport, and on the other hand, on the environmental levy, which in turn may depend on the use of fossil fuels and carbon-dioxide emissions [8]. This hybrid approach and its implementation can be successfully used for other objective functions, including those named in Section 3. For the numerical examples from Section 7, in addition to the objective function formulated as above, the objective functions where F1'=F1+F2 (sum (1a) and (1b), whereas F2'=V (total capacity of distribution centers) were formulated.

$$\begin{aligned} F1 = {} & \sum_{e=1}^{E} F_e \cdot Tc_e + \sum_{n=1}^{N}\sum_{e=1}^{E}\sum_{d=1}^{D}(A_{n,e,d} \cdot Xb_{n,e,d} + \sum_{k=1}^{O}\sum_{m=1}^{M}(K1_{n,e,k,d} * X_{n,e,k,d,m})) + \\ & \sum_{e=1}^{E}\sum_{m=1}^{M}\sum_{d=1}^{D}(G_{e,m,d} \cdot Yb_{e,m,d} + \sum_{k=1}^{O} K2_{e,m,k,d} \cdot Y_{e,m,k,d}) + \sum_{n=1}^{N}\sum_{k=1}^{K}(C_{nk} \cdot \sum_{e=1}^{E}\sum_{d=1}^{D}\sum_{m=1}^{M} X_{n,e,k,d,m}) \end{aligned} \tag{1a}$$

$$F2 = \sum_{d=1}^{D} Od_d (\sum_{n=1}^{N}\sum_{e=1}^{E} Xb_{n,e,d} + \sum_{e=1}^{E}\sum_{m=1}^{M} Yb_{e,m,d}) \tag{1b}$$

### B. *Two typical objective functions, F1 and F2, are commonly*

The model was based on constraints (2) .. (23). Constraint (2) specifies that all deliveries of product *k* produced by the manufacturer *n* and delivered to all distributors *e* using mode of transport *d* do not exceed the manufacturer's production capacity. Constraint (3) covers all customer *m* demands for product *k* through the implementation of delivery by distributors *s* (the values of decision variables $Y_{e,m,k,d}$). The flow balance of each distributor *e* corresponds to constraint (4). The possibility of delivery is dependent on the distributor's technical capabilities (5). Time constraint (6) ensures the terms of delivery are met. Constraints (7), (8), (9), (10), (11) guarantee deliveries with available transport taken into account. Constraints (12), (13), (14) set values of decision variables based on binary variables $Tc_e$, $Xa_{n,s,k,d,m}$, $Ya_{e,m,k,d}$, $X_{n,e,k,d,m}$, $Y_{e,m,k,d}$. Dependencies (15) and (16) represent the relationship based on which total costs are calculated. In general, these may be any linear functions. The remaining constraints (17) .. (23) arise from the nature of the model (MILP). A detailed description of the constraints and their formalization have been presented in [16],[17].

### C. *Model transformation*

The ability to transform the problem using CLP is one of the most important features of the hybrid programming framework. Due to the nature of the decision problem (adding up variables in the objective function and constraints), the constraint propagation efficiency decreases dramatically. Constraint propagation is one of the most important methods in CLP affecting the efficiency and effectiveness of the CLP and hybrid programming framework (Fig. 1). For that reason, research into more efficient and more effective methods of constraint propagation was conducted. The results included different representation of the problem and the manner of its implementation. The classical problem modeling in the CLP environment consists in building a set of CLP predicates with parameters. While modeling problem (1) .. (23), quantities *n, e, k, d1, d2, m* and decision variables $X_{n,e,k,d1,m}$ and $Y_{e,m,k,d2}$ were predicate parameters (Fig. 3a). The process of finding the solution may consist in using the constraints propagation methods, labeling of variables and the backtracking mechanism [15]. The quality of constraints propagation and the number of backtrackings are affected to a high extent by the number of parameters that must be specified/labeled in the given predicate. In the models presented above, the classical problem representation included four parameters (Fig. 3b): n, e, d1, d2, and two decision variables $X_{n,e,k,d1,m,j}$, $Y_{e,m,k,d2}$. Considering the domain size of each parameter, the process is complex and time-consuming. The idea was to transform the problem by changing its representation without changing the very problem. All permissible routes were first generated by CLP3 predicate based on the fixed data (factories, distributors, mode of transport etc.) and a set of orders (Fig. 4a), then the specific values of parameters *n, e, k, d1, d2*, m were assigned to each of the routes. Thereby only one parameter of the transformed decision variable $X^T_{n,k,e,j,d1,d2}$ (deliveries) had to be specified (Fig. 4b). This transformation fundamentally improved the efficiency of the constraint propagation and reduced the number of backtracks.

***Solution* (Objective_Function, parameters) :-**
**parameters /O_n,P,M,D,F,Tu,Tu,Oq,X,Y,T/**

Fig 3a. Representation of the problem in the classical approach-the main search predicate

***Solution*** **($V_{FC}^{1}$,o_1,$p_1$,$m_1$,_,_,_,_,12,_,_,8)**
**($V_{FC}^{1}$,o_2,$p_2$,$m_4$,_,_,_,_,20,_,_,6)**
**($V_{FC}^{1}$,o_4,$p_3$,$m_2$,_,_,_,_,12,_,_,9)**
**…**

Fig. 3b. Representation of the problem in the classical approach- the process of finding a solution

***Solution_hybrid* (Objective_Function, parameters) :-**
**parameters /route_n,P,M,D,F,Tu,Tu,Oq,$X^T$, T/**

Fig. 4a. Representation of the problem in the hybrid approach- the main search predicate

***Solution_hybrid*** **($V_{FC}$,((route_1,f1,p1,c1,m1,s1,s1,5,10,100,_8)**
**(route_2,f1,p2,c1,m1,s1,s1,5,15,120,_12),**
**(route_2,f1,p2,c1,m1,s2,s1,5,15,80,_8),**
**(route_4,f2,p2,c2,m1,s1,s2,5,12,20,_6),…), …)**

Fig. 4b. Representation of the problem in the hybrid approach- set of feasible routes

Symbols used in descriptions are presented in Table III.

TABLE III.
INDICES, SYMBOLS USED IN THE REPRESENTATION OF THE PROBLEM

| Symbol | Description |
|---|---|
| $V_{FC}$ | Value of the objective function calculated on the basis of the vector of parameters. |
| O_n | Order number. |
| P | Products, P ∈ {$p_1$,$p_2$, ... ,$p_o$}. |
| M | Customers, M ∈ {$m_1$,$m_2$, … ,$m_m$}. |
| D | Distributors, D ∈ {$c_1$,$c_2$, … ,$c_e$}. |
| F | Factories, F ∈ {$f_1$,$f_2$, … ,$f_n$}. |
| Tu | Transport unit, Tu ∈ {$s_1$,$s_2$, … ,$s_l$}. |
| T | Delivery time/period. |
| Oq | Order quantity. |
| X/Y/$X^T$ | Delivery quantity. |
| route_n | Routes name-number. |
| , | Separates predicate parameters. |
| :- | Separates predicate heading from its definition. |
| _ | Unknown value of the variable. |

The obtained multi-objective optimization model after the transformation (MOOPT) has different decision variables and different constraints than those in the MOOP (1) .. (23). Some of the decision variables are redundant; other variables are subject to aggregation. This results in a very large reduction in their number. Decision variables before and after the transformation are shown in Table IV. The transformation also reduces or eliminates some of the constraints of the model. Thus constraints (4), (6), (12), (13), (14), (15) and (16) present in the MOOP (1) .. (23) are redundant in the MOOPT. Balance constraint (4) is unnecessary because the route defines the specific distribution center. Only those routes are generated that meet the time constraints, therefore constraint (6) does not make sense. Binarity ensures whether or not the route occurs, thus constraint (12) is redundant. Reduction of certain variables also affects the reduction of constraints, hence lack of constraints (13), (14) in the model. Constraints (15) and (16) are unnecessary, because the delivery costs are now calculated for the entire route.

TABLE IV.
DECISION VARIABLES USED IN THE MOOP AND TRANSFORMED MOOPT MODELS

| MOOP | MOOPT | Description of the decision variables after the transformation |
|---|---|---|
| $X_{n,e,k,d1,m}$<br>$Y_{e,m,k,d2}$ | $X^T_{n,k,e,m,d1,d2}$ | Decision variable $X^T$, unlike the initial decision variables X,Y, is generated only for technologically possible indices combinations. It defines the allocation size of product k to the route of deliveries. |
| $Xa_{n,e,k,d,m}$ | unnecessary | After transformation replaced by the appropriate factor for the route - generated by the CLP. |
| $Xb_{n,e,d}$ | $Xb_{n,e,d}$ | Without change, the same sense. |
| $Ya_{e,n,k,d}$ | unnecessary | After transformation replaced by the appropriate factor for the route - generated by the CLP. |
| $Yb_{e,m,d}$ | $Yb_{e,m,d}$ | Without change, the same sense. |
| $Tc_e$ | $Tc_e$ | Without change, the same sense. |

After the transformation in the MOOPT model, the objective functions F1 and F2 were re-formulated. New objective functions, F1T (A1a) and F2T (A1b) were obviously formulated using new decision variables (Table IV) and calculated parameters by CLP (Table V). These parameters were determined as a result of constraint propagation and the transformation itself using CLP2 and CLP3. Owing to these quantities, it is possible to introduce to the MOOPT model additional constraints (A2) ... (A7). These constraints affect the efficiency of the search for a solution by narrowing down the search area. Table VI describes these constraints.

$$\begin{aligned} F1T = &\sum_{e=1}^{E} Tc_e \cdot F_e + \sum_{n=1}^{N}\sum_{e=1}^{E}\sum_{d=1}^{D}(Xb_{n,e,d} \cdot Ksc_{n,e,d}) + \\ &\sum_{e=1}^{E}\sum_{m=1}^{M}\sum_{d=1}^{D}(Yb_{e,m,d} \cdot Ksm_{e,m,d}) + \\ &\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{e=1}^{E}\sum_{m=1}^{M}\sum_{d_1=1}^{D}\sum_{d_2=1}^{D}(X_{n,k,e,m,d_1,d_2} \cdot Kz_{n,k,e,m,d_1 d_2}) \end{aligned} \tag{A1a}$$

$$F2T = \sum_{d=1}^{D} Od_d \cdot (\sum_{n=1}^{N}\sum_{e=1}^{E} Xb_{n,e,d} + \sum_{e=1}^{E}\sum_{m=1}^{M} Yb_{e,m,d}) \tag{A1b}$$

$$\sum_{n=1}^{N}\sum_{e=1}^{E} Xb_{n,e,d} + \sum_{e=1}^{E}\sum_{m=1}^{M} Yb_{e,m,d} \geq R\,min_d \text{ for } d = 1..D \tag{A2}$$

$$\sum_{n=1}^{N}\sum_{e=1}^{E} Xb_{i,s,d} + \sum_{e=1}^{E}\sum_{m=1}^{M} Yb_{e,m,d} \leq R\,max_d \text{ for } d = 1..D \tag{A3}$$

$$\sum_{n=1}^{N}\sum_{e=1}^{E}\sum_{d=1}^{D} Xb_{n,e,d} \geq Min\_F\_C \tag{A4}$$

$$\sum_{e=1}^{E}\sum_{j=1}^{M}\sum_{d=1}^{D} Yb_{e,m,d} \geq Min\_D\_C \tag{A5}$$

$$\sum_{n=1}^{N}\sum_{e=1}^{E}\sum_{d=1}^{D} Xb_{n,e,d} + \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{d=1}^{D} Yb_{e,m,d} \geq Min\_TU \tag{2}$$

$$\sum_{e=1}^{E} Tc_e \geq Cn \tag{A7}$$

TABLE V.
CALCULATED FIGURES FOR MOOPT

| Symbol | Description |
|---|---|
| $Rmin_d$ | Minimum number of transport units d (CLP – propagation). |
| $Rmax_d$ | Maximum number of transport units d (CLP – propagation). |
| Min_F_C | Minimum number of transport units in the route Factories – Centers (CLP – propagation). |
| Min_D_C | Minimum number of transport units in the route Distributors – Customers (CLP – propagation). |
| Min_TU | Minimum number of transport units (CLP – propagation). |
| Cn | Minimum number of active centers (CLP – propagation). |
| Ksc | The fixed cost of delivery from manufacturer to distributor by using transport mode (CLP-calculated based on fixed data). |
| Ksm | The fixed cost of delivery from distributor to customer by using transport mode (CLP-calculated based on fixed data). |
| Kz | The variable cost of delivery (CLP-calculated based on fixed data). |

TABLE VI.
ADDITIONAL CONSTRAINTS FOR MOOPT

| Constraints | Description |
|---|---|
| A2 | narrowing the size of the transport unit domain from the bottom |
| A3 | narrowing the size of the transport unit domain from the top |
| A4 | the minimum number of all transport unit types necessary for the shipment from the factory to the distribution center |
| A5 | the minimum number of all transport unit types necessary for the shipment from the from the center to customers |
| A6 | the minimum number of transport units in routes |
| A7 | the number of working distribution centers |

### *D. Decision support*

Implementation of the presented models using implementation hybrid programming framework can support decision-making in the following practical areas of supply chain (not limited to the following):

- the multi-optimization of total cost of the supply chain (objective functions F1 and F2/ F1' and F2' -Table VII) in the form of the Pareto-optimal solution set (Fig.5.);
- analysis of delivery costs with environmental cost optimization;
- the selection of the transport fleet number, capacity and modes for specific total costs (Fig.6, Fig.8);
- the sizing of distributor warehouses and the study of their impact on the overall costs (objective functions F1' and F2' -Table VII and Fig.7);
- the sizing production capacity and the study of their impact on the overall costs;
- the selection of transport routes for elements of the Pareto-optimal solution set;

## VII. NUMERICAL EXPERIMENTS AND ANALYSIS

In order to verify and evaluate the proposed hybrid approach and implementation platform, many numerical experiments were performed. All the examples relate to the supply chain with seven manufacturers (*n=1..7*), three distributors (*e=1..3*), ten customers (*m=1..10*), three modes of transport (*d=1..3*), and twenty types of products (*k=1..20*). The numerical data were taken from the trans-regional distributor FMCG and transportation fleet parameters available online. Experiments began with four examples of P1 .. P4 for the optimization MOOP model (1) .. (23). The examples differ the number of orders (No). The first series of experiments was designed to show the benefits and advantages of the presented approach. For this purpose the model was implemented in both the hybrid programming framework (MOOPT) and mathematical programming environment (MOOP). In the next stage of the experiments, the objective function was changed (Section 6.A). Its first element F1' refers to the total cost of deliveries whereas F2' defines the distributor's total storage capacity available.

Due to the nature of the optimization problems considered here, a new algorithm based on the ε-constraint method was proposed in the final phases of MP1 and MP2. This algorithm (Appendix A) helped determine a set of Pareto-optimal solutions. The algorithm has been implemented in LINGO by using meta-modeling and programming language LINGO package.

The detailed results for a 20-order example (P4) are shown in Table 7. Figures 5 and 6 show the corresponding sets of Pareto-optimal solutions for a 20-order example. It is evident that only the hybrid approach provides the results within the acceptable time (Table VIIIA and VIIIB).

This is possible owing to the reduction of the problem size and, in particular, transformation of the problem with the use of methods from both environments, MP and CLP. For the illustrative examples discussed here, a 23-fold reduction in the number of constraints (C) was obtained with an over 130-fold reduction in the number of decision variables (V). This gives the size of the combinatorial problem calculated as VxC reduced by more than 2600 times. Comparison of the results (Table 8a and 8b) from the hybrid approach with those from mathematical programming indicates that, first, when the same model and data are used in the classical manner (mathematical programming), the Pareto-optimal solution was obtained only for the smallest example P1 (5 orders). Second, for larger examples P2, P3, P4, it was hardly possible to find at least one point of the Pareto-optimal solution in acceptable time (computing stopped after 600 s). Comparing the proposed approach to modeling and solution in only CLP environment is pointless due to the nature and weak capacity of the CLP relative to the optimization of problems where many variables are added up, which is illustrated in [16].

TABLE VII.
THE DETAILED RESULT FOR EXAMPLE P4 (NO=20)

| PP | F1 | F2 | PP | F1' | F2'=V |
|---|---|---|---|---|---|
| 1 | F1≥26540 | F2=3775 | 1 | F1'=0 | F2'≤0 |
| 2 | F1≥29345 | F2=3725 | 2 | F1'=0 | F2'≤500 |
| 3 | F1≥32150 | F2=3725 | 3 | F1'=0 | F2'≤1000 |
| 4 | F1≥34955 | F2=3655 | 4 | F1'=0 | F2'≤1500 |
| 5 | F1≥37760 | F2=3655 | 5 | F1'=45195 | F2'≤2000 |
| 6 | F1≥40565 | F2=3655 | 6 | F1'=35705 | F2'≤2500 |

| | | | | | |
|---|---|---|---|---|---|
| 7 | F1≥43370 | F2=3655 | 8 | F1'=35705 | F2'≤3000 |
| 8 | F1=33670 | F2≥3655 | 9 | F1'=35705 | F2'≤3500 |
| 9 | F1=26540 | F2≥4158 | 10 | F1'=35705 | F2'≤4000 |
| 10 | F1=26540 | F2≥4661 | | | |
| 11 | F1=26540 | F2≥5164 | | | |
| 12 | F1=26540 | F2≥5668 | | | |
| 13 | F1=26540 | F2≥6171 | | | |
| 14 | F1=26540 | F2≥6075 | | | |

PP-the number of element of Pareto-optimal set F1, F2, F1', F2'-objective function

### TABLE VIIIA. THE PARAMETERS OF THE PROCESS OF FINDING A SET OF PARETO-OPTIMAL SOLUTIONS FOR ILLUSTRATIVE EXAMPLES (OBJECTIVE FUNCTIONS F1 AND F2)

| Example | No | MP-based approach (MOOP) | | | Hybrid programming framework (MOOPT) | | |
|---|---|---|---|---|---|---|---|
| | | T | V | C | T | V | C |
| P1 | 5 | 53 | 30369 | 16177 | 7 | 223 | 693 |
| P2 | 10 | ----* | 30369 | 19122 | 10 | 264 | 697 |
| P3 | 15 | ----* | 30369 | 20067 | 14 | 287 | 702 |
| P4 | 20 | ----* | 30369 | 21012 | 44 | 318 | 703 |

* calculations stopped after 600 s, not found even one point from the set of Pareto-optimal solutions

T-solution time, V- the number of integer variables, C- the number of constraints

### TABLE VIIIB. THE PARAMETERS OF THE PROCESS OF FINDING A SET OF PARETO-OPTIMAL SOLUTIONS FOR ILLUSTRATIVE EXAMPLES (OBJECTIVE FUNCTIONS F1' AND F2')

| Example | No | MP-based approach (MOOP) | | | Hybrid programming framework (MOOPT) | | |
|---|---|---|---|---|---|---|---|
| | | T | V | C | T | V | C |
| P1 | 5 | 70 | 30369 | 16179 | 11 | 223 | 691 |
| P2 | 10 | ---- | 30369 | 19124 | 14 | 264 | 695 |
| P3 | 15 | ---- | 30369 | 20069 | 21 | 287 | 700 |
| P4 | 20 | ---- | 30369 | 21014 | 36 | 318 | 701 |
| P5 | 25 | ---- | 30369 | 21959 | 67 | 357 | 707 |

* calculations stopped after 600 s, not found even one point from the set of Pareto-optimal solutions

T-solution time, V- the number of integer variables, C- the number of constraints

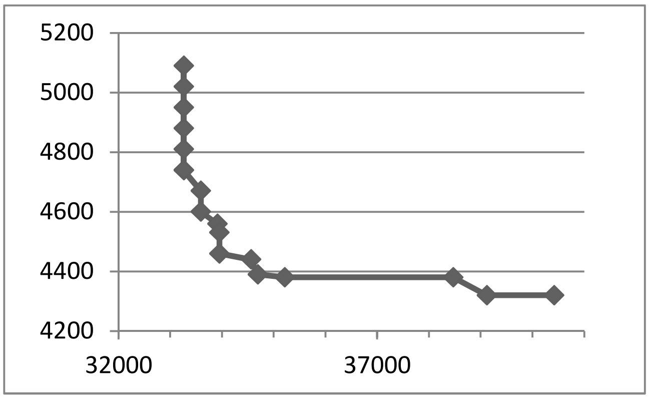


Fig. 5 A set of Pareto-optimal solutions for illustrative example P4 (No=20, vertical axis F2 horizontal axis F1)

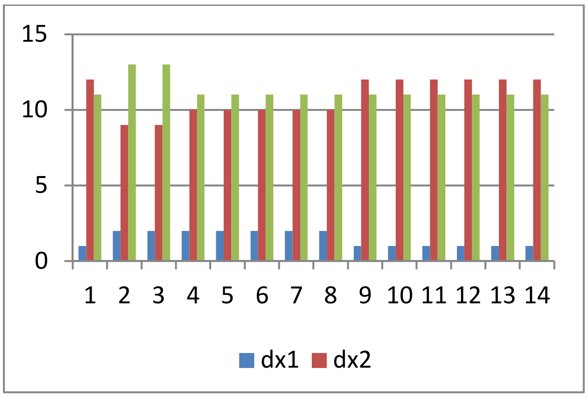


Fig. 6 The use of mode of transport for illustrative example P4 (No=20, F1, F2, vertical axis the number of mode of means of transport-dx1,dx2,dx3, horizontal axis Pareto set of points -PP)

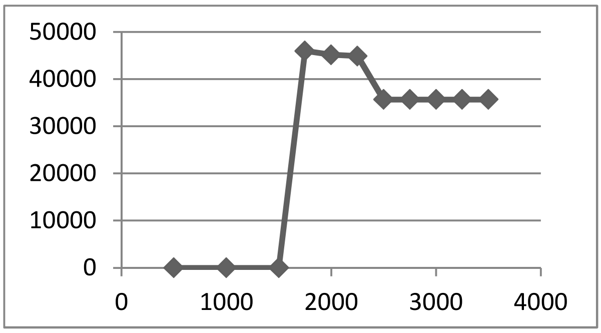


Fig. 7 A set of Pareto-optimal solutions for illustrative example E4 (No=20, vertical axis F2' horizontal axis F1')

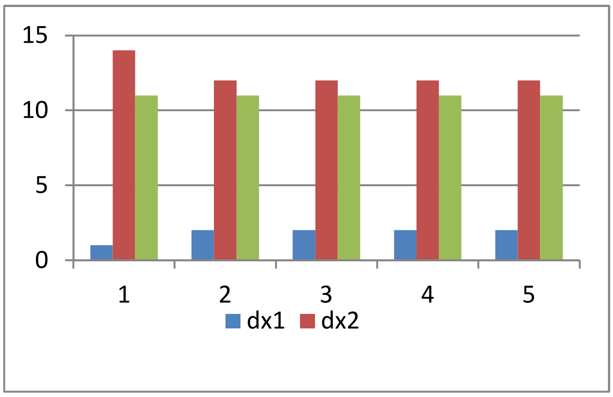


Fig. 8 The use of mode of transport for illustrative example P4 (No=20, F1',F2', vertical axis the number of mode of means of transport-dx1,dx2,dx3, horizontal axis Pareto set of points –PP for F1'>0 )

## VIII. CONCLUSIONS

This hybrid programming framework is especially significant and suited for multi-objective optimization, where a slightly changed single-objective problem has to be solved multiple times and where the modeling efficiency and ease are essential. The efficiency of the proposed approach is based on the reduction of the combinatorial problem. This means that using the hybrid approach practically for all models of this or a similar class, the same or better solutions are found even up to two hundred times faster (the optimal instead of the feasible solutions). Another element contributing to the high efficiency of the method is a possibility to determine the values or ranges of values for some of the decision variables (predicate CLP2). The presented transformation of the problem (predicate CLP3), characteristic of the problems that have the structure as in Fig. 2, is an important aspect of this approach. It should be emphasized that with this approach it is possible not only to solve optimization problems faster, but also to solve much larger problems than in the [24]. The proposed solution is highly recommended for all types of decision problems in supply chain or for other problems with a similar structure. This structure is characterized by the constraints of many discrete decision variables and their summation. Furthermore, this method can model and solve problems with logical constraints. Therefore the implementations in the form of hybrid platform can be applied to various practical decision problems in the area of logistics, transport, production, scheduling or project management. In addition to the undoubted effectiveness of the proposed declarative hybrid approach, we should underline the possibility of modeling decision problems.

Further work will focus on running the optimization models with non-linear and other logical constraints, uncertainty, fuzzy logic [25] etc., numerical test with hybrid models (HM) and different scheduling problems and resource allocation and activity coordination in the supply chain [26]. It is also planned to implement the framework in the form of cloud applications [28].

## APPENDIX ALGORITHM FOR FINDING A SET OF PARETO-OPTIMAL SOLUTIONS.

```
Enter the step (for how many points divided
interval)
Number_of _steps =? /input value/
Solve (min objective F1
  subject to
  the constraints of 1 - 27 (primary problem)
  or A1-A7 (problem after transformation)
)
Save the F1min-determined value of the objective
function
F2max = F2
Solve (min objective F2
  subject to
  the constraints of 1 - 27 (primary problem)
  or A1-A7 (problem after transformation)
)
Save the F2min-determined value of the objective
function
F1max = F1
There are designated intervals of a function F1
and F2
(F1min, F1max) and (F2min,F2max)
Optimization of F2
discretization = (F1max - F1min) / Number_of
_steps
cutting_off = F1min
i = 1
WHILE (cutting_off <F1max)
{
 Solve (min F2
  subject to
   the constraints of 1 - 27 (primary problem)
   or A1-A7 (problem after transformation)
   F1> = cutting_off
 )
 save the pareto-optimal point
 F2 (i) - optimal value of the objective function
 F1(i) = F1
 cutting_off cutting_off + = discretization
 i = i +1
}
Optimization of F1
discretization = (F2max - F2min) / Number_of
_steps
cutting_off = F2min
WHILE (cutting_off <F2max)
{
 Solve (min F1
  subject to
   the constraints of 1 - 27 (primary problem)
   or A1-A7 (problem after transformation)
 F2> = cutting off
 )
 save the point
 F1 (i) - optimal value of the objective function
 F2 (i) = F2
 Cutting_off=cutting off + discretization
 i = i +1
}
F1=F1 or F1'
F2=F2 or F2'
```

## REFERENCES

[1] Beamon B.M, Supply chain design and analysis: models and methods, International Journal of Production Economics 55, 281–294, 1998.

[2] Mula J., Peidro D., Diaz-Madronero M., Vicens E., Mathematical programming models for supply chain production and transport planning, European Journal of Operational Research, 204, 377–390, 2010.

[3] Gunasekaran A., Ngai EWT., Modeling and analysis of build-to-order supply chains. Eur J Oper Res 2009;195(2):319–34, 2009.

[4] Howard M, Miemczyk B.J, Graves A., Automotive supplier parks: an imperative for build-to-order. J Purch Supply Manage 2006, 12, 91–104, 2006

[5] Krajewski L., Wei J.C., Tang L.L., Responding to schedule changes in build-to-order supply chains. J Oper Manage 2005;23(5):452–69, 2005.

[6] Abdolhossein S., Napsiah I., Norzima Z., Ariffin M. K. A., Nezamabadi-pour H., Mirabi H., A Multiobjective Optimization Model in Automotive Supply Chain Networks, Mathematical Problems in Engineering, vol. 2013, Article ID 823876, doi:10.1155/2013/823876, 2013.

[7] Minor P., Hertwin O., Elías O.B., Ruben T.O., Luis M., Variations in the Flow Approach to CFCLP-TC for Multiobjective Supply Chain Design, Mathematical Problems in Engineering, vol. 2014, Article ID 816286, doi:10.1155/2014/816286, 2014.

[8] Seyed M, Al-e-Hashem J.M., Aryanezhad M.B, Sadjadi S.J., An efficient algorithm to solve a multi-objective robust aggregate production planning in an uncertain environment, The International Journal of Advanced Manufacturing Technology, January 2012, Volume 58, Issue 5-8, 765-782, 2012.

[9] Shankar L., Basavarajappa B., Chen S., Jason C.H, Kadadevaramath, Rajeshwar S., Location and allocation decisions for multi-echelon supply chain network - A multi-objective evolutionary approach, Expert Systems with Applications, Volume 40, Issue 2, 1 February 2013, 551–562, 2013

[10] Che Z.H., Chiang C.J., A modified Pareto genetic algorithm for multi-objective build-to-order supply chain planning with product assembly, Advances in Engineering Software 41, 1011–1022, 2010.

[11] Wang F., Lai X., Shi N., A multi-objective optimization for green supply chain network design, Decision Support Systems, 51, 262–269, 2011.

[12] Erenguc S. S., Simpson N. C., Vakharia A. J., Integrated production/distribution planning in supply chains: an invited review, European Journal of Operational Research, 115, 219–236, 1999.

[13] Collette Y., Siarry P., Multiobjective Optimization, Principles and Case Studies, Springer, IX, 293 p, 2003.

[14] Apt K., Wallace M., Constraint Logic Programming using Eclipse, Cambridge University Press, 2006

[15] Rossi F., Van Beek P., Walsh T., Handbook of Constraint Programming (Foundations of Artificial Intelligence), Elsevier Science Inc. New York, NY, USA, 2006

[16] Sitek, P., Wikarek, J., A hybrid approach to modeling and optimization for supply chain management with multimodal transport, IEEE Conference: 18th International Conference on Methods and Models in Automation and Robotics (MMAR), 777-782, 2013.

[17] Sitek, P., A hybrid CP/MP approach to supply chain modelling, optimization and analysis, Federated Conference on Computer Science and Information Systems (FedCSIS), 2014.pp.1345-1352. DOI: 10.15439/2014F89.

[18] Sitek, P., Wikarek, J.: A Hybrid Approach to the Optimization of Multiechelon Systems. Mathematical Problems in Engineering vol. 2015, Article ID 925675, 12, pages, 2015. doi:10.1155/2015/925675.

[19] Milano M., Wallace M., Integrating Operations Research in Constraint Programming, Annals of Operations Research, vol. 175 issue 1, 37-76, 2010.

[20] Achterberg T., Berthold T., Koch T., Wolter K., Constraint Integer Programming. A New Approach to Integrate CP and MIP, Lecture Notes in Computer Science, Volume 5015, 6-20, 2008.

[21] Bockmayr A., Kasper T., Branch-and-Infer, A Framework for Combining CP and IP, Constraint and Integer Programming Operations Research/Computer Science Interfaces Series, Volume 27, 59-87, 2004.

[22] Eclipse - The Eclipse Foundation open source community website, www.eclipse.org, 2015

[23] LINDO Systems - Optimization Software: Integer Programming, Linear Programming, Nonlinear, www.lindo.com, 2015

[24] Sitek P., Wikarek J., Cost optimization of supply chain with multimodal transport, Federated Conference on Computer Science and Information Systems (FedCSIS), 1111-1118, 2012.

[25] M. Relich and W. Muszynski, “The use of intelligent systems for planning and scheduling of product development projects”, Procedia Computer Science, vol. 35, pp. 1586–1595, 2014.

[26] Grzybowska K., Selected Activity Coordination Mechanisms in Complex Systems, J. Bajo et al. (Eds.): PAAMS 2015 Workshops, CCIS 524, Springer International Publishing Switzerland, pp. 1–11, 2015.

[27] Grzybowska K., Kovács G., Logistics Process Modelling in Supply Chain – algorithm of coordination in the supply chain – contracting, International Joint Conference SOCO’14-CISIS’14-ICEUTE’14, Advances in Intelligent Systems and Computing, Vol. 299, pp. 311-320, 2014

[28] Bąk, S., Czarnecki R., Deniziak S. (2013). Synthesis of Real-Time Cloud Applications for Internet of Things. Turkish Journal of Electrical Engineering &Computer Sciences, DOI: 10.3906/elk-1302-178.

[29] Bocewicz, G., Nielsen, I., Banaszak, Z. (2014). Iterative multimodal processes scheduling. Annual Reviews in Control 38(1), 113-132.

DOI: 10.15439/2015F218
ACSIS, Vol. 5

# Graph based approach to the minimum hub problem in transportation network

J. Owsiński, J. Stańczak, A. Barski, K. Sęp
Systems Research Institute,
Polish Academy of Sciences
ul. Newelska 6,
01-447 Warszawa, Poland
Email: {owsinski, stanczak,
Aleksy.Barski, sep} @ibspan.waw.pl

P. Sapiecha
The Faculty of Electronics
and Information Technology,
Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
Email: {sapiecha@tele.pw.edu.pl}

***Abstract*—*In this paper we consider a hub location problem in a real multimodal public transportation network. This problem is also known as the park-and-ride problem. Hubs stations are special facilities that serve as switches in such a network. In practice the set of hubs has a strategic importance, because all of the traffic that passes through the network can be controlled by these elements. From the theoretical point of view, the minimal hub problem is NP hard. Two different approaches to this problem are presented. The first group of methods bases on the greedy algorithms. In the second group the evolutionary strategy is used. The computational results for these algorithms proved a significant efficiency, what can be clearly expressed in terms of an input data reduction and also in quality measure values for the obtained solutions of this problem.***

## I. Introduction

LET US consider a real public transportation network. This network can be described as a graph [3], [8], [9]. In this graph, each tram/bus stop corresponds to one vertex and any two vertices are adjacent iff they belong to at least one common public transport line. Hence, the set of vertices along a route forms a connected subgraph of the entire graph. Each vertex in this graph is characterized/labeled by a set of numbers of tram/bus lines passing through it.

In such a graph, a hub set is a subset of vertices, such that any two vertices are connected by a path whose vertices lie in it. Let us define the minimum hub set problem as the problem of finding the hub set of minimal cardinality for a given graph. This collection has a strategic importance for the transportation system, because all the traffic that passes through the network can be controlled by these vertices. Communication hubs are excellent candidates for park-and-ride (P&R, sometimes P+R) points ([4]) with good connections to the center and other parts of the city and this idea is a basis of presented work. This problem is well known and is commonly used in industry, in particular in such areas as transport, telecommunication, and distributed computing [1], [2], [11], [13], [16].

Let us observe that we can try to find the minimum hub set for a given network in two different ways. This problem would be directly reduced to the task of designation the minimum dominating set [6], [14], [15]. Alternatively, we have a possibility of creating a hypergraph, such that its set of vertices is the same as in the original network, and for each set of vertices, which correspond to stops lying along a tram/bus route, we form the hyperedge. In this case, we reduce the considered problem to the search for the minimal transversal in a constructed hypergraph.

## II. Basic mathematical definitions

This section provides some basic notation. A **graph** is a representation of a set of objects, where some pairs of objects are connected by links. The interconnected objects are represented by mathematical abstractions called vertices, and the links that connect some pairs of vertices are called edges. More formally, a **graph** is an ordered pair G=(V, E) comprising a set V of **vertices** or **nodes** together with a set E of **edges**, which are 2-element subsets of V (E is a subset of VxV). An **undirected** graph is the one in which edges have no orientation. The edge (a, b) is then identical to the edge (b, a).

Let N(v)={u ∈ V: (v, u)∈E} be an open **neighborhood** for a given vertex v.

A **dominating set** for a graph G is a subset D of V such that every vertex not in D is adjacent to at least one member of D. This problem is strongly related to a problem well known in computational geometry, the **art gallery problem.** The **domination number** $\gamma(G)$ is the number of vertices in a smallest dominating set for G. The k-**dominating set problem** concerns testing whether $\gamma(G) = k$ for a given graph G and natural number k; it is a classical NP-complete decision problem in computational complexity theory (Garey & Johnson 1979) [3], [8], [9].

**Theorem** (Ore): If G=(V, E) is a graph without isolated vertices, then the complement of a minimal dominating set of G is also a dominating set of G. This implies that every such graph has two disjoint dominating sets and hence, $\gamma(G) \leq ½$ Card(V) [3], [8], [9].

**Theorem** (Arnautov 1974, Payan 1975, Alon 1990) : If G=(V, E) is a graph with minimum degree d > 1, then $\gamma(G) \leq [(1 + \ln(d+1))/(d + 1)]$ Card(V) [3], [8], [9].

978-83-60810-66-8/$25.00©2015, IEEE

An **independent set** is a set of vertices in a graph, such that no two of them are adjacent. The size of such a set is called the **independence number** of G, and is denoted $\alpha(G)$. The problem of finding the set such that $\alpha(G) = k$ is called the k-**independent set problem** and is an NP-complete decision problem. The dominating sets are closely related to independent sets (e.g. the 8-**Queens Problem** Puzzle). Namely, an independent set is also a dominating set if and only if it is a maximal independent set, so any maximal independent set in a graph is necessarily also a minimal dominating set.

A **hypergraph** is a generalization of a graph in which an edge can connect any number of vertices. Formally, a hypergraph H is a pair H=(X,F), where X is a set of elements called **vertices**, and F is a set of non-empty subsets of X called **hyperedges**. Let F be a subset of P(X)\{Ø}, where P(X) is the power set of X and F(x) = {f ∈F: x∈f } for x ∈ X. A hypergraph is also called a **set system** or a family of sets drawn from the universal set X. The **rank** of hypergraph H is the size of the largest hyperedge in H. A **set covering** of a hypergraph H=(X, F) is a subfamily C of F, such that the union of hyperedges from C equals the universe of vertices. A **transversal** (or **hitting set**) of a hypergraph H=(X, F) is a subset T of X that has a nonempty intersection with every edge. The notions of hitting set and set covering are equivalent. The **decision versions** of **hitting set** and **set covering** problems are NP-complete.

For any given graph G=(V, E) with V={1, 2,..., n}, construct a hypergraph H=(X, F) as follows: the universe X is V, and the family of hyperedges F is {F1, F2, ..., Fn} such that Fv consists of the vertex v and all vertices adjacent to v in G. Hence, if D is a dominating set for G, then S={Sv: v∈D} is a feasible solution of the set cover problem, with Card(C)=Card(D). Conversely, if S={Sv: v∈D} is a feasible solution of the set cover problem, then D is a dominating set for G, with Card(D)=Card(C).

The **greedy algorithm** for set covering chooses the sets according to one rule: at each stage, choose the hyperedge that contains the largest number of uncovered elements. This algorithm actually achieves an **approximation ratio** Card(C)/Card(Opt) (Opt – is an optimal set covering) of **h(rank)**, where h(n) is the n-th harmonic number. This value is approximately given by: ***O*((*1*+ *log*(*Card*(*V*)))**. We can construct dual algorithm for hitting set problem for which performance ratio is: ***O*((*1*+ *log*(*Card*(*F*)))**.

**Algorithm 0 – Greedy set covering method**

```
input:   a hypergraph H=(X, F);
output:   a set covering C;
 U := X; C := Ø;
 while( C ≠ X )do{
    select S from F such that maximizes Card(S ∩ U);
    U := U \  S;
    C := C ∪ {S};
  }
return: C;
```

It is interesting that there exists a pair of **polynomial-time reduction** between the **minimum dominating set problem** and the **minimum set covering problem** (Kannan 1992 pp.108–109). These reductions show that an efficient algorithm for the minimum dominating set problem would provide an efficient algorithm for the set cover problem and vice versa. According to the above presented facts, the greedy algorithm provides a factor 1+ log(Card(V)) approximation of a minimum dominating set.

**Algorithm 1 – Greedy hitting set method**

```
input:    a hypergraph H=(X, F);
output:   a hitting set – transversal T;
T := Ø; E := F;
while( E ≠ Ø )do{
   select x from X such that maximizes Card(F(x) ∩ E);
   T := T ∪ {x} ;
   E := E \ F(x);
}
        return: T;
```

Consider another approach to the hitting set problem, I.e. **algorithm 2** (**MSBT**) [12]. It seeks the vertices with the lowest degree and removes them from the set of vertices. If (without a removed vertex) the vertex set is not a transversal, then this vertex should be added to the transversal under construction, and the edges incident with it are eliminated from the hypergraph – they are deemed to be covered.

Yet another method (**algorithm RSBT**) [12] seeks a maximum transversal in the sense of cardinality. Contrary to the MSBT algorithm, the RSBT removes vertices with the biggest degree as long as it can. All the rest is the same as in the MSBT algorithm.

**Algorithm 2 - MSBT**

```
input:     a hypergraph H=(X, F);
output:   a hitting set – transversal T;
T :=  Ø V := X; Q := X; E := F;
while( V ≠  Ø & E ≠  Ø) do
 {      x := vertices with the lowest degree; V := V \ {k};
    if( V is not transversal of hypergraph (Q, E) )
    then {  T := T ∪  {x}; E := E \ F(x); V := V \ { v □ V: F(v) = Ø};}
       else
          for( each edge covers by exact one vertex v )do
              {T := T ∪  {v}; E := E \ F(v) ; V := V \ {v};}
       Q:=V;
 }
```

The algorithms MSBT and RSBT are the algorithms designed by authors as a complement of the Algorithm 1.

Let us consider a **reduction** from the dominating set problem to the set covering problem. For any given graph G=(V, E) with V={1, 2,..., n}, construct a hypergraph H=(X, F) as follows: the universe X is V, and the family of hyperedges F is {F1, F2, ..., Fn} such that Fv consists of the vertex v and all vertices adjacent to v in G. Hence, if D is a dominating set for G, then S={Sv: v∈D} is a feasible solution of the set cover problem, with Card(C)=Card(D). Conversely, if S={Sv: v∈D} is a feasible solution of the set cover problem, then D is a dominating set for G, with Card(D)=Card(C).

According to the above presented facts, the **greedy algorithm** provides a factor 1+ log(Card(V)) approximation of a **minimum dominating set.** Additionally, Raz and Safra (1997) show that **no** algorithm can achieve an approximation factor better than clog((Card(V)) for some c > 0 unless P = NP. Fomin, Gradoni and Kratsch (2009) show an **exact** algorithm which can be used to find a minimum dominating set in time *O*((1.5264)^n) and polynomial space. In parallel, a faster algorithm, using *O*((1.5048)^*n*) time was found by van Rooij, Nederlof and van Dijk (2009).

## III. Evolutionary algorithm

Graph problems such as graph coloring, TSP, graph partitioning, maximum clique search, etc. (**NP-hard optimization problems**) are often solved using computational intelligence methods due to the lack of efficient polynomial algorithms. Among them, the **evolutionary algorithms** (EAs) are often applied; thus, it seems fully justified to use the EA in the present graph transformation problem. The EA approach is quite different than the described earlier hypergraph method. The EA method tries to find hubs, which are strongly connected (in the sense of high capacity of connections – see formula (4)) among them and allow for fast mass travel to the center of the city and other hubs. It is very likely that such hubs are good candidates for the park-and-ride locations.

In our approach, information about the transformed graph is stored in an array of data describing all connections among graph nodes – public communication stops. This array is an adjacency matrix of undirected graph with stored values – weights, denoting the capacities of connections.

The idea of our approach is based on the hub and spoke paradigm. The spokes, which in this case represent public transport stops of minor importance, constitute groups of nodes connected with their hubs. The size of the subgraph of hubs is known in advance. Hubs (should) represent important public transport stops with fast and frequent connections with other important stops and the center of the city.

Algorithm 4. shows how the typical EA work, but this general framework requires several improvements to work efficiently and thus a specialized evolutionary method, developed by authors is used to solve the problem.

To adjust the EA to the solved problem it is necessary to apply proper, problem-specific **encoding** of solutions (sometimes called also population members, individuals or even agents), to develop specialized **genetic operators** tailored for the analysed data structure and the solved problem and, finally, to formulate the problem-dependent **fitness function** to be optimized by the algorithm.

**Algorithm** 4 - **The standard evolutionary algorithm**

```
input:   a given problem;
output:  solution of problem;
{   while( not stop condition )do
        { reproduction and modification
          of solutions using genetic operators;
          valuation of obtained solutions;
          selection of individuals for the next generation;
        }
}
```

The first step to obtain efficient evolutionary method is to choose an appropriate encoding method. It is obvious that there are plenty of possible solutions with different advantages and drawbacks. In presented approach the solution is stored as a vector of selected hubs with lists (or variable length vectors) of attached to them spokes.

An important problem in developing an efficient EA is the design of **genetic operators** for the adopted data structure, taking into account the constraints imposed on solutions. In the case here considered, the standard crossover and mutation operators are not proper, so the problem-specific, specialized operators must be prepared to efficiently solve the problems considered. When one element of a modified solution (for instance one node) is to be moved to another place (e.g. to the cluster of another hub), it must first be checked if it has a connection with this new hub. If not, the operation is canceled to avoid creation of infeasible solutions. Altogether, here, the set of genetic operators is:

- mutation – exchange of randomly chosen nodes in different sets of spokes,
- relocation of a randomly chosen node to a different set of spokes,
- exchange of a randomly selected hub for randomly selected spoke.

Application of several specialized genetic operators requires a method of selecting and executing them in all iterations of the algorithm. In the approach used in [17] it is assumed that an operator that generated good results for some population member should have bigger probability of execution for its possible offspring and more frequently affect them than the remaining operators. But it is very likely that the operator that improves one individual and its descendants may give worse effects for other individuals, because of its location in the domain of possible solutions. Thus, every individual should have its own preferences and could be treated as an agent, whose role is to select and call one of the evolutionary operators (or to perform an action) to obtain improvement of solution (or its income) as high as possible. In the EA here used, every individual has an additional vector of floating point numbers, besides the encoded solution. Each number corresponds to one genetic operator and is a measure of its quality (a quality factor). The higher the factor, the higher the probability of calling and executing the operator. Simple normalization of the vector of quality coefficients turns it into a vector of operator execution probabilities. This set of probabilities is also an expression of experience of every individual and according to the experience gathered one can maximize the chances of its offspring to survive.

The method of computing quality factors is based on **reinforcement learning**. When the selected by the individual-agent $i^{th}$ operator is applied, this can be regarded as an agent performing action $a_i$ leading to a new state $s_i$, which, in this case, is a new solution. The **agent** (or individual, solution or population member) receives reward or penalty depending on the quality of the new state (solution). The aim of the agent is to perform the actions, which give the highest long term discounted cumulative reward $V^*$, maximizing its chances to have offspring in next generations:

$$V^{\Pi}=E_{\Pi}\left(\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\right) \quad ,(1)$$

$$V^*=\max_{\Pi}\left(V^{\Pi}\right) \quad .(2)$$

The following formula, derived from (1) and (2), is used for evaluation purposes:

$$V(s_{t+1})=V(s_t)+\alpha\left(r_{t+1}+\gamma V^*(s_{t+1})-V(s_t)\right) \quad ,(3)$$

where:

*Π* - the set of possible strategies of the agent,
*V*⏭ - the discounted cumulative reward obtained using strategy *Π*,
*E* - expected value,
*k* - consecutive time steps,
*t* - current time,
$V(s_t)$ - quality factor or discounted cumulative reward,
$V^*(s_{t+1})$ - estimated value of the best quality factor
(in our experiments we take the value attained
by the best operator),
$\alpha$ - the learning factor,
$\gamma$ - the discount factor,
$r_{t+1}$ - reward for the best action, equal to the improvement
of the quality of solution after execution
of the evolutionary operator.

In the here presented experiments, the values of $\alpha$ and $\gamma$ were set to 0.1 and 0.2, respectively.

The **fitness function** in the EA is closely connected with problem specific quality function. The fitness function evaluates the members of the population. It is a modified (scaled, translated, etc.) problem quality function, prepared for computational purposes in the EA. The quality function directs evolutionary computations to obtaining the proper graph structure. In the considered problem the quality function is formulated to direct the EA towards the desired structure of potential P&R, taking into account weights among the graph vertices. Several quality functions can be used, depending on input data (binary, integer or real) or what kind of set of P&R nodes one wants to obtain. The formula presented in this paper was obtained on the basis of experiments. Usually, such formulas contain a penalty part for the potential invalid or improper structure of the obtained solutions.

For the **hub and spoke** structure with the predetermined number of kernel nodes the quality function promotes solutions where a rather small subgraph of hubs is (almost) fully connected and the sets of spokes attached to their hubs have medium sizes:

$$\max Q=\sum_{i=1}^{n}\left(a*\sum_{j=1}^{k_i}w_{ij}+b*w_{iC}+c*\sum_{m=1}^{k-k_i}w_{im}\right) \quad , \qquad (4)$$

where:

$w_{ij}$ – weight between candidate for P&R (hub) and its subgraph
of communication stops (spokes),
$w_{iC}$ – weight between candidate for P&R (hub) and
the center of the city[1],
$w_{im}$ – weight between candidate for P&R (hub)
and remaining communication stops,
*n* – predetermined (constant) number of hubs (candidates for P&R)
in the solution,
$k_i$ – number of nodes attached to $i^{th}$ hub,
*k* – number of nodes in the whole graph,
*a, b, c* – non-negative constants values that emphasize the influence
of corresponding factor.

## IV. Data base & computational environment

For the computational experiments a set of data files was obtained from the **www page**: http://www.ztm.waw.pl/. The description of the structure of public transport in Warsaw and full list of bus, subway and tramway stops and all lines of these means of transport are collected in the data files. The data include more than 5600 stops, but the majority of them are stops located not far from each other, with the same names but different minor numbers (like Centrum 01, Centrum 02,...), which indicate the opposite direction or different transport mean or extensions for bigger numbers of vehicles boarding passengers simultaneously. Thus, first the stops were aggregated: all of them with the same names and different minor numbers were treated as one. This allowed to reduce the size of the communication network to 1883 stops.

We prepared computation experiments for this article based on typical **personal computer** (with an Intel i5 4 x3.2 GHz microprocessor) running Linux OS. All our programs were written in the C language and compiled with the g++ compiler application.

## V. Data reduction

At the second step, some preliminary computations for the considered data allowed to reduce the size of the problem even more. This **preprocessing** was oriented to **reduce** input data (filtering method). We can treat the graph G as a representation of Warsaw **transportation network** being an input data for our programs.

Looking at this graph one can observe that it is possible to remove some of the redundant edges. It is clear that if two vertices, a and b, are **adjacent** in G and the sets of bus/subway/tram lines passing through them are the same, then we can apply the edge (a, b) contraction operation. An **edge contraction** is an operation which **removes** an edge from a graph while simultaneously **merging** the two vertices that it previously joined. More precisely, the edge (a, b) is removed and its two incident vertices, a and b, are merged into a new vertex c, where the edges incident to c correspond to edges incident to either a or b. Obviously, it is possible to apply these contraction operations several times, in polynomial time. After executing this procedure an obtained graph consists of 1003 vertices (reduction of 880 vertices).

## VI. The results of presented approaches

Given the above, we can try to find the **minimum hub set** for the given network by applying the presented algorithms.

[1] We decided to emphasize a central communication stop in the city as a virtual aim of the majority of commuters.

At the beginning we construct a hypergraph, such that its set of vertices is the same as it is in the input network. Let for each set of vertices, which correspond to stops lying along a tram/bus route, form a hyperedge. In this case, we **reduce** the problem considered to the **minimal transversal** construction for this hypergraph. Therefore, two different approaches are possible - deterministic and random:

(**a**) application of **greedy methods**
(in different versions);

(**b**) use of **evolutionary algorithms.**

In order to compare these methods we also made computations with and without the preliminary reduction. In the first approach **(greedy method),** algorithm 2 found 35 vertices. The **RSBT** found smaller transversal – 140 vertices instead of 152 vertices found without the reduction. Considering the obtained results it looks like the reduction could be used for the EA approach. There was no change as to the result obtained in the evolutionary algorithm, but there was a significant reduction of time needed for computation.

Due to a smaller graph of public communication stops considered each iteration of EA last significantly shorter, while the number of them remained unchanged. For the greedy method the time of computation is very short (less than 1 s.) with and without the reduction, thus it is no use to compare them with EA and with or without the reduction.

TABLE 1.
COMPUTATIONAL RESULTS

| Number of hubs | Time of 10 000 EA iterations (average of 10 trials) | |
|---|---|---|
| | Before reduction | After reduction |
| 50 | 5 854 s | 2 478 s |
| 100 | 13 433 s | 5 895 s |
| 152 | 27 231 s | 14 256 s |

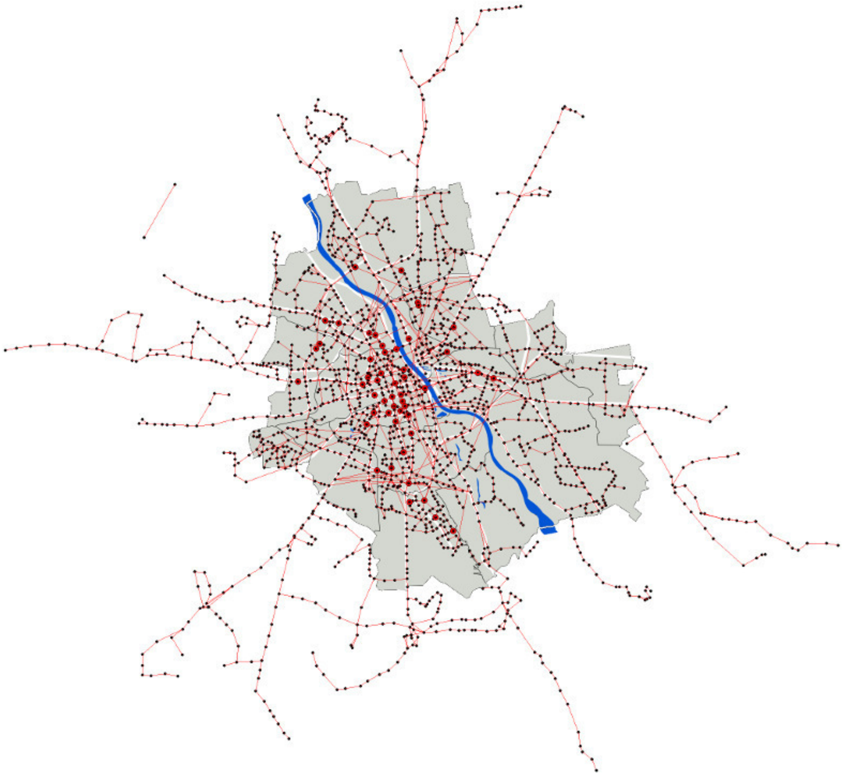


**Figure 1**: Result obtained by the EA with the imposed number of 50 hubs

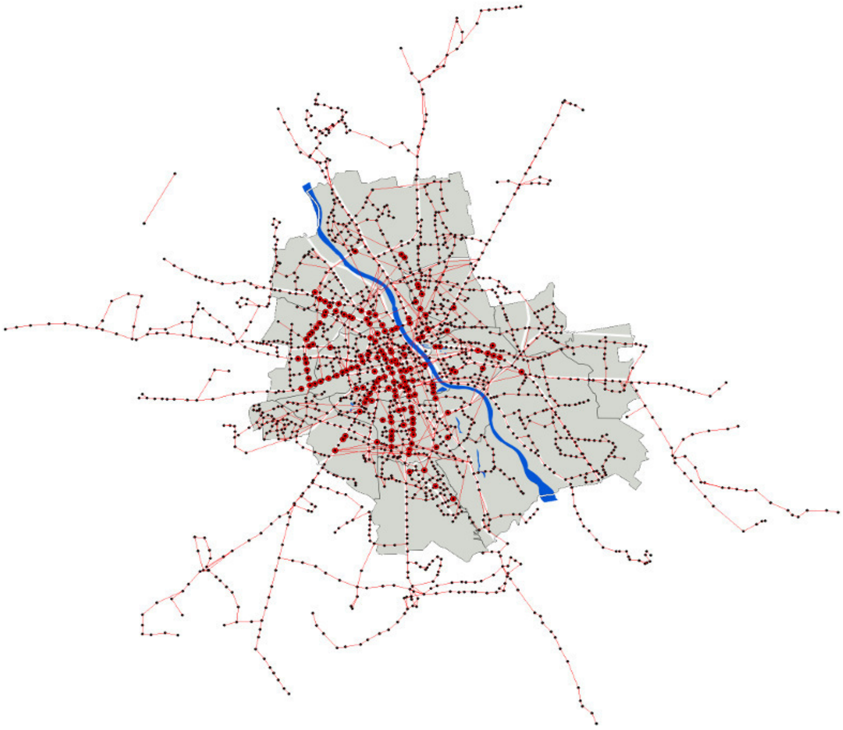


**Figure 2**: Result obtained by the EA with the imposed number of 152 hubs

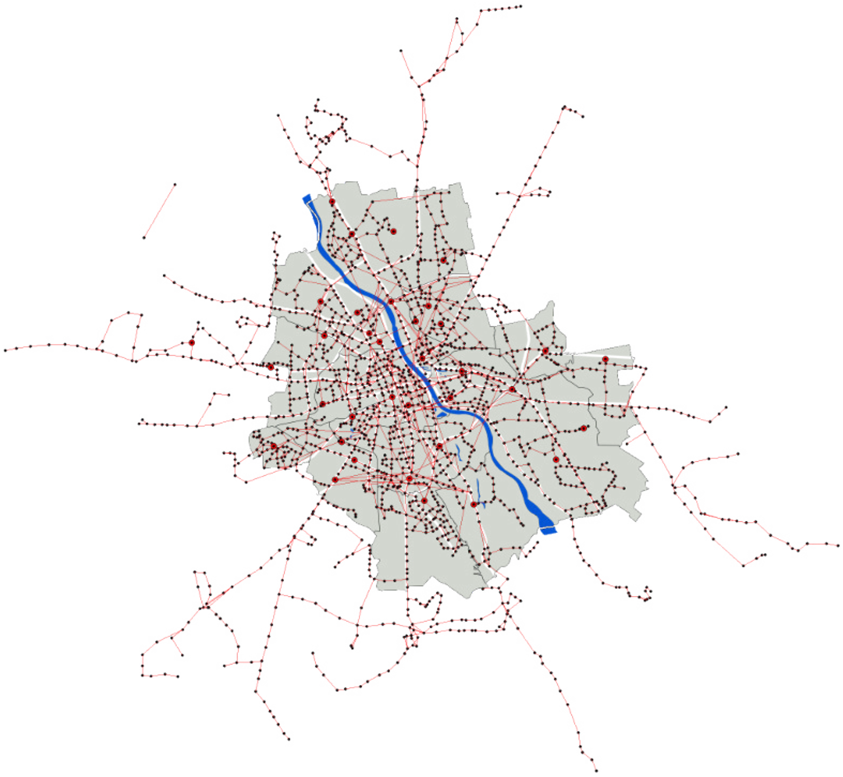


**Figure 3**: Result obtained by the MSBT algorithm - 35 points selected

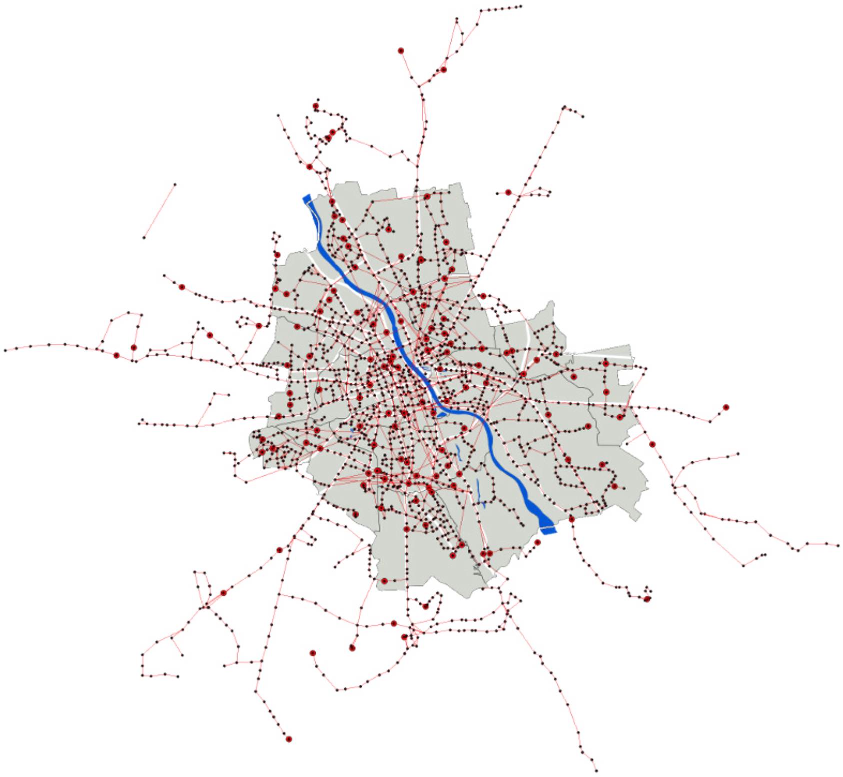


**Figure 4**: Result obtained by the RSBT algorithm - 140 points selected

## VII. Conclusions

This article is devoted to the problem of locating hubs in a public transportation network. Two different approaches to this problem are presented. Namely, the first group of considered methods are based on greedy strategies, and the second group, using evolutionary strategy. In the computational experiments the model of the transportation network of Warsaw was analysed. The obtained results showed for this model the potential P&R set of 34 stations using the MSBT greedy method and of 140 potential points using the RSBT greedy method. Using the EA with imposed number of hubs, satisfactory solutions for 50, 100 and 152 predetermined numbers of candidates were obtained. Additionally, it must be concluded that the graph reduction could be used especially for the evolutionary approach as a filtration method significantly reducing the size of the solved problem. Almost always the preliminary reduction of the stop number speed up the computational process by about 46%. According to the quality of solutions obtained with and without the preliminary reduction it appears that the use of the proposed preprocessing method seems to be justified.

This work is the first step towards obtaining a more complete model of communication in Warsaw, which would be a base to find and project communication hubs with proper places for P&R facilities.

## References

[1] Aros-Vera F., Marianov V., Mitchell J., p-Hub approach for the optimal park-and-ride facility location problem. European Journal of Operational Research 226, 2013,

[2] Bonato A., Lozier M., Mitsche D., Perez-Gimenez X., Prałat P., The domination number of on-line social networks and random geometric graphs, Theory and Applications of Models of Computation: 12th Annual Conference, Singapore, 2015

[3] Ding-Zhu Du, Peng-Jun Wan, Connected Dominating Set: Theory and Applications, Springer Optimization and Its Applications, 2012,

[4] Farhan B., Murray A. Siting Park-and-Ride facilities using a multi-objective spatial optimization model. Computers & Operations Research 35, 2008,

[5] Gartner B., Matousek J., Approximation Algorithms and Semidefinite Programming, Springer, 2012,

[6] Gogas P., Papadimitriou T., Matthaiou M., A Novel Banking Supervision Method using the Minimum Dominating Set, The Rimini Centre for Economic Analysis, 2014,

[7] Greetham D., Poghosyan A., Charlton N., Weighted -rate dominating sets in social networks, University of Reading, UK, 2014,

[8] Haynes T., Hedetniemi S., Slater P., Fundamentals of Domination in Graphs, Marcel Dekker Inc., 1998,

[9] Henning M., Yeo A., Total Domination in Graph, Springer Monographs in Mathematics, 2013,

[10] Jin-Hua Zhao, Habibulla Y., Hai-Jun Zhou, Statistical Mechanics of the Minimum Dominating Set Problem, Journal of Statistical Physics, 2015,

[11] Krishnam R., Indukuri R., Penumathsa S.V., Dominating Sets and Spanning Tree based Clustering Algorithms for Mobile Ad hoc Networks, International Journal of Advanced Computer Science and Applications, Vol. 2, No.2, February 2011,

[12] Mażbic-Kulma B., Sęp K., Some approximation algorithms for minimum vertex cover in a hypergraph. Computer Recognition Systems 2. Springer-Verlag, Berlin-Heidelberg, pp. 250-257. Series: Advances in Soft Computing, 2007,

[13] Milenkovic T., Memisevic V., Bonato A., Przulj N., Dominating Biological Networks, PLoS ONE, 2011,

[14] Molnár F. , Sreenivasan S., Szymanski B. K. & Korniss G. Minimum Dominating Sets in Scale-Free Network Ensembles, Nature, 2015,

[15] Naixue Xiong, Xingbo Huang, Hongju Cheng, and Zheng Wan, Energy-Efficient Algorithm for Broadcasting in Ad Hoc Wireless Sensor Networks, Sensors 2013,

[16] Nettleton D. F., Data mining of social networks represented as graphs, Computer Science Review, Elsevier, 2013,

[17] Stańczak J. (2003) Biologically inspired methods for control of evolutionary algorithms. Control and Cybernetics, 32(2), pp. 411-433.

# 21$^{th}$ Conference on Knowledge Acquisition and Management

KNOWLEDGE management is a large multidisciplinary field having its roots in Management and Artificial Intelligence. Activity of an extended organization should be supported by an organized and optimized flow of knowledge to effectively help all participants in their work.

We have the pleasure to invite you to contribute to and to participate in the conference "Knowledge Acquisition and Management". The predecessor of the KAM conference has been organized for the first time in 1992, as a venue for scientists and practitioners to address different aspects of usage of advanced information technologies in management, with focus on intelligent techniques and knowledge management. In 2003 the conference changed somewhat its focus and was organized for the first under its current name. Furthermore, the KAM conference became an international event, with participants from around the world. In 2012 we've joined to Federated Conference on Computer Science and Systems becoming one of the oldest event.

The aim of this event is to create possibility of presenting and discussing approaches, techniques and tools in the knowledge acquisition and other knowledge management areas with focus on contribution of artificial intelligence for improvement of human-machine intelligence and face the challenges of this century. We expect that the conference&workshop will enable exchange of information and experiences, and delve into current trends of methodological, technological and implementation aspects of knowledge management processes.

## Topics

The following group topics, concerning both theory and applications, will be included (unavoidably incomplete):

- Knowledge discovery from databases and data warehouses
- Methods and tools for knowledge acquisition
- New emerging technologies for management
- Organizing the knowledge centers and knowledge distribution
- Knowledge creation and validation
- Knowledge dynamics and machine learning
- Distance learning and knowledge sharing
- Knowledge representation models
- Management of enterprise knowledge versus personal knowledge
- Knowledge managers and workers
- Knowledge coaching and diffusion
- Knowledge engineering and software engineering
- Managerial knowledge evolution with focus on managing of best practice and cooperative activities
- Knowledge grid and social networks
- Knowledge management for design, innovation and eco-innovation process
- Business Intelligence environment for supporting knowledge management
- Knowledge management in virtual advisors and training
- Management of the innovation and eco-innovation process
- Human-machine interfaces and knowledge visualization

## Event Chairs

**Hauke, Krzysztof,** Wroclaw University of Economics, Poland

**Nycz, Małgorzata,** Wrocław University of Economics, Poland

**Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland

**Pondel, Maciej,** Wroclaw University of Economics, Poland

## Program Committee

**Andres, Frederic,** National Institute of Informatics, Tokyo, Japan

**Chmielarz, Witold,** Warsaw University, Poland

**Christozov, Dimitar,** American University in Bulgaria, Bulgaria

**Goluchowski, Jerzy,** University of Economics in Katowice, Poland

**Helfert, Markus,** Dublin City University, Ireland

**Jan, Vantienen**

**Jelonek, Dorota,** Faculty og Management of Czestochowa University of Technology

**Korczak, Jerzy,** Wrocław University of Economics, Poland

**Kowalczyk, Ryszard,** Swinburne University of Technology, Melbourne, Victoria, Australia

**Ligęza, Antoni,** AGH University of Science and Technology, Poland

**Mach-Król, Maria,** University of Economics in Katowice, Poland

**Mercier-Laurent, Eunika,** IAE Lyon3, France

**Nalepa, Grzegorz J.,** AGH University of Science and Technology, Poland

**Sobińska, Małgorzata,** Wroclaw University of Economics

**Surma, Jerzy,** Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States

**Zhelezko, Boris,** Belorussian State Economic University, Belarus

## Organizing Committee

**Marciniak, Katarzyna,** Wroclaw University of Economics, Poland

DOI: 10.15439/2015F010
ACSIS, Vol. 5

# What technology for efficient support of sustainable development?

Eunika Mercier-Laurent
University Jean Moulin, Lyon 3
Lyon, France
eunika.mercier-larent@univ-lyon3.fr

*Abstract*—**Sustainable development is among the greatest challenges of this century. Sustainability and development are apparently opposite. The traditional approaches try to manage the planet protection without taking the best from technology while eco-innovation focus mainly on smart transportation, smart use of energy and water and waste recycling. Many new concepts such as Smart, Intelligent, Green or Wise City were invented to promote … old technology. All these initiatives use data bases and machine to machine communication. The AI approaches and techniques may be very useful not only for process optimization but also for simulation before doing and education of new behaviors. This talk will present the principal advantages of using AI for balancing planet protection and technological progress.**

*Index Terms*—**Sustainable develoment, Intelligent technology, e-co-innovation, knowledge.**

## I. Introduction

TECHNOLOGICAL progress is considered as a powerful engine of economy but, despite partial recycling, it generates a lot of waste and contributes to climatic changes. Computers and smartphones are quickly out of date; new software versions are available only on new devices. Most of these devices are not eco-designed.

The quickly growing population of the cities represents a lot to feed it is also a huge global market facilitated by communication technology, e-business, and transportation means. Google business model based on advertisements generates intellectual and visual pollution and is an important theft of time – time to find how to close the advertisement window or 30 seconds if this facility is not provided. The mass media promotes the “to have more and to show” mentality, through advertisements and various entertainment programs. “Buy more, throw away and buy new” are the engines of today business. Opening to business and quick development of China and other Asian and South American countries offering the low labor cost increases the relocation-out from origin countries, mainly US and Europe, in search of quick business – in aim to offer more for less and increase the firms’ income. This way of doing increases also pollution in developing countries and the unemployment in Europe and other developed countries. Besides, we have to recycle products, often of poor quality, made somewhere else and travelling around the world by ships and airplanes.

Innovation is among contributors to planet disaster, because the inventors and designers think about functionality, shape, look and attractiveness and not about the overall behavior inside the environment. Business people want to sell more; nowadays a global market is addressed without thinking about the right values and right benefits.

The eco-innovation movement claims to generate new businesses and jobs but the use of technology in the related activities is moderate. Corporate Social Responsibility aims to balance environmental, economic and social impacts with a little use of traditional technology. Based mainly on norms, it aims in having an impact on products design via Product Lifecycle Management (PLM) software. Intelligent technology has a significant role to play in balancing all these components. A condition however is a capacity of different thinking and managing all available knowledge and experience that apply.

After describing main trends some perspective will be given on another way of doing and the role of intelligent technology in the transformation of society.

## II. From sustainability to sustainable development

In biological systems sustainability means long life. Systems components influence each other and the balance of the whole system is the condition to survive. Human activities have affected the balance of natural ecosystems [4, 5, 6].

Sustainable development is a principle that many people claim to have invented. The definition of the European Union is following:

“Sustainable Development stands for meeting the needs of present generations without jeopardizing the ability of futures generations to meet their own needs – in other words, a better quality of life for everyone, now and for generations to come. It offers a vision of progress that integrates immediate and longer-term objectives, local and global action, and regards social, economic and environmental issues as inseparable and interdependent components of human progress.”

“Sustainable development will not be brought about by policies only: it must be taken up by society at large as a principle guiding the many choices each citizen makes every day, as well as the big political and economic decisions that have to be taken. This requires profound changes in thinking, in economic and social structures and in consumption and production patterns”[2].

Sustainable development can be seen as a process for meeting human development goals while maintaining the ability of natural systems to provide the required natural resources and ecosystem services upon which the current

978-83-60810-66-8/$25.00©2015, IEEE

economy and society depend. The greedy economic system is not compatible with sustainable development. In this context sustainable development is oxymora – it is impossible to be sustainable without deep understanding of our natural ecosystems (multidisciplinary knowledge) and without radical change of behaviours, objectives, values, political and economic system. Sustainable development is an attitude to learn and cultivate.

Various actions aiming in changing the way of acting and doing things are engaged, mainly by individuals and organizations. Many of them are supported by social networks and have potential to influence behaviours. For example, Uber platform (and some others) connects people offering services with those who need them. It is quick and effective, but out of the current economic practices (taxes, companies…) and raise conflicts between “old” and “new”. Serious games try to push people to act [7].

Introducing the environmental principles into a design of products and services is a step forward. However the traditional PLM tools should evolve to take into consideration a new way of doing [8] by using simulators and optimizers. Technology is able to provide a considerable help, but the way of thinking should evolve to global, holistic and system [9]. For example, aeronautic and automotive industries focus on lightening weight and reducing carbon footprint, while other aspects can also be considered.

Corporate Social Responsibility requires the integration of environmental aspects into design. While it constraints companies to take care about social and environmental impact, they should also focus on economic performance. Environmental norms (ISO 26000) are very heavy; it is impossible for a small company to check and respect all these norms without loosing business. The intelligent technology can help checking and optimizing things; such systems are prototyped [12].

To remain sustainable, development requires the intensive use of available knowledge – individual and collective, from related domains, from the past and currently gained from experience.

## III. From eco-innovation to e-co-innovation

According to European Union [3] there are many fields and activities concerned, such as air, chemicals, green public procurement, industry, marine and coast, biodiversity, noise, soil, urban development, waste and water. The areas that should be greened are agriculture, employment, energy, fisheries, research, trade, transport…All are potential source of growth and job creation. But in existing system and with current goals, attitudes and practices, it can not work as expected. The greedy economic system aims in producing and selling more. The industrial giants impose their products and absorb competitors.

Circular economy is promoted by some wise economists and recently by European Union: “This means re-using, repairing, refurbishing and recycling existing materials and products. What used to be regarded as ‘waste’ can be turned into a resource. The aim is to look beyond waste and to close the loop of the circular economy. All resources need to be managed more efficiently throughout their life cycle” [16]. How they face the current practices of planned and perceived obsolescence [17]?

Circular economy is just new business concept, but nothing new – it has been practice in Poland in the 1950s and it is still practiced in poor countries. Despite generating activities, recycling uses the energy and generates pollution in many cases. Repairing involves education of new know how and another design, that should be encouraged – those of modular and reparable and reusable products (back to the 1950). AI approaches and techniques are very well suites to do that.

Another approach, connecting the best from technology and the best human capacity is those of e-co-innovation, considering all components, see Figure 1.

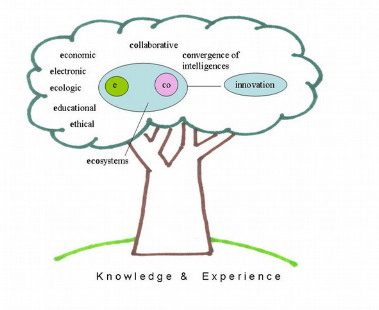


Fig. 1. e-co-innovation principles [9]

E-co-innovation is based on past and present knowledge as well as on a vision for the future. Its facets are ecological, economic, educational, electronic and ethical. It is collaborative and enables a convergence of intelligences. Inspired by nature, respecting natural and digital ecosystems it is centered on humans; such dynamics will certainly contribute to avoiding unbalance and a loss of competitiveness leading to economic, ecological and social decline.

## IV. Contribution of AI – example of wise city

The recent trends of Smart City, Intelligent City, Green and recently Wise City aim in providing services to citizens of becoming bigger and bigger cities. Actually, the population of cities is growing for different reasons, among them job offer, access to services, culture, immigration…This kind of initiatives is based on linear approximation that in 2050 70% of world population will be living in the Megapoles. The other factor is the ambition of politics to grow their territory. As consequence regional economy collapses in many cases and should be reinvented using local talents and resources.

The concept of Smart City was probably born in early 2000s. In aim to reboost their existing technology IBM changed in 2010 a slogan to Smarter Planet. Cisco as network provider is also involved. Intelligent City tries to be different and defines integrated model of intelligent economy, transportation, environment, governance and services for citizens [13]. The city of Nice (French Riviera) runs annually a conference on Innovative City [14]. Growing cities want to offer well being to citizens by initiating Green City which main interest is in architecture, gardening and social cohesion [15].

Among the lastest trends the Wise City of Hong Kong brings together large companies: Schneider Electric, Alstom, Veolia, Thales and KPMG, no room for small ones. KPMG claim to be integrator of all offers. They focus on improvement of living conditions: pollution, access to clean water, waste treatment…Hong Kong with over 7 million of habitants produces 15,000 tones of waste daily and uses 60% more water than New York. Buildings are heavy consumers and air pollution is 53% local.

The Figure 2 gives some details on main project components which are: smart energy, smart mobility, smart water, smart public services, smart buildings, smart data centers and smart integration. Public services offer seems very poor. We can notice that the offers for energy and water are a part of smart buildings. Data Center is certainly related to city management and we hope there is only one for all applications proposed by partners.

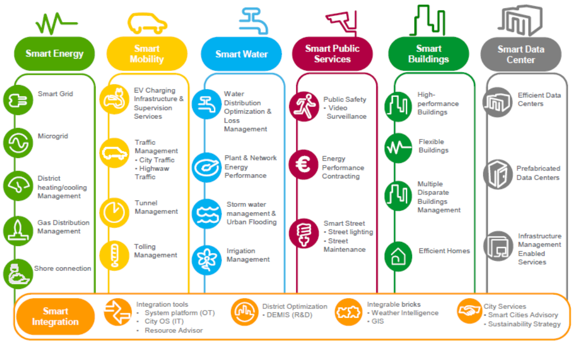


Fig. 2. Principal components of Wise City (source Schneider Electric, 2012)

Waste processing is not shown in this figure, because it is managed by another company. Such a separation generates more business for each company. They do not try to build an optimal (and cheaper) solution for the city in function of their needs; they just adapt their respective offers. The analysis of current situation and real needs is not published.

### A. *What could be done*

The Digital Agenda for Europe [15] points out the links between domains: "Information and Communication Technologies (ICTs) enable us to see the connections between seemingly disparate issues, like transport and energy or health and economic growth, and help us find comprehensive solutions, for example in the European Innovation Partnerships on Smart Cities and Communities and on Active and Healthy ageing" [15]. Such thinking is vital for building integrated and reusable solutions.

"Sensors and embedded software make objects increasingly interactive. Computer simulation makes a direct contribution to sustainability issues, particularly with regard to environmental protection, the scarcity of raw materials, and the emergence of a low-carbon economy. Smart city ambition is offering their inhabitants increased comfort, employment and economic development."

Since over 50 years now Artificial Intelligence developed methods and techniques that are now embedded in many decision support systems, diagnostic and simulation tools, educational software, innovative electronic commerce, data, text and image mining tools, creativity "amplifiers", robots and drones.

Combination of Knowledge Management method and Artificial Intelligence techniques are available to guide and enhance the architecture of an adapted, innovative and reusable system for city management. The comprehension and common discovery of needs may help innovating the both city and involved companies. Implication of local people (citizens and companies), sharing their talents, knowledge and experience may be beneficial for all. Knowledge base of experience of smart cities around the word will avoid the waste on energy and money for developing again and again the similar and not connected data bases.

New attitudes can be influenced via peaceful and knowledgeable games such as intelligent purchase for less waste, reusing the packaging, intelligent use of water and energy, avoiding or combining the transportations and other are a part of fun learning for all. Behind there is strategy, problem solving, thinking and multisystem agents.

While robots were conceived to copy human intelligence there are some useful robots designed to help human as industrial or surgery robots, those able to diagnostic and fix complex equipments in places difficult or dangerous to access, such as control and monitoring cabinets, nuclear power plants or high voltage stations.

Despite of available technology for distance meetings the number of people travelling is still growing. Optimization techniques such as constraint programming are helpful for optimizing travel. However air transportation companies have their own hub and an intelligent optimizer of travels is still missing. The available "optimizers" such as kayak, sky scanner and others works for their clients and not for travelers.

Smart cities open their data and wish to make business on it. Data is as raw materials, but intelligent access to relevant information allows the large-scale dissemination, analysis and use of data for the benefit of consumers and citizens. Analytics are mainly used to find information in large amount of data. The AI techniques of knowledge discovery, such as neural networks, genetic algorithms, induction or other multi-strategy machine learning hybrid tools are available but underused.

Offering the targeted services, corresponding to the real user's needs could be a great value. The effective matching of offer and demand in many domains saves our time and help in capturing opportunities. Smartphones and future devices embedded with machine learning techniques shall learn from real-time interaction with the user and not from navigation (too many errors) or published profile only.

## V. Conclusion and Future Work

Only few examples of what can be done on city level were presented here. The main rule is to apply suitable knowledge management approach, involve the end users from the beginning and use the best adapted techniques. The principle of modularity, reusability and genericity is still valuable for defining and building an effective and replicable system that really support city management and offer expected services to citizens, organizations and companies being a part of the city (holistic approach).

The evaluation of the impact of city activities may lead to right metrics for measuring progress and leadership. However the balance between the use of technology and human capacity should be preserved. Technology producers have tendency to produce software and devices that think instead of the human and take decisions for him/her. This kind of applications may replace human at the long run and reduce his cognitive capacity. That's why we have to produce the applications that enhance human intelligence without switching it off.

The future work is about evaluating 6D impact (technological, environmental, economic, social, cultural and cognitive) of technology.

## REFERENCES

[1] G. H. Brundtland "Our Common Future", United Nations, March 1987

[2] EU - sustainable development http://ec.europa.eu/environment

[3] EU eco-innovation http://ec.europa.eu/environment/ecoap/about-eco-innovation/index_en.htm

[4] C. Folke, C.S. Holling, C. Perrings "Biological diversity, Ecosystems and the Human Scale", Ecological applications 6(4), pp 1018-1024, Ecplogical society of America, 1996

[5] A. Lenkowa "Oskalpowana ziemia", Omega, Wiedza Powszechna, Warszawa, 1969

[6] E. Mercier-Laurent "The Innovation Biosphere. Planet and Brains in Digital Era, Wiley, ISBN.

[7] Water footprint, http://www.empreinteh2o.com/

[8] E. Mercier "Innovation Ecosystems: Example of Eco-design Process, KAM & AI4KM, Fedscis 2013, Krakow

[9] E. Mercier-Laurent "Innovation Ecosystems", Wiley, ISBN.

[10] United Nations https://sustainabledevelopment.un.org

[11] http://wisecity.hk/

[12] Z. Feng, M.Rio, R. Allais, P.Zwolinski, T. Reyes, L. Roucoules, E. Mercier-Laurent, N. Buclet "Toward a Systemic Navigation Framework to Integrate Sustainable Development into the Company", Journal of Cleaner Production, Volume 54, pp 199–214,1 September 2013.

[13] Ville Intelligente http://www.smartgrids-cre.fr

[14] Innovative City http://www.innovative-city.com

[15] "ICT for Societal Challenges. Digital Agenda for Europe", Publication Office of the European Union, Luxemburg, 2013

[16] Circular Economy http://ec.europa.eu/environment/circular-economy/

[17] Planned Obsolescence and Perceived Obsolescence - https://www.youtube.com/watch?v=N2KLyYKJGk0

DOI: 10.15439/2015F175
ACSIS, Vol. 5

# Management of requirements for the projects of the diagnostic systems

Marcin Amarowicz
Silesian University of Technology
Institute of Fundamentals of Machinery Design
Konarskiego 18A, 44-100 Gliwice, Poland
Email: marcin.amarowicz@polsl.pl

*Abstract*—**The paper deals with the problem of designing diagnostic systems for technical objects such as power plant, machining etc. The new approach for this problem that is based on the methodology that are used in the design of machines and some aspects of requirement engineering that derived from the software area is considered in the article. In general case the requirements may be used to formalized notation of the project need, that described the designed diagnostic systems. The evaluation of gathered set of requirements may be done by using the specialized expert systems, that may be developed as part of proposed approach. The whole process related to gathering, writing and evaluation of requirements for the purpose of designing diagnostic system is presented in the article.**

## I. Introduction

DURING the operation of technical objects, exists the danger of occurrence various kinds of events, usually summary called as faults, that may have the negative consequences to the considered technical object and its immediate vicinity. These faults may be caused by different factors including the constructional errors committed during the design process, production errors associated with inaccuracy of technological processes, operational errors related to noncompliance with the prescribed terms of operation of the object and obsolescence errors that are the results of natural changes of object's state e.g. decreasing of strength etc., [15]. Probability of occurrence of particular faults and their consequences are interpreted as technical risk associated with the operated object. From the mathematical point of view, the technical risk may be written as:

$$R_i = P_i \cdot c_i, \tag{1}$$

where $P_i$ is the probability of occurrence of $i$-th fault and $c_i$ is the possible costs/consequences connected with the occurrence of this fault. In order to ensure the proper operation of a technical object it is necessary to minimize the technical risk. This can be achieved by diagnose the current technical state of the considered object. This process can be perform by using the diagnostic system that is designed specially for this object.

In technical diagnostics area it is assumed, that the technical state of an object is a function of faults [10]. Therefore, in order to correctly determine the state of the object it is necessary to detect particular faults. Process of their detection is performed based on available information about the existing relationships between the value of selected signals, recorded on considered technical object, and the possible faults. The occurrence of specific values of signal, values of several signals or e.g. values of selected features of these signals (e.g. medium, standard deviation etc.) is called a symptom of this fault. In the simplest case, the information about the exceeded of permissible value of some signal (e.g. exceeding the permissible rms value of vibration signal in the bearing node) could be the symptom of the fault [13].

In the structure of typical diagnostic systems, the two parts, the measurement and the software one can be distinguished [3]. The main purpose of the measurement part is to record the value of selected signals and save them, according to the assumed scheme, in the database e.g. in OPC servers. The software part of diagnostic system is responsible for the appropriate detection of faults. It consists of analysis of recorded signals, including the designation of the necessary features of these signals, and the inference about the technical state of the object. In many cases, this process is carried out by using advanced computational algorithms including the artificial intelligence methods. The results of performed analysis are presented (depending on the class of the diagnostic system) through various techniques e.g. sound and light alarms, synoptic screens, sms and email notifications etc.

The purpose of this article is to present the new approach to the designing of the diagnostic systems. This approach will be based on general methodology that are used in the design of machines and the introduction to it a some modifications. They are necessary because the specificity of the field which is the technical diagnostics must be taken into account [1], [2].

## II. Diagnostic systems design method

In general case the process of design machines is a multistage. It begins from the definition of the need. At the general level of detail it expresses the expectations that should be met by the developed technical object. It is important to note, that the need cannot impose any restrictions and itself cannot define any potential solutions. In the next stage, with using various techniques such as brainstorming, teamwork, morphological table etc., the set of possible solutions is generated. After appropriate analysis and evaluation, the optimal solution is selected from this set. This one best meets the evaluation criteria and housed in imposed or adopted project limitations [7].

978-83-60810-66-8/$25.00©2015, IEEE

Taking into account that the need to develop a diagnostic system is the main project's need, it may be seen that the direct use of this approach to the design of diagnostic systems, encounters to the some problems at the stage of development possible solutions of this system. This is due to the necessity to take into account the wide knowledge about technical object, the ways of its diagnosis (methods of measurement, signal analysis etc.), accepted limitations etc., [5]. For this reason, it is not possible to define the set of potential solutions of the diagnostic system directly from the defined need. This problem can be solved by using an additional project's stage, during which the formalized description of the need is prepared. Due to the fact that the diagnostic system contains the software part that is responsible for analyzing of the signals and inference about object state, it seems reasonable to use certain design techniques that are directly derived from the software engineering area. The sets of requirements that describe the expected functionality of the designed systems are one of them.

In the process of designing the diagnostic systems requirements can be used to formalize the notation of the project's need. Within the defined need, some characteristic functionalities that are necessary from the diagnostic system point of view may be extracted. These functionalities can be accurately characterized by applying the requirements with the hierarchical structure. Stored in such way the need, would be better expressed/described expectations about the potential diagnostic system solutions, because it will take into consideration the specificity of the field which is the technical diagnostics and the characteristics of the technical object for which there is a need to develop a diagnostic system. Based on such formalized need it is possible to develop/generate a set of possible solutions of the diagnostic system. In the next stage of the design process, these solutions will be evaluated on the basis of the set of established criteria, wherein these criteria may be defined/derived from a set of defined requirements.

Individual requirements may be defined based on the available diagnostic and operation knowledge, knowledge about the considered technical object etc. An important advantage of the use of requirements in the design process is the ability to define the individual requirements by independent persons, so-called domain experts. This is particularly important in the area of technical diagnostics, due to the fact that the modern technical objects are very often complicated and the design the diagnostic system by the small group of people may be very difficult.

## III. Requirement engineering

In literature, there are available many different definitions of the term requirement e.g. in [12], [16] or [17]. For the purpose of presented studies, the following statement may be assumed: *Requirement is a formalized description of function/atributte, that designed object should realize/meet.* Just as in the literature exist many definition of the term requirement, the different divisions and classifications of requirements exist also. However, two groups of requirements i.e. functional and nonfunctional requirements are usually distinguished. Functional requirements describe the basic functions of the system, i.e. the way in which the system operates, colloquially what the system should to do. This, how the particular functions of the system should be implemented is described by the nonfunctional requirements. Very often there are also called as a quality requirements due to the fact that they describe the operation of the system (implementation of particular functional requirements) from the quality point of view. Repeatedly, an overall evaluation of the developed system and the recognition that the system has been developed according to the assumptions depend from them [12], [16].

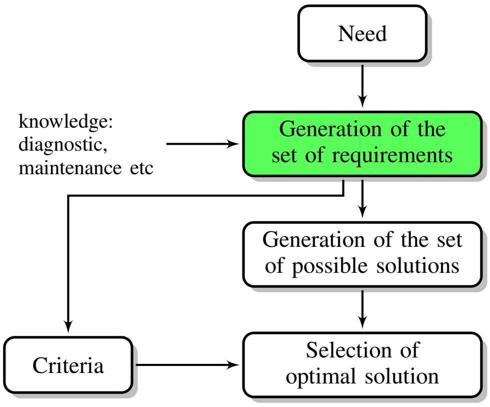

Need
knowledge:
diagnostic,
maintenance etc
Generation of the set of requirements
Generation of the set of possible solutions
Criteria
Selection of optimal solution

Fig. 1. Scheme of the proposed diagnostic systems design process

## IV. Notation of the set of requirements

The use of requirements for the purpose of formalized description of the need, requires the introduction some additional assumptions, that are not seen in the software engineering area. These assumptions are connected with the necessity of adoption of a specific manner of notation of defined requirements. This notation is caused by the specific properties of the area that is the technical diagnostics. Namely, the detection of individual faults is considered as the major functionality of the diagnostic system. In order to achieve this task, certain diagnostic rules (symptoms of faults) should be applied. Each diagnostic rule may be realized with using some measurement methods e.g. measurement and analysis of temperature, vibration measurement etc., which include various types of sensors, measurement cards etc. All of these elements must be saved as a component of the established definition of the requirement.

Taking into account the mentioned remarks, each functional requirements $req_i$ may be recorded as

$$req_i =< c_r, attr_r, pf_r, drule >, \qquad (2)$$

where:
$c_r$ – content of requirement, $attr_r$ – attributes of requirement, $pf_r$ – preference factor, $drule$ – diagnostic rule.
The content of the requirement is the description of the functionality of the diagnostic system. It is related with the necessity of detection on specific fault e.g. *The necessity to*

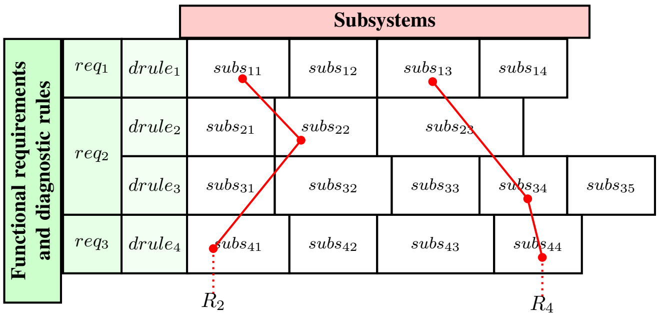


Fig. 2. Morphological table with the set of possible solutions of diagnostic system project

*detect breakage of the gear tooth*. For each requirement it is possible to assignment some set of attributes,

$$attr_r = \{attr_{r1}, attr_{r2}, ...\}, \tag{3}$$

which are used to detail the description of this requirement and are the basis for evaluating the usability of the functionality which are described by this requirement. *Level of reduction the probability of fault occurrence described by the requirement*, or *level of reduction of costs associated with the occurrence of this fault* may the examples of these attributes. For each functional requirement, is possible to assign some value from the range $< 0; 1 >$, called the degree of preference. It expresses the need to take into account this requirement in the final solution of the diagnostic system. Preference factor may be considered as a subjective assessment of the suitability of the given requirement to the achieve by the diagnostic system the established goal, which is the minimization of technical risk. For each functional requirement is possible to define the set of possible diagnostic rules:

$$drule = \{drule_1, drule_2, ...\}, \tag{4}$$

which are used to detect the specific fault. Each diagnostic rule is expressed as:

$$drule_i =< c_{dr}, attr_{dr}, pf_{dr}, subs, ko > . \tag{5}$$

Scheme of marks is similar like in the case of requirements. So $c_{dr}$ is a verbal description of the rule, e.g. *exceeded the root mean square value of vibration signal*, $attr_{dr}$ is a set of possible attributes, that may be used to accurately describe each of the rule e.g. *level of the false alarms*, and $pf_{dr}$ is a degree of preference assigned to each rule. Each of the diagnostic rules may be fulfilled by using different measurement systems, for the purpose of this study, called subsystems. Each subsystem $subs_i$ may be expressed as:

$$subs_i =< elem, attr_s, pf_s >, \tag{6}$$

where:
$elem$ – subsystem elements, $atrr_s$ – attributes of subsystems, $pf_s$ – preference factor.
Each element of subsystem may be considered as:

$$elem_i =< elem_{type}, elem_{value}, attr >, \tag{7}$$

where $elem_{type}$ is a type of element, e.g. relative vibration sensor, measurement card (DAQ card) etc. and $elem_{value}$ is a possible items of the specific element type, that is, e.g. specific sensors from catalogues of certain producent. These elements may be described by numerous attributes $attr$ e.g. measurement range, sensitivity, cost etc. For the purpose of limiting the applicability of the selected elements included in the subsystems, it is possible to define a set of limiting criteria $ko$ that are assigned to each diagnostic rules $drule_i$. Diagnostic rules and subsystems assigned to them may be regarded as non-functional requirements defined for selected functional requirements, that describe the necessity to detection of individual faults.

Requirements (i.e. hierarchically saved elements $req_i$, $drule_i$ and $subs_i$), that are gathered by using the proposed method, describe in the formalized manner the project's need. Based on this formalized need, it is possible to generate the set of possible solutions of the diagnostic system. To perform this operation it is necessary to use the catalog of possible elements that are the part of subsystems $subs_i$, and take into account the limiting criteria. In this way, for the defined combination of functional requirement - diagnostic rule, is generated a set of possible subsystems. These three elements i.e. functional requirement, diagnostic rule and subsystem are the component of the project of the diagnostic system. Morphological table [18] (see example in Fig. 2) may be used for easier presentation the set of possible solutions of this diagnostic system.

In order to use the proposed method of formalization of project's need, it is necessary to answer on two fundamental questions. First of all, how to acquire the necessary set of requirements. And the second, how to evaluate the generated set of possible solutions so as to choose the optimal solution. The answer to these two questions is contained in two successive chapters.

## V. Requirements acquisition

The process of gathering the requirements for the purpose of design diagnostic system is not an easy task. As in the software engineering area [12], [16], it depends on many factors, including the availability of knowledge about the possible faults of object, methods of their diagnosis etc., [5]. The process of obtaining the relevant requirements, should therefore be adaptive aided by using different kinds of knowledge banks, catalogues of elements, etc. In general case, for each technical object it is possible to develop a so-called catalog of hazardous events, which includes the possible faults of its individual components. For individual fault, it is possible to estimate the risk (probability/frequency of occurrence of the event, the potential costs of the consequences of this event) associated with their occurrence. Based on this set of faults, it is possible to define the set of diagnostic rules, that include typical symptoms of individual faults. For the purpose of describing the subsystems that are used to provide various diagnostic rules, it is possible to create a catalog of elements in the form of pairs: element type - a list of elements.

Using data banks prepared in this way, is possible to perform the process of requirements acquisition, that assumes the defining of following elements that are included in the definition of requirements, in hierarchical order, i.e. functional requirement, diagnostic rule and subsystem, whereby during this stage, it is possible to use the pre-defined knowledge banks. It is assumed that, depending on the needs and the technical possibilities (see Sec. VII) for individual objects may be assigned different attributes. These attributes are the basis for assessment of the suitability of individual requirements, and also allow to generate a set of possible solutions of the diagnostic system. Additionally, it is assumed that during the stage of defining the diagnostic rules, it is possible to define the limits (limiting criteria) that must be met by the subsystems used to implement the specific diagnostic rule. However, the limiting criteria must refer to the attributes assigned to individual elements.

An important advantage of using this approach is the possibility of a separate preparation knowledge banks and acquiring the set of requirements. Both of these stages can be performed not only independently, but also by different persons, such as domain experts. The entire process of obtaining requirements can be further simplified by making the decomposition of the technical object and define the requirements for such separate components.

## VI. Requirements evaluation

In the software engineering area the process of evaluation of obtained requirements takes many forms. However, it usually amounts to the negotiation between the customer that order the project and its potential developer. Each of them articulate own needs and expectations. After a series of conversations, they together agree a final set of established requirements [8]. This approach, due to the several factors cannot directly be used in the case of designing the diagnostic systems. The first of them is the lack of explicit customer for the requirements negotiation process. A person who contracts or expresses the need to perform diagnostic system in many cases not have complete knowledge about the considered technical object. Therefore, this person will not be able to correctly assess the whole set of requirements. Additionally, in many cases the knowledge about technical object, particularly the knowledge about the manner of its diagnosis may not be fully accessible and good quality. So, it is necessary to apply a some software system, that will be support the process of evaluation of requirements based on the collected data. This system may be treated as a virtual customer, that represents the technical object and articulates the necessary requirements.

The overall objective of the stage of the requirements evaluation, is to determine the answer for two fundamental questions, namely:

- what should be the functionalities of the diagnostic system, i.e. which faults should be detected by the diagnostic system and which diagnostic rules should be applied for this purpose,
- what should be the structure of the diagnostic system, i.e. which subsystems that carrying out specific diagnostic rules should be applied and which items from available element types should be used.

For the purpose of carrying out the mentioned above activities, it seems reasonable to develop an expert system. It allows not only to carry out the process of inference about the usefulness of individual requirements, taking into account the existing data/information, but also the assessment of individual solutions of the diagnostic system in order to select an optimal solution.

The basic components of a typical expert system is a knowledge base, database, inference algorithm and the user interface, through which is provided the access to the individual components of system [14]. The use of expert system for the described purposes, requires the develop of the knowledge base in the appropriate form. This is due to the necessity to take into account the numerous set of attributes, based on which will be performed the process of evaluating of individual requirements. For the purpose of developing the knowledge base the graphic models e.g statement networks [4] or belief network [9], [11] can be used. In considered case the belief networks are used.

Belief network is a directed graph $G$ in which the set of nodes $N$ and set of directed edges $V$ may be distinguished. Nodes of this network are used to represent the available information on observed facts and made observations. Causal relationships between the selected facts are represented by defined edges between the relevant nodes. The size of the impact of individual nodes to each other is determined in the conditional probabilities tables that are defined for each of nodes.

In the presented case, the structure of required belief networks is constant. Example of the structure of the required network that is necessary to represent the knowledge associated with the functional requirements is shown in Fig. 3. Nodes of four classes i.e. *requirements*, *attributes for requirements*, *criteria*, and *auxiliary nodes* exist in this structure. The attributes and criteria for attributes are the input nodes, while the functional requirements are the output nodes. As the values of the *attributes* nodes, qualitative data collected at the stage of defining requirements are introduced. *Criterion* node determines whether the attribute is important from the point of view of the project of a diagnostic system. Conditional probability tables are fixed for nodes within each group and may be elaborated based on available literature data or expert's knowledge. In the same way the network for diagnostic rules can be defined. As a result of the inference process, for the individual functional requirements and diagnostic rules the preference factor is assigned. The value of this factor indicates about whether the certain solution (requirement or diagnostic rule) is desirable in the designed diagnostic system.

The stage of assessment of subsystems is the next process that can be performed by using expert system. It is assumed, that based on the obtained database of elements and defined constraints imposed on the subsystems, is possible to generate very large set of potential subsystems. The process of their evaluation involves the use of optimization functions, that

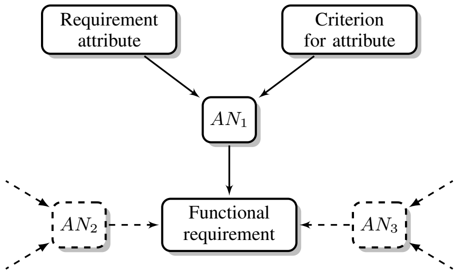


Fig. 3. Structure of belief network for considered task, $AN_i$ – auxiliary node

looking for the best solution. The most commonly used functions minimize or maximize the value of one or more attributes assigned to the individual elements e.g. cost minimization. In this way the set of desirable functional requirements and rules and within their the optimal subsystems are determined. The optimal structure of the diagnostic system is generated by the choice of a certain subset of functional requirements (e.g. those for which the appointed degree of preference is greater than the certain predetermined value) and iterative merge of selected subsystems. The iterative merge assumes, that is possible to omit certain elements of subsystems, for example measurement cards, because they are already elements of another subsystems and can be used repeatedly. This approach allows in many cases to simplify the structure of the diagnostic system as well as to reduce its cost.

## VII. Short example of proposed approach

For the purpose of presentation described in this article the method of aided the development of diagnostic systems process, one simple example will be presented. A simple technical object which is typical rotating machine equipped with the ball bearings is considered. For this object it is possible to define the needs that may be written as: *determine the current technical state of the object*. As it has already been mentioned, this need may be met by developing appropriate diagnostic system. For such defined problem, it is possible to begin the process of defining requirements. As a result of this process, the set of functional requirements may be defined. Examples of these requirements may be given as follows:

- $req_1$ – the need to detect the wear of the bearing raceway,
- $req_2$ – the need to detect crack of the outer raceway of the bearing.

These requirements may be described by the following attributes:

- the level of reduction of the probability of fault occurrence,
- the level of reduction of cost associated with occurrence of fault,
- downtime of the object after fault occurrence,
- the applicability of the requirement in other projects of the diagnostic systems.

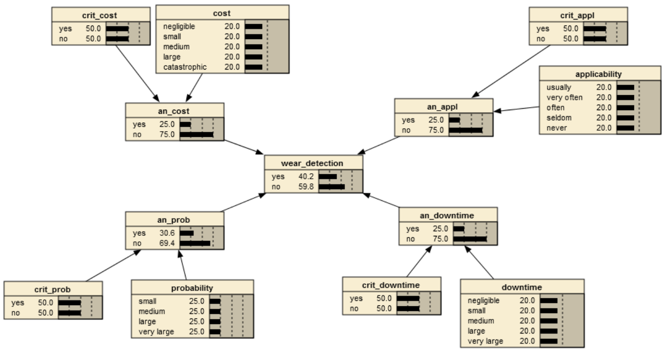


Fig. 4. Belief network with elementary data

Considering only the requirement $req_1$, it is possible to define for it a diagnostic rule written as, $drule_1$ – *exceeding the allowable root mean square value of vibration signal in the bearing*, and the limiting criterion for the subsystem on the content: *measuring range up to 10kHz*. Diagnostic rule $drule_1$ can be realized by the subsystem of the structure: absolute vibration sensor, connecting cables, measuring card and software for signal analysis. Fragment of a possible catalog of available sensors may include for example such elements as: CMSS 2100, CMSS 2100T, CMSS 2200, CMSS 780C, CMSS 2100 etc. all of these sensors are produced by the SKF company. Described process of requirements defining should be carried out in the same way for the remaining functional requirements. For the purpose of the presentation the inference process with using the belief networks the Netica environment was used. The defined statements network for the elementary data is shown in Fig. 4. The network structure is consistent with the assumptions shown in Fig. 3 and described in Sec. VI. After entering to the network the data related with: the probability of occurrence of considered fault, possible cost of the fault occurrence and information about the necessity to take into account some attributes, the value of preference factor for the considered requirement will be appointed (see Fig. 5). The next stage of the assessment involves the selection of the optimal form of the subsystem. For this purpose, based on the available sets of items and the defined limiting criterion, the set of possible solutions is generated. In this set the solution characterized by the minimum cost is now sought.

Taking into account that for the considered technical object, it is possible to define large set of requirements, the described the way of proceeding should be carried out for each of them. At the end, it is necessary to carry out the optimization of the whole structure of diagnostic system, to exclude repeating items.

## VIII. Environment for diagnostic systems design process

Presented in this article, the method of elaborating of the projects of the diagnostic systems, may be aided by using a dedicated software environment. Four main modules, i.e. requirements acquisition module, requirements evaluation module, presentation module as well as database may be dis-

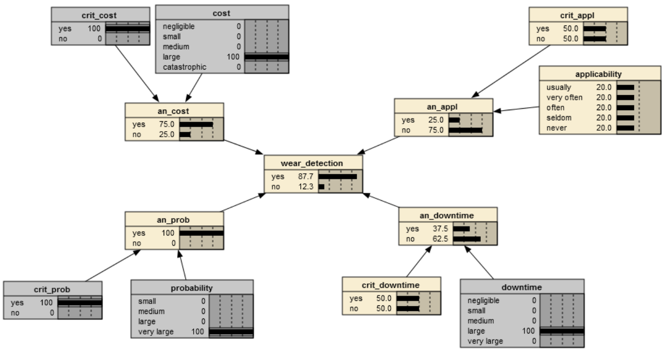


Fig. 5. Belief network after entering the obtained data

tinguished in the structure of this environment. Requirements acquisition module is responsible for collecting and storing requirements that describe developed diagnostic system. It cooperates with the banks of knowledge, in which are recorded information about known collections of faults, possible diagnostic rules, and known catalogues of elements in the form of pairs type of element - element. Based on the defined functional requirements and diagnostic rules the knowledge base for expert system is generated. The inference process will be carried out by using the package *REx* that is developed in R programming language [6]. The preference factot is assigned to each functional requirements and diagnostic rules as the result of inference process. The process of evaluating the individual subsystems and integration of the overall project of the diagnostic system will be performed in the evaluation module. For the purpose of presentation of the results the presentation module will be used, while the individual data, knowledge, etc. will be stored in the suitably developed database. Currently this environment is under construction.

## IX. Conclusion

The process of designing the diagnostic systems is a very complex task. In many cases these systems are prepared for the critical machines e.g. turbines of power plant, etc. that appropriate diagnosis is essential. During the design process of these systems is necessary to take into account numerous factors including available knowledge, existing limitations etc. In order to properly management of the available data, it is necessary to support the process of defining the diagnostic systems. For these purposes the set of requirements may be used. The used requirements may formalize the notation of the main project need. The formalized notation of the need assumes that exist three elements i.e. functional requirements, diagnostic rules and subsystems. Each of these elements can be developed based on existing catalogues of faults, measurement systems or catalogues of elements of those systems. Based on the formalized project need it is possible to generate the set of possible solution of the diagnostic systems. Because the evaluation of these solution, there is the selecting of the optimal solution is a challenging task, the additional software environment so-called expert system may be used. As a result of the use of expert system the optimal solution of the designed diagnostic system is isolated. Each of the necessary stages of the proposed approach may be performed by using specialized environment which is currently being developed. Future work will be related to the testing of this environment and removing possible inconveniences and drawbacks of the proposed method.

## Acknowledgment

Described herein are selected results of study, supported partly from the budget of Research Task No. 4 implemented under the strategic program of The National Centre for Research and Development in Poland entitled *Advanced technologies of generating energy*.

## References

[1] M. Amarowicz, Simulators for defining of requirements for diagnostic systems, In J. Korbicz and M. Kowal, (Eds.), *Intelligent Systems in Technical and Medical Diagnostics*, vol. 230 of Advances in Intelligent Systems and Computing, Springer Berlin Heidelberg, 2014, pp. 187–198, http://dx.doi.org/10.1007/978-3-642-39881-0_15.

[2] M. Amarowicz, Diagnostic Systems Design Process With Using The Sets Of Requirements, *Diagnostyka – Applied Structural Health, Usage and Condition Monitoring,* vol. 15, no. 2, 2014, pp. 19–26.

[3] T. Barszcz, *Systems for monitoring and diagnostics of machines* (in Polish), ITE–PIB, Kraków, 2006.

[4] W. Cholewa, Dynamical statement networks. In J. Awrejcewicz (Ed.), *Springer Proceedings in Mathematics & Statistics*, vol. 93 of Springer Proceedings in Mathematics Statistics. Springer International Publishing, 2014, pp. 351–361, http://dx.doi.org/10.1007/978-3-319-08266-0_25.

[5] W. Cholewa and M. Amarowicz, Acquisition of requirements for diagnostic systems, *Diagnostyka – Applied Structural Health, Usage and Condition Monitoring*, vol. 13, no. 2, 2012, pp. 23–30.

[6] W. Cholewa, T. Rogala, P. Chrzanowski, M. Amarowicz, Statement Networks Development Environment REx. In: P. Jędrzejowicz, N. Nguyen, K. Hoang (Eds.) *Computational Collective Intelligence. Technologies and Applications*. LNCS, Springer, Heidelberg. vol. 6923, pp. 30–39 http://dx.doi.org/10.1007/978-3-642-23938-0_4.

[7] J. Dietrych, *System and construction* (in Polish), WNT, Warszawa, 1978.

[8] P. Grünbacher and N. Seyff, Requirements Negotiation. In: A. Aurum and C. Wohlin, (Eds.) *Engineering and Managing Software Requirements*, Springer, Berlin Heidelberg, 2005, pp. 143–162, http://dx.doi.org/10.1007/3-540-28244-0_7.

[9] F. V. Jensen, *Introduction to Bayesian Networks*, Springer, Berlin 1997.

[10] J. Korbicz, J. M. Kościelny, Z. Kowalczuk and W. Cholewa (Eds.), *Fault Diagnosis. Models, Artificial Intelligence, Applications*, Springer-Verlag, Berlin, 2004, http://dx.doi.org/10.1007/978-3-642-18615-8.

[11] T. Koski and J. Noble, *Bayesian Networks. An Introduction*, John Wiley & Sons, New York, 2001.

[12] D. Leffingwell and D. Widrig, *Managing Software Requirements: A Use Case Approach*, 2 edition. Pearson Education, Boston, 2003.

[13] S. Legutko, *Fundamentals of operation of machines and devices* (in Polish). Wydawnictwa Szkolne i Pedagogiczne, Pozna«, 2004.

[14] J. *Liebowitz The Handbook of Applied Expert Systems*, CRC Press, Boca Raton, 1997.

[15] P. O'Connor and A. Kleyner, *Practical Reliability Engineering*, 5 edition, Wiley, 2011.

[16] I. Sommerville and P. Sawyer, *Requirement Engineering: A Good Practice Guide,* John Wiley & Sons, 1997.

[17] K. Wiegers and J. Beatty, *Software Requirements,* 3 edition, Microsoft Press, 2014.

[18] F. Zwicky, *Discovery, invention, research through the morphological approach*, MacMillan, 1969.

DOI: 10.15439/2015F328
ACSIS, Vol. 5

# Process Approach to Management Knowledge Objects Managerial Expertise for Distance Learning

Krzysztof Hauke
Wroclaw University of Economics
Komandorska 118/120,
53-345 Wrocław, Poland
Email: krzysztof.hauke@ue.wroc.pl

***Abstract*— One of the basic features of the information society is a constant acquisition of information. This information can be obtained from various sources. One source is the Internet environment. In this environment, you can use the standard communication of information on the web. The disadvantage of this approach is the lack of assessment of knowledge by the recipient. If you want to professionally lead the learning process then we have to use distance learning technologies. The use of distance learning systems, forcing the recipient to the following activities: acquire new knowledge and its evaluation. Note, however, that the knowledge in the course offered must be current. Then it makes sense to use distance learning systems.**

***The paper will be presented process approach to management knowledge objects managerial expertise for distance learning systems. Process management is carried out in the following steps. knowledge acquisition, knowledge Locating, developing knowledge, preservation of knowledge, knowledge sharing, knowledge utilization. The use of learning objects taking into account the process approach to improving the quality of the courses offered in the form of e-learning.***

***Keywords:* e-learning, knowledge, knowledge management, knowledge object, knowledge management process approach.**

## I. Introduction

The traditional model of the teacher - student knowledge transfer becomes quite a big limitation. The limitation is due to lack of time and logistic reasons. However, keep in mind that this model is good especially when you consider news aspect of their contents. In order to try to alleviate the constraints are involved with information technology solutions such as e-learning. From the point of view of the process of communicating, content society is already so educated, that it does not have much of a problem. The problem, however, lies in the interior of such a system. Didactic units are created by traditional human team. Immediately after completion of the work related to the availability of content is generally rated very well. However, after some time, a substantive assessment related to the content will be increasingly lower until it can be quite inadequate to the environment and the existing state of knowledge in the field. In order to avoid this uncomfortable situation a didactic unit or course must be constantly evaluated. Any derogation as to the content on a regular basis must be corrected to prevent such a situation that learners will use the outdated knowledge without this awareness. However, to improve the content can not be consistently implemented from the outset. Such an approach leads to a significant prolongation update the content of the course. The speed associated with upgrading the content is inherent to the process of knowledge management. If you consider elements of knowledge management in this course including activities associated with the creation or update it, will no longer be time-consuming. Note, however, that implemented elements of knowledge management in e-learning systems will not eliminate the traditional mentor in preparation for the actual content of educational material [1] [2].

## I. Learning objects in educational materials

The information society is a new type of society, which is characterized by information processing. One of the sectors which is engaged in information processing is education. Traditionally provided education requires challenges for mentors returning knowledge through lectures, exercises, conservatories and laboratories. Progressive growth of information generated by the various institutions of this society compels members to look into the issue of the organization of knowledge. In the eighties of the last century we had seen very rapid growth in the use of information technology in teaching [3]. Mentors began to prepare classes with the use of tools. It was noted unsatisfactonary with these solutions related to the teaching process. There were tools (course management systems), which in a comprehensive manner solved the problems of teaching. However, even here there are problems. Each of the mentors worked out on an individual didactic material for the course. The development of such a course would be individual to the issue and was of the author. In addition, be aware of the dynamics of the environment from the point of view:

- time that provides more and more information in the following units of time,
- a place that generates information specific to their environment.

But the uniqueness of feature associated with the time to create such a course. The preparation of such teaching

978-83-60810-66-8/$25.00©2015, IEEE

material that takes into account any change of scene and adjusts to the time and place is very difficult. This problem can be solved by developing courses using the components. Components of the so-called learning objects (learning object - LO). The concept of learning objects was first described in the work of Gerard in 1967. Initially, this concept does not use the term object of knowledge. The first time the concept of learning object was used 1994 by Wayne Hodgins, who created a working group Computer Education Management Association (CEdMa) dealing with objects of knowledge [12].

In the literature you can find many definitions of learning objects.

General Definition - the object of knowledge is any element constituting a whole (entity) in digital form or not, which can be used in the process of learning, teaching or training.

This learning object is the smallest independent structural elements comprising three components:

- purpose - specifying expected outcomes of learning / teaching;
- activity - that this part of the knowledge element, which achieves the objective pursued;
- evaluation - allows an indication of how the expected objective has been achieved.

According to Mr Shepherd - the object of knowledge is a small, digital reusable component that can be used - alone or in combination with others - using computer software by the creator of the content or independently by the learner, for educational purposes [5].

According to S. Mills - the object of knowledge can be used to meet the teaching and the intended results can be borrowed from a different educational environment. Knowledge reusable objects are associated with electronic educational sources that can be used by different learning environments [14].

Definition author - learning objects are self-contained and independent objects that describe a reality that can be used an infinite number of times in various educational courses.

The creation of such an object is very difficult. Creating a learning object should take into account quality of universality. This will ensure that the use of such a universal object in other educational materials.

The object of knowledge is an independent piece of information, which might function independently, designed for reuse, used to create a lecture held at a distance. The object of knowledge is associated with:

- content / object content,
- metadata (tags) that describe the object keywords,
- educational material management systems LMS / LCMS (Learning Management System / Learning Content Management System).

The figure below (Fig. 1) shows the relationship between the object of knowledge, metadata and educational material management systems.

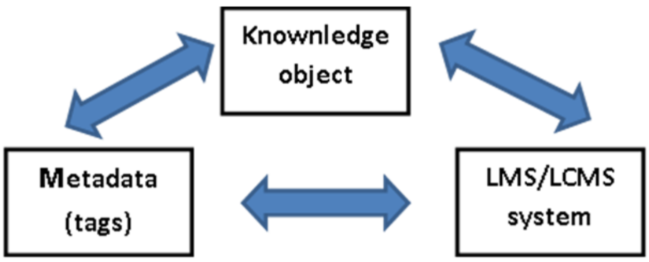


Fig. 1. Environment the object of knowledge

Source: Own study based on: http://www.itpedia.pl/index.php/Grafika:E-lear_5.jpg

From the above considerations the common elements of knowledge object definition can be derived::

- content - the purpose of training, content and actions in the area of knowledge transfer needed to achieve this goal and objective assessment reflecting the training,
- size or time needed to take advantage of knowledge - party knowledge that assimilation takes no more than 15 minutes,
- context and characteristics - knowledge that can operate independently and be delivered to the listener, if necessary, exactly on time and in sufficient quantity,
- labeling and remembering - the part of knowledge that describes a standardized set of tags.

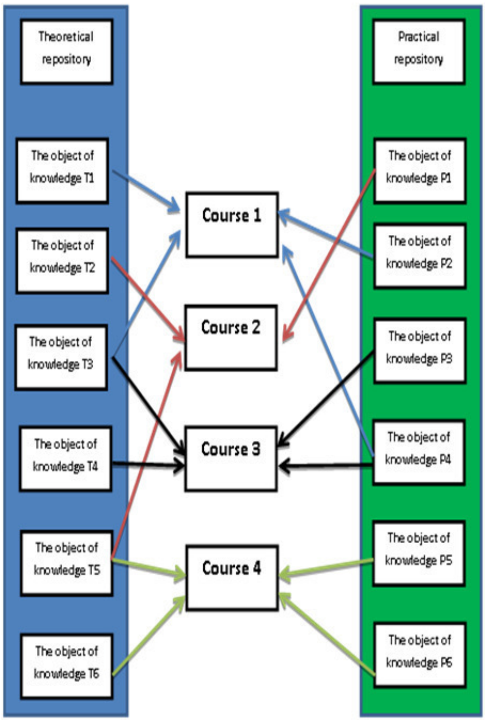

Theoretical repository
Practical repository
The object of knowledge T1
The object of knowledge T2
The object of knowledge T3
The object of knowledge T4
The object of knowledge T5
The object of knowledge T6
Course 1
Course 2
Course 3
Course 4
The object of knowledge P1
The object of knowledge P2
The object of knowledge P3
The object of knowledge P4
The object of knowledge P5
The object of knowledge P6

Fig. 2. Schematic combining learning objects into a finished course

Source: Own study based on http://www.itpedia.pl/index.php/Grafika:E-lear_6.jpg

The object of knowledge can be represented by: text, graphic image, animation, multimedia audio, video [4]. a combination of these elements [13].

Learning objects are managed via the Learning Management System (LMS). Learning objects can be grouped into thematic repositories individually determined by the management of these facilities.

Repository - the place for orderly storage of documents, all of which are designed for sharing. The concept identified with the main storage, the central, but designed in such a way that access to all of its resources was equally easy. In the era of information technology there is also a term used in relation to all sorts of digital resources (databases, set of packages or source code), for example on the Internet [14].

Repository relates primarily to storage, and no sharing. It is a storage at the same time: the main, central, current and easily accessible.

Knowledge repository - a repository of documents with specific subject matter, with a specific to the field of mechanisms that facilitate access to information, or / and mechanisms for synthesizing additional information based on content stored documents.

Repository of domain, also called thematic repository and the repository of knowledge, according to. Lexicon Distance learning: the repository is a storage place structured domain knowledge, intended for repeated use in different contexts. Knowledge is presented in the form of a portion called - LO (Learning Object) In other words, it is a specialized computer system, which performs the functions of storing, sharing and modifying structured domain knowledge.

Features a repository of knowledge

- completeness - chronological, thematic, etc,
- timeliness - the new data may change old,
- relationships - mapped data relationships,
- searchability - FTS, taxonomies,
- accessibility - digital form, the Internet, PDAs, Epaper.

In the picture above (Fig. 2) learning objects are divided into two repositories:

- practical (Pn) - contains objects resulting from the observation of reality, case studies, calculations, collected empirical material
- theoretical (Tn) - contains objects resulting from the theory related to the matter, definitions, concepts, structure, interpretations, which are contained in the books, published in the form of compact or posted on websites.

This approach allows the use of multiple learning object developed in various courses. It was only at the stage of designing the course designer can benefit from such a facility in order to obtain a particular course.

## II. The Evolution of Knowledge

The organization of the knowledge allows you to achieve new opportunities. Accordingly the strategy of its organization can contribute to a rapid response to stimuli from the environment. The strategy becomes a tool for organizing the accumulated knowledge from a particular individual. Be aware of the fact that the ordering of knowledge should be fully aware of the act and the result of a desire to dominate over its generation by a number of individual and collective objects. Mastering the body of knowledge related to the functioning of the matter may become the primary factor leading to success. We must realize that today's information society generates enormous amount of information even in the smallest unit of time. Organizing them as a whole would be an extremely difficult issue to pursue. Information technology organization is trying to help this knowledge as a whole. You can mention the concept of big data, which lets you store knowledge objects. At present, it is a challenge for the information society, which must determine the direction of its activities for the coming years. Currently, a better solution is to adopt a strategy for organizing knowledge in very limited areas, resulting from specialization. Such organized knowledge will be useful for potential people who would want to solve the problems of an individual or for the environment. Such an approach requires a focus on a very narrow area of activity. This leads to specialization of an individual or collective entity in the field of influence. The limited approach to knowledge organization for the area allows you to assign its high utility value. The big economic corporations using so organized knowledge can make decisions of a strategic nature. Of course, because of the dynamics of the environment when making strategic decisions should take into account current knowledge and adapt in activities that may result in the future.

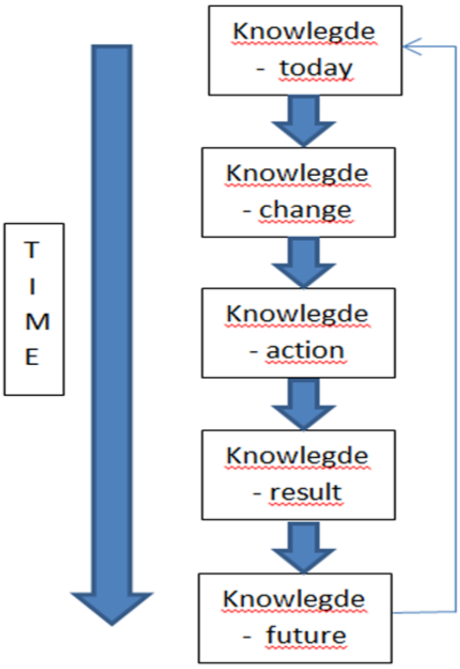


Fig. 3. The evolution of knowledge
Source: Own study

The process of evolution of knowledge is dependent on time. In the figure above (Fig. 3) the process of evolution is shown in five steps:

- Knowledge today - it is the state of knowledge on a specific state of our reality.
- Change of knowledge - a process based on the fact that by comparing the state of knowledge in the information system with real environment arising.
- Action on knowledge - declare that appeared in the environment or you may receive new knowledge forces administrators to system.
- Measurement of knowledge - at this stage the control of the state of knowledge in the information system of the update knowledge of the environment.
- Knowledge in future - it reaches the result of this step is to ensure comparability of knowledge in the system and the knowledge derived from the environment. After this stage, your "future" knowledge becomes "today" and may be used by the public.

The evolution of knowledge is an iterative process. Depending on the knowledge of frequency witch be very different. Knowledge can be changed in very short periods of time.

## III. Process approach to knowledge management

In a process approach the following are taken into account by G. Prost, S. Raub, K. Romhardt:

- locating knowledge,
- acquisition knowledge,
- developing knowledge,
- preserving knowledge,
- sharing knowledge,
- using the knowledge (Fig. 4)[11] [7]

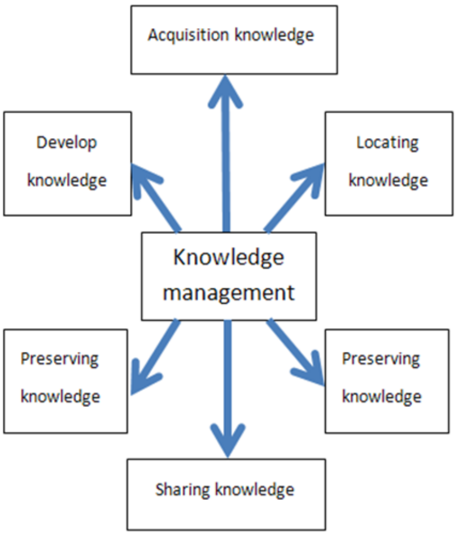


Fig. 4. Elements of knowledge management
Source: Own study [11]

**Locating of knowledge** - a basic function of solving the above problems. However, you will notice that the users in defining the demand for expertise in early stage have very big problems with identifying what knowledge they need.

**Acquisition knowledge** - knowledge can come from different sources. Each of the entities in the end generates a knowledge of the differentiated value.

**Developing knowledge** - it is the process of acquiring complementary expertise. The mere knowledge acquisition over time due to turbulence of the environment can lead to stagnation of the acquired knowledge.

**Preserving knowledge** - can work the right tools and means of media in order to replay, as well as activities undertaken in the field of detention experienced and skilled employees, as the most severe form of loss.

**Sharing knowledge and its dissemination** - this is one of the most difficult processes of knowledge management. Because of the development of the information society consciousness more and more we appreciate its value.

**Using the knowledge** - is a process that is used to achieve the objective of the organization. The organization wanting to accomplish their goal should seek to use the findings obtained.

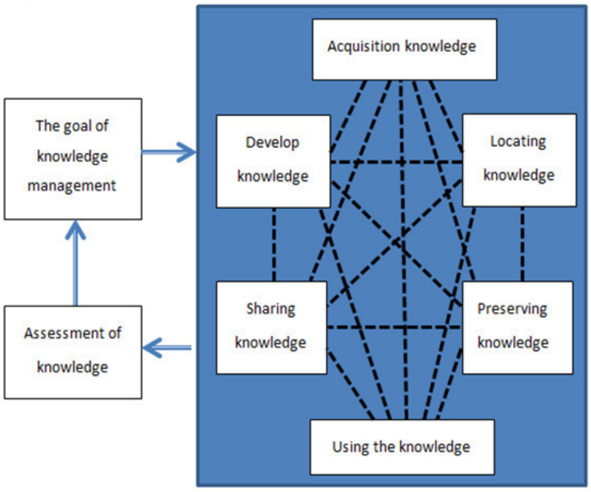


Fig. 5. Knowledge management environment
Source: Own study [11]

The activation of any of the six elements determines the assessment of knowledge. If the assessment of knowledge is inadequate to the knowledge of the environment it must be restored to its desired state - founded at the beginning of the creation of educational material.

The proposed approach will allow for an immediate reaction on outdated knowledge in the e-learning system.

## IV. LCMS as an example of a process approach to managerial knowledge management

Learning Content Management System (LCMS) is a software application that allows content management training. The LCMS helps you to create, use, locate, deliver, manage and improve training content.

The LCMS is able to locate and deliver to the end user the personalized training unit to cater to a single request, or to provide more elements of the course. Fig. 6 shows the managerial knowledge management process through e-learning system - Learning Content Management System (LCMS) [8].

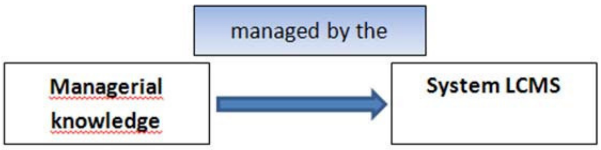


Fig. 6. Knowledge management environment
Source: Own study

Each course is a complex system, which (in the design) can be divided into small separate objects, called "learning objects" and that after the merge create the course content. Fig. 6 shows how you can create repositories managerial knowledge specific to the operating level, management and strategic management of the organization [6].

The basic components of the system LCMS are:

- **Repository of learning objects** - it is a centralized database in which content knowledge are stored and managed. Content can be made available to users as a single object of knowledge, in the form of larger modules or as an entire course, according to individual requirements.
- **Automated authentication application** - it is used when creating reusable learning objects, which are available from the repository using templates containing scenarios, training design principles. It supports both authors, slingshots automates the development of learning objects.
- **Dynamic publishing interface** - provides learning objects based on profiles, preliminary tests and / or user queries.
- **Administration** - it is used to manage lists of users sharing directories courses with courses, tracking and reporting the progress of the participants [9].

Functional modules of LCMS (Learning Content Management System):

- **The object repository** - the module object repository is a central database in which are stored all the elements included in the course. From this place objects that make e-learning training they are sent to the students.
- **The module automates build courses** - in this module to create the objects included in the rate (SCO - Sharable Content Objects) module facilitates work by providing templates, as well as the full list of existing objects. which can be re-utilization, processed, copied.
- **The distribution of courses** - foreign distribution module allows you to share courses to students according to established profiles. It also allows you to track the progress of the trainee and reports the results of exercise testing questions.
- **Administration module** - it is used to manage the process of learning: account management trainees, providing them with courses, tracking the progress of science and carry out other administrative tasks [10]. The following figure (Fig. 8) shows the overall process of creating the course.

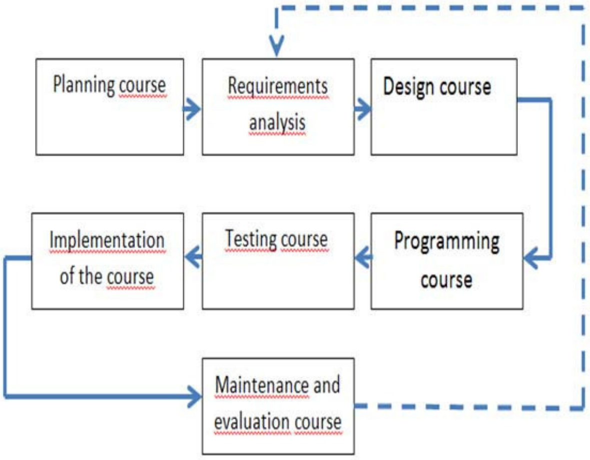


Fig. 8. The process of creating e-learning course
Source: Own study

In the process of building the e-learning course we should use the LCMS system. It will help to support the knowledge management function. Thanks to this course it will be able to quickly adapt to changes resulting from the dynamics of the environment. The figure below illustrates the LCMS system in the environment of knowledge management functions.

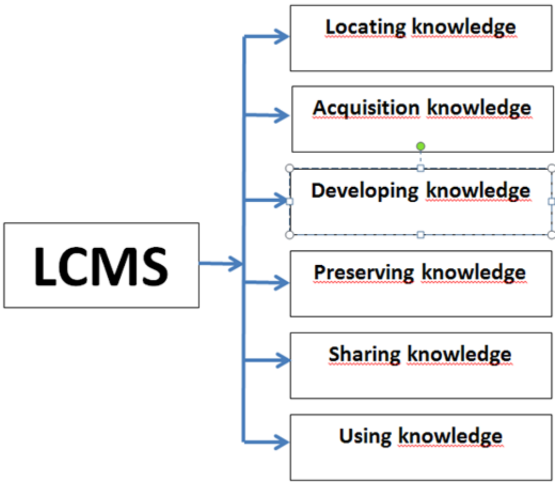


Fig. 9. LCMS in an environment of knowledge management function
Source: Own study

The following analysis will be carried out to present the LCMS system due to the components.

Components - **Repository of learning objects** - of the system LCMS supports:

- locating knowledge,

- acquisition knowledge,
- developing knowledge,
- preserving knowledge,
- sharing knowledge,
- using knowledge.

Components - **Automated authentication** application – of the system LCMS supports:

- locating knowledge,
- acquisition knowledge,
- developing knowledge,
- preserving knowledge,
- sharing knowledge,
- using knowledge.

Components - **Dynamic publishing interface** – of the system LCMS supports:

- acquisition knowledge,
- preserving knowledge,
- using knowledge.

Components - **Administation** – of the system LCMS supports:

- acquisition knowledge,
- develop knowledge,
- preserving knowledge,
- sharing knowledge,
- using knowledge.

The above analysis shows that the available LCMS system components can support knowledge management functions. As a result, knowledge management may be exercised by the systems to create e-learning courses, a mentor will be able to deal with the domain expertise in educational materials.

## V. Conclusion

Construction of knowledge and placing it in a repository that will be available to designers of e-learning will increase the quality of available educational materials. The disadvantages of data content in materials can be eliminated through a change in approach to lectures created. But it is not only the substantive dimension which must be taken into account. Although it is the most important. You should also pay attention to the efficiency of particular emphasis on funding. Stephen Downes in his work made a very simple economic calculations. These calculations should be taken into account by the designers of courses and organizations that are involved in creating e-learning courses.

These calculations are very simplistic but illustrate the scale of the economic problem. If you prepare one lesson which costs $ 100 and it will be used by 100 universities then the cost of this unit will be $ 1. If, however, the same teaching unit will be produced for each institution individually, the combined cost will be 10 000 $.

Financial calculations, however, should be in the background. Nothing can replace well-designed educational material with current professional knowledge. Recipients quickly assess the value of training that will be fully up to date, discussed the current problems of reality [9].

Managerial knowledge management is very difficult. Its variability over time is the cause of failure of e-learning in the process of transferring knowledge to the recipient. This can be eliminated by a proper design of such a course. The inclusion of knowledge management functions will create the course, which will be able to quickly incorporate the changes in the domain knowledge from the environment.

The use of a process approach to managerial knowledge management can contribute to the improvement of the course prepared for the recipient. Taking into account this approach with the capabilities of the system Learning Content Management Systems - LCMS implemented the tools to build the course which will help to improve the quality of substantial exchange of managerial knowledge.

## References

[1] Clarke A., 2007, E-learning nauka na odległość, Klebanowski M. (tłum), Wydawnictwo Komunikacji i Łączności, Warszawa.

[2] Dąbrowski M., Zając M.(red.), 2013: Rola e-edukacji w rozwoju kształcenia akademickiego, FPAKE, Warszawa.

[3] Hauke K., Owoc M.L., Schreurs J., Theunissen M., 2000, A multimedia Warehouse Supporting on line learning via Internet, Antwerp, Belgium.

[4] Hauke K., Owoc M.L., Gładysz T., 2001, Management of the Multimedia LearnigSpace, rozdział w Knowledge Acquisition and distributed learning in resolving managerial issues, red: Baborski A., Bonner R., Owoc M.L., Malardalen University (Sweden).

[5] Horton W., 2001: Designing knowledge object - Crafting reusable component for teaching, communicating, and entertaining, William Horton Consulting.

[6] Neven F., Duval E., 2002, Reusable learning objects: a survey of LOM-based repositories, Proceedings of the tenth ACM international conference on Multimedia.

[7] Nycz M. (red.), 2004, Generowanie wiedzy dla przedsiębiorstwa – metody I techniki, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

[8] Hauke K., 2014, Analiza wybranych polskich portali wiedzy wspomagających proces prowadzenia działalności gospodarczej, monografia: Wiedza w kreowaniu przedsiębiorczości, red: Kazimierz Perechuda, Iwona Chomiak- Orsa, Politechnika Częstochowska Wydział Zarządzania, Częstochowa.

[9] Hauke K., 2015, Zarys koncepcji zarządzania obiektami wiedzy menedżerskiej w systemach nauczania na odległość., monografia: Wykorzystanie potencjału współczesnych technologii informacyjnych w zarządzaniu organizacjami, red: Leszek Kiełtyka, Waldemar Jędrzejczyk, Wydawnictwo Politechniki Częstochowskiej, Częstochowa.

[10] Hauke K., Obiekty wiedzy w procesie nauczania na odległość – wybrane problemy, monografia: Informatyka 2 Przyszłości, Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, Warszawa.

[11] Probst G., Raub S., Romhardt K., 2000, Managing Knowledge. Building Blocks for Success, John Wiley & Sons.

[12] Internet portal: http://www.cedma.org/ [2015.05.05].

[13] Institute Electronic & Electrical Engineers Learning Technology Sub Committee (IEEE LTSC) L'Allier J., Frame of reference: NETg's Map to the Products, Their Structure and Core beliefs, http://www.netg.com/research/whitepapers/frameref.asp [2015.05.05].

[14] Internet portal: http://forumakad.pl/archiwum/2005/02/15-za-krazace_obiekty_wiedzy.htm [2015.05.05].

DOI: 10.15439/2015F292
ACSIS, Vol. 5

# Supervised Context Classification Methods for an Industrial Machinery

Mateusz Kalisch
Silesian University of Technology
Institute of Fundamentals of Machinery Design
ul. Konarskiego 18a, 44-100 Gliwice, Poland
Email: mateusz.kalisch@polsl.pl

***Abstract*—The paper describes a method of supervised context classification for an industrial machinery. The main objective of this study is to compare single and ensemble classifiers in order to classify groups of contexts which are based on an operating state of the device. The applied research was conducted with the assumption that only classic and well-practised classification methods would be adopted. The comparison study was carried out using real data recorded from an industrial machinery working underground in a mine in Poland. The achieved results confirm the effectiveness of the proposed approach and also show its limitations.**

## I. Introduction

THE INCREASING complexity of recent industrial objects causes that fault diagnosis is one of the most important directions of research in the fields of robotics and modern automatic controls [1], [2], [3]. There are a lot of areas where technical systems and processes are required to be safely and reliably operated, such as aircraft, spaceships, automotive or the mining industry. A large majority of the methods implemented for fault detection and isolation are based on simple approaches [4], because these are easy to implement and fast, but the final result can be unsatisfactory because of limitations e.g. too slow system reactions. More complex solutions can be used to achieve better results for more difficult cases but it can be impossible for even an expert to build this kind of the system. One solution is to create a fault detection and isolation system based on the classifiers which are used to prepare and classify datasets, but it is difficult to extract real data fragments connected with a faulty state, and as a consequence, the training data are not good enough for the classifiers. Another solution is context based reasoning [5]. A system based on this approach can be focused on the list of the contexts which are connected with e.g. working conditions of an examined device. Simpler models of the classifiers and more efficient results of the fault detection and isolation process can be considered as advantages of the system based on the context. However there are some problems connected with this approach like when and what kind of context occurs in a specific period of time and how to use context based approach in the fault detection and isolation process.

The rest of the paper is organized as follows. In Section 2 the context based approach with regards to machine learning is described. The next section includes a detailed description of the proposed method. In particular, there are investigations of the classification methods. Section 4 contains a case study description and the more interesting results of the verification experiments. The last section is devoted to concluding remarks and suggestions for future work.

## II. Context in machine learning

In a classification task, it is possible to distinguish three types of features: primary, contextual and irrelevant [6]. Primary features are useful for classification, without regard to the other features. The irrelevant features are not useful for classification, either when combined with the other features or when they are considered alone. Contextual features cannot be used directly by a classifier, but can be useful when they are combined with other features. The primary features can be also divided into context-sensitive and context-insensitive features. In the case of a machine diagnosis, the context variable could be connected with a number of factors e.g. weather conditions. In another paper [5], the author used contextual variables such as humidity, barometric pressure and external temperature for a gas turbine engine diagnosis. Speech recognition is another example of an area which can use contextual features to improve the efficiency of classification [7]. A speaker's sex, nationality or age may have a strong influence on the relevance of various features but without the primary features the contextual features are useless for these methods. Another type of contextual variable is unknown context which can be identified from data by means of a method based on an evolutionary algorithm [8], [9]. In the final implementation, the context can be acquired directly from the data base or distinguished from the data by the classifier.

The contextual variable is a continuous or discrete variable connected with a specific object. In the case of a discrete contextual variable, the contextual value is equal to one of all the available contextual variants describing this variable. The contextual variant can be obtained from a continuous contextual variable by using the classifier. In the case of an expert system, the context may be connected with text information, where the first part of the message is connected to the contextual variable and the object related to this variable. The second part of the message is connected to the contextual variant. An example of a contextual message could be the contextual variable which refers to wind velocity. The first part of the message connected with the variable might be *Wind*

978-83-60810-66-8/$25.00©2015, IEEE

*velocity is* and the second part (connected with the variant) might be *too low* or *too high*.

In the literature some of the concepts for the usage of the context with machine learning algorithms are described [10], [11]. Peter Turney in [6], [5] described five strategies which show how context can be used: *Contextual normalization*, *Contextual expansion*, *Contextual classifier selection*, *Contextual classification adjustment* and *Contextual weighting*. The extraction of the context during a reasoning process is one part of the context based method of the classification. The context can be available as an additional variable in the dataset or can be hidden in the data. In the second approach, it is necessary to implement an algorithm which can extract this context from the data in the dataset. The context can occur as single context but also as group of contexts, where each context from the group can also occur independently.

## III. Methods of context classification

In the next part of the article, the author describes two methods of classification for groups of contexts. Each context can be described as a binary variable whose value is equal to 0 or 1. All contexts can be connected together and be presented as a decimal value obtained from the binary representation of all contexts (e.g. six binary contexts connected together can be presented as a binary value 010010, which is equal to 16 in decimal notation). It is possible to define a list of all the available combinations of contexts in the group and to create a list of all possible decimal values. This approach (Figure 1) lets us create only one multi-class classifier and the final result of the classification can be decoded to a binary representation to see the state of each binary context in the group.

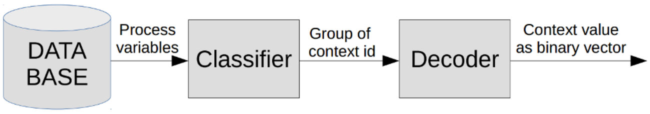


Fig. 1. A scheme for context classification using a single classifier

The advantage of the first method is that the result of the classification can be connected with only one of the known combination of the contexts, because all possible combinations are defined in the training data. The second method (Figure 2) is based on a bank of the binary classifiers where each of them is trained to detect different contexts. When six contexts are available during the reasoning process, it is necessary to implement six binary classifiers in the scheme.

In both methods, each context in the group can be used independently for fault detection and an isolation system. However in the second method, there is a possibility of reaching a result which is not correct. A more detailed description of this problem is presented in the next section of the article.

### A. Used classifiers

In this paper the author compares four different classifiers based on various approaches: *Bayesian Network*, *Naive Bayes*, *Decision Tree* and *Artificial Neural Network*. Each of these classifiers returns a label for a chosen class and a degree of belief for all predicted classes. The best result occurs when one of the classes is characterised by the belief level equal to 1 and the rest of them are equal to 0. This gives us a 100% certainty that a new element should be classified as belonging to this particular class. In the next subsections, more precise descriptions of the selected methods are given.

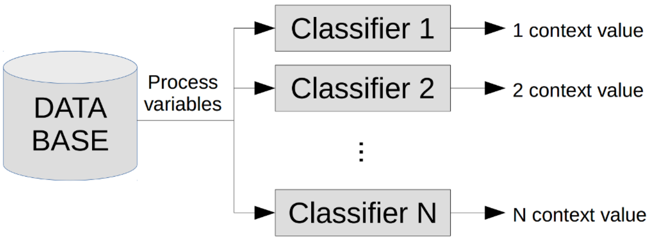


Fig. 2. A scheme of context classification using group of binary classifiers

*1) Bayesian Network:* Bayesian Network, also called *Belief Network* or *Casual Network*, is a graphical model for representing the conditional independences between a set of random variables. Each node in the network represents a variable [12], [13]. Each connection between the nodes is represented by the Bayesian equation 1 where $P(d_i|V_1, \cdots, V_n)$ is often known as the posterior probability of $d_i$ given $V_1, \cdots, V_n$. $P(V_1, \cdots, V_n|d_i)P(d_i)$ is often referred to as the likelihood of $d_i$ given $V_1, \cdots, V_n$, $P(d_i)$ is the prior or marginal probability of $d_i$ and $P(V_1, \cdots, V_n)$ is a normalizing term.

$$P(d_i|V_1, \cdots, V_n) = \frac{P(V_1, \cdots, V_n|d_i)P(d_i)}{P(V_1, \cdots, V_n)} \quad (1)$$

*2) Naive Bayes:* Naive Bayes is a simple probabilistic classification method which is based on Bayesian theory. However, the Naive Bayes classifier considers each of the existing features independently. Taking into account this assumption, the Bayesian equation (1) can be transformed to (2), where the denominator of the equation is replaced by a constant $C$ and the conditional probability is calculated by the multiplication.

$$P(d_i|V_1, \cdots, V_n) = C \cdot P(V_1|d_i) \cdot ... \cdot P(V_n|d_i) \cdot P(d_i) \quad (2)$$

The degrees of beliefs for the classification results are equal to the probability values obtained from the Bayesian equation.

*3) Decision Tree:* A Decision Tree is a classifier based on a tree-like graph created by nodes and the connections between them, where each end node is called a *leaf* and the rest of the nodes have conditions. The result of a decision tree application depends on a chosen leaf. In the algorithm, different split evaluation criteria (e.g. ratio gain in C4.5; information gain in ID3; the Gini impurity measure in CART; etc.) can be used [14], [15]. The confidence levels for the classification results are calculated separately for all leaves of the tree during the learning process. Sometimes, when the learning data is very complex, the results of the decision tree may be uncertain since some of the leaves may be connected to more than one class. The class which is described by more elements then others (in a specific leaf) is chosen as the main class for this leaf. The ratio between the number of elements for available classes is used to calculate the of probability for each class for the leaf.

*4) Artificial Neural Network:* The classifier is a feedforward neural model in which multiple layers of neurons with nonlinear activation functions allow the network to learn nonlinear or linear relationships between input and output vectors [16]. In this paper, a multiple-layer network is used which consists of three layers including $n^1$ neurons in the input layer, with $n^2$ and $n^3$ neurons in the first and the second hidden layers, respectively. In this case, the neural computation can be represented by the following equation:

$$y = \mathbf{LW}^3\mathbf{f}^2\left(\mathbf{LW}^2\mathbf{f}^1\left(\mathbf{LW}^1\mathbf{u}+\mathbf{b}^1\right)+\mathbf{b}^2\right)+b^3 \qquad (3)$$

where $\mathbf{LW}^{\{1,2,3\}}$ correspond to the weight matrices of the input layer and the first/second hidden layer, $\mathbf{b}^{\{1,2,3\}}$ are vectors of the biases, $\mathbf{u}$ is the input signal, and $\mathbf{f}^{\{1,2\}}$ are nonlinear transform operators consisting of tangensoidal activation functions.

## IV. Case Study

A longwall shearer working in a coal mine in Poland is the subject of this study. Longwall mining is underground mining where a long wall of materials is removed in a single slice. The longwall mining method extracts ore along a straight front having a large longitudinal extension. The mining technology involves a longwall shearer, a machine 15 metres long, and weighting 100 tonnes, that has picks attached to two drums which rotate atat a speed of $30{-}40 rev/min$. A longwall face is the mined area from which the materials are extracted. The shearer removes coal by traversing the face at approximately 25 minutes intervals. Traditionally, longwall mining equipment is controlled manually, and the face is aligned in a straight line [17], [18].

### A. Data analysis

Available datasets consist of 36 signals including values of the currents, oil and water pressures, temperatures and rotational speeds of the left and right drums of the longwall shearer. Redundant signals were removed from the dataset after a statistical analysis and the final number of signals was reduced to 21. One of the variables was the operational state which contained information about the current state of the longwall shearer. Information for this variable is represented as binary value and each bit is connected with a specific state. The available dataset covering a few days was divided into the smaller datasets connected with single days. In each dataset the author calculated the number of empty rows and the rows containing data. The results of this calculation are presented in Table I.

TABLE I
Relation between the number of rows containing data and the size of the full dataset

| Date | Rows containing data | Total number of rows | Ratio |
|---|---|---|---|
| 18th October | 64783 | 106385 | 61% |
| 19th October | 72337 | 102470 | 71% |
| 20th October | 79719 | 105248 | 76% |
| 21st October | 77200 | 103746 | 74% |
| 22nd October | 44746 | 67080 | 66% |

It can be seen that all the datasets contained empty rows (with no values) and the size of these gaps was between 24% and 39% of all data.

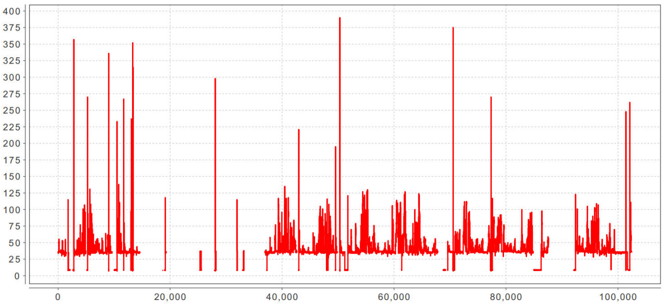


Fig. 3. Average current value of the drive engine

Figure 3 shows that gaps are placed in different fragments of the dataset and the lengths of the fragments of empty data and those filled with data are various. For the dataset from 19th October it is possible to distinguish 28 fragments filled the with the continuous data. Table II presents how many of these fragments filled with data lasted for specific periods of time.

TABLE II
Relation between the number of fragments of useful continuous data and their duration

| Duration | Number of fragments |
|---|---|
| 0 - 1 min | 2 |
| 1 - 10 min | 12 |
| 10 - 60 min | 7 |
| 1 - 3 h | 7 |

The author considered only the 7 datasets with the largest number of the samples. They contained in sequence 5031, 5362, 6461, 7351, 7680, 9937 and 10998 samples, so the duration of the series was between one hour and seven minutes and about two hours and thirty minutes. The higher number of the samples can delivered more samples connected with each operational state, providing more opportunities for the classifiers trained on this data to work properly.

### B. Operating states of the considered device

In the article operating states of the longwall shearer are considered as contexts described in the first part of the article. There are six operating states represented by binary value (0 or 1):

1) Breakdown,
2) Warning,
3) Operation of drives,
4) Drives turned off,
5) Drive to the left,
6) Drive to the right.

In the study presented in this article the operating states of the longwall shearer were recorded in the dataset by a monitoring system. Sometimes this kind of information is not recorded in the data base and it is necessary to discover it or to add additional information as defined by an expert. The operating state is available in the dataset as a decimal value and it is important to convert it to a binary representation to extract information about each operating state. Table III presents the list of the considered states and the possible combinations of them. The first row shows lists the states, where each state can be equal to 0 or 1. The first column presents a list of all possible combinations of the states represented by decimal values (4, 5, 8, 20, 36, 38) and their binary representations are presented in the central area of the table. The numerical values in the first row of the table correspond to the labels of the operating states in the list presented above.

TABLE III
BINARY REPRESENTATION OF ALL CONSIDERED COMBINATIONS OF THE OPERATING STATES

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 4 | 0 | 0 | 1 | 0 | 0 | 0 |
| 5 | 1 | 0 | 1 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 1 | 0 | 0 |
| 20 | 0 | 0 | 1 | 0 | 1 | 0 |
| 36 | 0 | 0 | 1 | 0 | 0 | 1 |
| 38 | 0 | 1 | 1 | 0 | 0 | 1 |

Some of the combinations of the states are not correct and they cannot be considered as possible combinations of the states, e.g. it is not possible to set the bit numbers 5 and 6 to 1 at the same time, because bit 5 is connected with the task *Drive to the left* and bit 6 is connected with the task *Drive to the right*. As it is impossible to move the machine in both directions at the same time, but it is possible to stop the machine, then bit 5 and 6 are equal to 0.

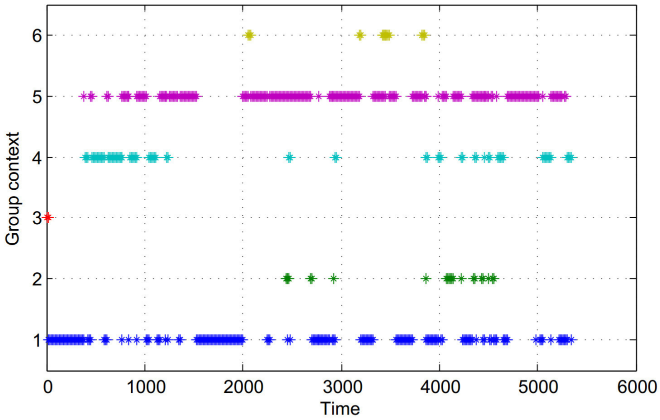


Fig. 4. Occurrence of possible groups of operating states in one dataset

Figure 4 presents the occurrence of the context groups in the fragment of the dataset where each context group id is connected with the following combinations of states:

1) *Operation of drives*,
2) *Breakdown* and *Operation of drives*,
3) *Drives turned off*,
4) *Operation of drives* and *Drive to the left*,
5) *Operation of drives* and *Drive to the right*,
6) *Warning*, *Operation of drives* and *Drive to the right*.

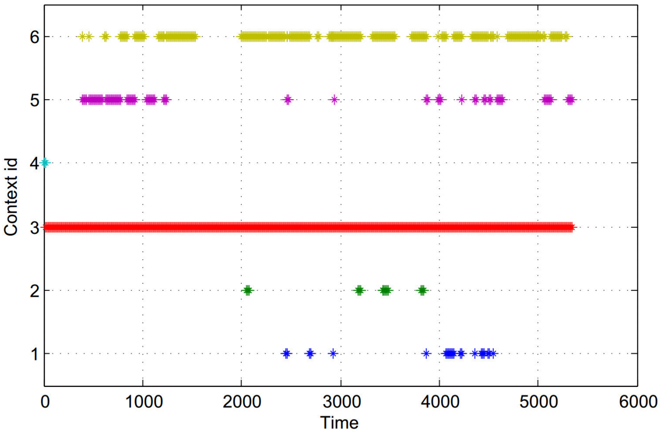


Fig. 5. Occurrence of considered operating states in the fragment of the dataset

Figure 5 shows the places were the specific states occurred in one of the considered datasets. It can be seen that the number of rows of data connected with each state is very various. ID values presented on the Y axis are connected with the list of considered operating states presented at the beginning of this section.

### C. Results

The author used seven different parts of datasets, all of them recorded on the 19th of October. The author compared the data using four various classifiers and two methods of classification. Each classifier in the first method (Figure 1) and all classifiers in the second method (Figure 2) were trained on one dataset and tested by the rest of them. Two measurements were considered to evaluate the effectiveness of the classification: accuracy and recall. Accuracy was the basic evaluation method of classification but its result may be not fully reliable in the cases where the data was not well balanced. The second measurement was used to reduce the influence of various numbers of states in the considered datasets.

To keep all results fully consistent, the final result of the second method (Figure 1) was considered correct only if all results of the binary classifiers (each classifier is connected with different state) were correct. Even if only one classifier made a mistake, the final result was treated as incorrect. This solution is fully comparable with first method, where the final result is presented as a group of the contexts.

Table IV shows the accuracy of all classifiers used in the two considered methods. The classifiers are presented in the table by their short names (DT - Decision Tree; NB - Naive Bayes; NN - Neural Network; BN - Bayesian Network). Each column is connected with a different training dataset and the values in the cells of the table show the average value of the test cases. It is clear that the bank of binary classifiers reached a much better result than the single multi-class classifier.

TABLE IV
ACCURACY RESULTS FOR ALL CLASSIFIERS AND METHODS

| Single classifier | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dataset id | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| DT | 78.1 | 42.8 | 68.9 | 57.7 | 63.6 | 76.6 | 65.4 |
| NB | 30.4 | 45.3 | 55.2 | 48.2 | 72.0 | 56.7 | 37.6 |
| NN | 51.4 | 53.6 | 70, 2 | 65.2 | 75.6 | 75.5 | 51.3 |
| BN | 51.9 | 27.6 | 33.1 | 37.3 | 46.0 | 48.5 | 51.5 |

| Bank of binary classifiers | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dataset id | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| DT | 89.4 | 86.8 | 89.9 | 90.8 | 91.0 | 88.5 | 89.4 |
| NB | 76.9 | 80.1 | 87.2 | 79.8 | 88.3 | 85.9 | 83.3 |
| NN | 83.3 | 86.7 | 92.6 | 82.9 | 92.4 | 92.1 | 88.8 |
| BN | 77.9 | 60.0 | 58.2 | 75.7 | 80.0 | 70.4 | 78.0 |

Accuracy is based on a ratio between all correctly classified rows and the number of all rows in the dataset, and it does not take into account the distinctness of each class in the testing dataset. The result based on the accuracy can not be used only as a measurement of the efficiency of the classification because of the unbalanced test datasets. The second measurement used in this test is mean recall. Recall was calculated for each class separately. The recall value is presented as a ratio of the number of all rows of data with a correctly predicted class to all data rows connected with a specific class. The final result is obtained as the average value of all recall values calculated for all available classes.

TABLE V
MEAN RECALL VALUES FOR ALL CLASSIFIERS AND METHODS

| Single classifier | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dataset id | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| DT | 68.3 | 46.6 | 57.7 | 51.3 | 55.3 | 64.9 | 61.1 |
| NB | 27.0 | 29.7 | 35.8 | 29.2 | 46.8 | 35.3 | 32.1 |
| NN | 36.1 | 31.6 | 55.1 | 36.6 | 47.2 | 37.9 | 44.5 |
| BN | 35.8 | 27.4 | 27.9 | 35.8 | 37.9 | 38.9 | 43.0 |

| Bank of binary classifiers | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dataset id | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| DT | 82.4 | 73.2 | 79.8 | 84.3 | 79.4 | 81.8 | 82.5 |
| NB | 58.1 | 61.1 | 63.9 | 59.0 | 65.9 | 64.1 | 58.2 |
| NN | 64.6 | 65.9 | 64.2 | 60.1 | 73.4 | 57.2 | 55.3 |
| BN | 65.2 | 53.1 | 60.1 | 66.9 | 66.8 | 67.8 | 66.1 |

The method based on the bank of binary classifiers reached much better results than the single classifiers. It can be seen that the values for mean recall are worse than the results for accuracy. This proves that the classifiers trained on unbalanced datasets were not evaluated properly by the accuracy measurement. The results for accuracy show that only two classifiers in each column reached the best results interchangeably: Decision Tree and Neural Network. The rest of the classifiers almost always reached worse results than the two mentioned above. The mean recall value shows that the algorithm based on a Decision Tree is able to work more properly with the unbalanced data, and in all columns it reached the best result. The classifier based on a Neural Network had a tendency to ignore classes with smaller numbers of examples.

TABLE VI
ACCURACY OF CLASSIFICATION OF EACH STATE OBTAINED BY DECISION TREE

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 99.81 | 100.00 | 100.00 | 100.00 | 98.69 | 96.91 |
| 2 | 90.81 | 98.59 | 99.94 | 99.94 | 65.50 | 73.37 |
| 3 | 91.50 | 95.75 | 99.97 | 99.97 | 48.45 | 77.60 |
| 4 | 96.87 | 98.67 | 99.95 | 99.95 | 52.84 | 83.17 |
| 5 | 97.04 | 98.59 | 99.71 | 99.71 | 71.09 | 79.44 |
| 6 | 99.42 | 98.16 | 100.00 | 100.00 | 30.14 | 81.99 |
| 7 | 96.82 | 98.89 | 98.63 | 98.63 | 65.01 | 75.22 |

Table VI shows the accuracy result for the classification of each operating state (columns 1 to 6) for all available datasets (rows 1 to 7) by the bank of the binary classifiers based on a Decision Tree. Each value shows the result of one classifier from the bank, e.g. the value in the third row and second column (89.67) presents the primary accuracy result obtained by the binary classifier (based on a Decision Tree) whose task was to distinguish the operating state called *Warning* (the label of column 2). The label of each row indicates the dataset which was used during the verification test. It can be seen that the accuracy value for the first four states is high but for states 5 and 6 it is lower (except for the first row, because the classifier used in this example was trained by the first dataset).

TABLE VII
RECALL OF THE CLASSIFICATION OF EACH STATE OBTAINED BY A DECISION TREE

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 97.42 | 100.00 | 100.00 | 100.00 | 95.98 | 96.72 |
| 2 | 73.59 | 99.29 | 97.92 | 97.92 | 76.02 | 72.10 |
| 3 | 50.40 | 89.67 | 99.99 | 99.99 | 62.51 | 77.00 |
| 4 | 56.08 | 94.92 | 93.18 | 93.18 | 66.33 | 80.30 |
| 5 | 50.50 | 83.83 | 95.15 | 95.15 | 77.75 | 76.95 |
| 6 | 58.11 | 75.11 | 100.00 | 100.00 | 53.32 | 82.11 |
| 7 | 48.84 | 77.29 | 91.48 | 91.48 | 61.15 | 71.93 |

The results for a recall of the same situation (Table VII) shows which states are more difficult to isolate and which are not. The classifier reached a high level of efficiency for the 3th and 4th state (*Operation of drives* and *Drives turned off*). The efficiency for the 2nd state (*Warning*) was a little bit lower. The classifier had some problems with the identification of the 5th and 6th states (*Drive to the left* and *Drive to the right*) and the reason for this problems could be the lack of clear information about the direction of movement of the longwall shearer in the dataset. There is no signal which could clearly show the direction of the movement. The classifier reached the worst results for the 1st state (*Breakdown*) because the number of examples connected with this state was very small and the classifier was not able to distinguish this state properly in the test dataset.

## V. CONCLUSIONS

It is possible to use different methods of classification to implement basic schemes of context identification. The author was able to increase the efficiency of classification by implementing groups of binary classifiers, instead of using a single multi-class classifier. The final decision of the presented schemes can be used in fault detection and isolation models implemented in an expert system.

The main advantage of the first method (Figure 1) of context classification is its simplicity. It requires only one multi-class classifier, and its result is always connected with a correct combination of states. But the classifier used in this method always reached a worse result than the classifiers used in the second method (Figure 2). Additionally the first scheme needed classified dataset which contained all possible combinations of states and sometimes it is impossible to prepare this kind of training dataset because some of the combinations might not occur in the recorded data. In the second method, it is not necessary to create a dataset with all possible combinations of contexts because each context is classified separately. It is therefore easier to prepare the appropriate training data. The scheme of this method is more complex, but the classification results are significantly more accurate than the results of the first method (Figure 1). Nonetheless, the results of this scheme cannot be fully correct because of the possibility of impossible combinations of contexts as a result of the classification (e.g. the longwall shearer moving in both directions at the same time).

### A. Future work

The next step in future research will be connected with other methods of context classification based on ensemble classifiers and meta-classification. Another step is the implementation of the described and future methods inside a fault detection and isolation system, in order to increase the quality of the system in comparison to a solution working without context. It is necessary to see how strong the influence of the context classifier is on the final result of the diagnosis system, because low efficiency of the context classifier could be a reason for high uncertainty levels of the final decision used in the fault detection and isolation system.

## Acknowledgment

The research presented in the paper was partially financed by the National Centre of Research and Development (Poland) within the frame of the project titled „Zintegrowany, szkieletowy system wspomagania decyzji dla systemów monitorowania procesów, urządzeń i zagrożeń" (in Polish) carried out in the path B of Applied Research Programme - grant No. PBS2/B9/20/2013. The part of the research was also financed from the statutory funds of the Institute of Fundamentals of Machinery Design.

## References

[1] F. Caccavale and L. Villani, *Fault Diagnosis and Fault Tolerance for Mechatronic Systems: Recent Advances*, ser. Springer Tracts in Advanced Robotics. Springer Berlin/Heidelberg, 2003. [Online]. Available: http://dx.doi.org/10.1007/3-540-45737-2

[2] J. M. Kościelny, *Diagnostyka zautomatyzowanych procesów przemysłowych*. Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2001.

[3] R. J. Patton, P. M. Frank, and R. N. Clark, *Issues of Fault Diagnosis for Dynamic Systems*. Springer-Verlag Berlin and Heidelberg, 2000. [Online]. Available: http://dx.doi.org/10.1007/978-1-4471-3644-6

[4] P. F. Odgaard and J. Stoustrup, "Results of a Wind Turbine FDI Competition," in *Proceedings of the 8th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes*, A. Zaragoza, Ed., Aug. 2012, pp. 102–107. [Online]. Available: http://dx.doi.org/10.3182/20120829-3-mx-2028.00015

[5] P. D. Turney, "Exploiting context when learning to classify," in *Proceedings of the European Conference on Machine Learning*, ser. ECML '93. London, UK, UK: Springer-Verlag, 1993. ISBN 3-540-56602-3 pp. 402–407. [Online]. Available: http://dl.acm.org/citation.cfm?id=645323.649588

[6] P. Turney, "The management of context-sensitive features: A review of strategies," in *Proceedings of the ICML-96 Workshop on Learning in Context-Sensitive Domains*, 1996, pp. 60–65. [Online]. Available: http://dx.doi.org/10.1.1.51.3784

[7] G. Widmer, "Tracking context changes through meta-learning," in *Machine Learning*. Kluwer Academic Publisher, 1996, pp. 259–286. [Online]. Available: http://dx.doi.org/10.1023/A:1007365809034

[8] A. Timofiejczuk, "Context based diagnostics of rotating machinery," in *7th International Conference on Acoustical and Vibratory Surveillance Methods and Diagnostic Techniques (Surveillance 7), 29-30 October 2013, Chartres, France*, 2013, pp. 1–9.

[9] ——, "Identification of diagnostic rules with the application of an evolutionary algorithm," *Maintenance and Reliability*, no. 1, pp. 11–15, 2008.

[10] J. A. Jakubczyc, "Contextual classifier ensembles." in *BIS*, ser. Lecture Notes in Computer Science, W. Abramowicz, Ed., vol. 4439. Springer, 2007, pp. 562–569. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72035-5_44

[11] D. Munoz, J. A. Bagnell, N. Vandapel, and M. Hebert, "Contextual classification with functional max-margin markov networks." in *CVPR*. IEEE, 2009, pp. 975–982. [Online]. Available: http://dx.doi.org/10.1109/CVPRW.2009.5206590

[12] Z. Ghahramani, "Hidden markov models." River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2002, ch. An Introduction to Hidden Markov Models and Bayesian Networks, pp. 9–42.

[13] R. Daly, Q. Shen, and J. S. Aitken, "Learning bayesian networks: approaches and issues." *Knowledge Eng. Review*, vol. 26, no. 2, pp. 99–157, 2011. [Online]. Available: http://dx.doi.org/10.1017/S0269888910000251

[14] A. Lile, "Analyzing e-learning systems using educational data mining techniques," *Mediterranean Journal of Social Sciences*, vol. 2, no. 3, pp. 403–419, 2011. [Online]. Available: http://dx.doi.org/10.5901/mjss.2011.v2n3p403

[15] F. Akthar and C. Hahne, *RapidMiner 5, Operator Reference*. www.rapid-i.com, 2012.

[16] S. Haykin, *Neural Networks: A Comprehensive Foundation*, ser. 2nd edition. Prentice Hall International, 1999.

[17] A. Oksan Altun, I. Yilmaz, and M. Yildirim, "A short review on the surficial impacts of underground mining." *Scientific Research and Essays*, vol. 21, no. 5, pp. 3206–3212, 2010.

[18] G. A. Einicke, J. C. Ralston, C. O. Hargrave, D. C. Reid, and D. W. Hainsworth, "Longwall mining automation an application of minimum-variance smoothing." *IEEE Control Systems*, vol. 28, no. 6, pp. 28–37, 2008. [Online]. Available: http://dx.doi.org/10.1109/MCS.2008.929281

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1673–1681

DOI: 10.15439/2015F26
ACSIS, Vol. 5

# Faults diagnosis using self-organizing maps: A case study on the DAMADICS benchmark problem

Andrzej Katunin, Marcin Amarowicz, Paweł Chrzanowski
Silesian University of Technology,
Institute of Fundamentals of Machinery Design
Konarskiego 18A Street, 44-100 Gliwice, Poland
Email: {andrzej.katunin,marcin.amarowicz,pawel.chrzanowski}@polsl.pl

***Abstract*—This paper deals with a method of faults detection and identification based on the clusterization of the multiple diagnostic signals. Various types of faults and character of their occurrence were simulated using DAMADICS Benchmark Process Control System. A great advantage of the applied approach based on self-organizing (Kohonen) maps is that even the smallest differences in signals allow for detection, isolation and identification of type of occurred faults with respect to the healthy condition of the investigated system based on the unsupervised learning. It was shown that in some cases the faults, which are undetectable during monitoring of simple heuristic and statistical parameters and other previously applied methods, are recognizable when the approach based on self-organizing maps is applied. The case studies presented in this paper show the faults detection procedure as well as clusterization of types and successful classification of almost all the unique faulty states of the investigated system.**

## I. Introduction

An increasing usage of control and automation systems in industrial applications influences on development of novel methods of fault diagnosis, particularly their detection, isolation and identification. Moreover, the industrial requirements for such procedures are to be sensitive even for the early stage of development of faults and should ensure the implementation of them in on-line monitoring systems. The specificity of the physical nature of some faults requires the application of new methods in order to detect and classify them with the lowest possible false alarms.

Recent studies in the area of the fault diagnosis are based on many different approaches. They can be classified, in general, into the quantitative and qualitative approaches. The most simple quantitative approach is an application of statistical measures to the diagnostic signals, observation and inference about a fault presence basing on these measures [1]. However, there is a large amount of research done with application of qualitative fault diagnosis, primarily using soft computing methods. Several authors used such methods in the fault diagnosis, in particular in [2], [3] the genetic programming for observer-based fault diagnosis and evolutionary learning of fuzzy models were used, while the authors of [4] used a fuzzy qualitative reasoning approach for the fault detection problem. A lot of studies are concerned with artificial neural networks (ANN) application to such a class of problems. In [5] the authors applied the group method of data handling (GMDH) neural networks, while in [6], [7] the authors used a neuro-fuzzy approach and ANN with decision trees, in [8] the hidden Markov model for the fault diagnosis was applied.

One of the novel and promising diagnostic tools is the group of approaches based on Kohonen self-organizing maps (SOM). SOM are very often used in problems of the analysis of large data structures e.g. in the problems of clustering or classification [9], [10], [11], [12], image processing [13], [14], [15], robotics [16], [17], time series forecasting [18], [19], [20] and faults detection and identification [21], [22], [23].

In the presented paper the authors analyzed and discussed the overall analysis of all faults using SOM-based approach, which diagnostic signals are able to be generated using the DAMADICS benchmark. The only previous study [24], which was carried out on DAMADICS data using the SOM-based approach, presented a possibility of application of SOM to the faults clusterization, but the results are limited to the analysis of three selected faults. In the present study the process signals were simulated basing on real measurements of diagnostic signals on the actuator system during the evaporisation process in the Lublin sugar factory (Poland) [25], [26]. In order to detect, isolate and identify the faults simulated using DAMADICS benchmark actuator system the SOM-based approach was applied.

## II. Description of the problem and the methodology

### A. The DAMADICS benchmark

The investigated problem of the faults detection and identification was based on sets of diagnostic signals, generated using the DAMADICS benchmark actuator system, which simulated various types of possible faults. The scheme of the investigated system is presented in Fig. 1 [27]. The actuator system consists of the pneumatic servo-motor $S$, a control valve $V$ and a positioner $P$. These three parts of a system are composed by a set of measured diagnostic signals: external controller output $CV$, flow sensor measurement $F$, input $P1$ and output $P2$ valve pressure, medium temperature $T$ and the rod displacement $X$.

Using the DABLib Simulink®-Matlab® library 19 fault types with a variable number of fault intensities (three stages

The publication is financed from the statutory funds of the Faculty of Mechanical Engineering of the Silesian University of Technology.

978-83-60810-66-8/$25.00©2015, IEEE

TABLE I: Types of faults simulated in the DAMADICS benchmark

| Symbol | Fault |
|---|---|
| $f1$ | Valve clogging |
| $f2$ | Valve sedimentation |
| $f3$ | Valve erosion |
| $f4$ | Increased of valve or bushing friction |
| $f5$ | External leakage |
| $f6$ | Internal leakage |
| $f7$ | Medium evaporation or critical flow |
| $f8$ | Twisted servo-motor's piston rod |
| $f9$ | Servo-motor's housing or terminals tightness |
| $f10$ | Servo-motor's diaphragm perforation |
| $f11$ | Servo-motor's spring fault |
| $f12$ | Electro-pneumatic transducer fault |
| $f13$ | Rod displacement sensor fault |
| $f14$ | Pressure sensor fault |
| $f15$ | Positioner supply pressure drop |
| $f16$ | Increase of pressure on valve inlet |
| $f17$ | Pressure drop on valve inlet |
| $f18$ | Fully or partly opened bypass valves |
| $f19$ | Flow rate sensor fault |

of abrupt faults and the incipient fault) can be modelled, which gives total of 45 cases including the case with no faults occurrence. One should consider that not every fault simulation has physical foundation of existence, thus for some types of faults a limited number of intensity subcases existed. The complimentary list of faults that can be modelled and which were used in further analyses was presented with a description in Table I and Table II [27].

TABLE II: Faults able for simulation in the DAMADICS benchmark

| No. | Small* | Medium* | Large* | Incipient |
|---|---|---|---|---|
| $f1$ | X | X | X | |
| $f2$ | | | X | X |
| $f3$ | | | | X |
| $f4$ | | | | X |
| $f5$ | | | | X |
| $f6$ | | | | X |
| $f7$ | X | X | X | |
| $f8$ | X | X | X | |
| $f9$ | | | | X |
| $f10$ | X | X | X | |
| $f11$ | | | X | X |
| $f12$ | X | X | X | |
| $f13$ | X | X | X | X |
| $f14$ | X | X | X | |
| $f15$ | | | X | |
| $f16$ | X | X | X | |
| $f17$ | | | X | X |
| $f18$ | X | X | X | X |
| $f19$ | X | X | X | |

* Abrupt faults

### *B. Previous studies related to diagnosis using DAMADICS benchmark*

The problem of faults identification based on DAMADICS benchmark has been investigated by several scientific teams. The first studies were performed by the authors of the benchmark using various methods of faults detection, isolation and identification. One can notice the general tendency in these papers where the authors take into consideration only selected cases of faults. The authors of [28] performed the study on faults detection and isolation based on the analysis of residua using binary diagnostic matrices for single sets of signals for a given faults case as well as the pairs and triplets of such cases. The authors of [29] used timed automata approach and considered three selected faults in their identification procedure. The authors of [30] focused on the abrupt large faults available in the DAMADICS benchmark for faults detection using a spectral estimation approach. Other approaches were to consider GMDH neural networks [5], interval observers [31] fuzzy classifiers for fault detection and isolation [32], structural analysis [33] in order to evaluate fault isolability, etc. The only study, which considered the whole set of faults possible to simulate in the DAMADICS benchmark, was performed by the authors of [34]. They used an approach of qualitative reasoning coupled with fuzzy neural networks and basing on their results only two cases were not detected and isolated.

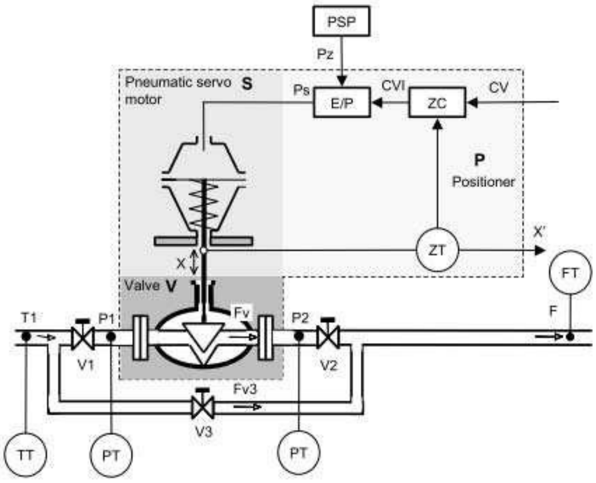


Fig. 1: The scheme of actuator system

The main goal of the studies on the DAMADICS benchmark was to find an appropriate methodology, which allows for detection, isolation and identification of all types of faults available in the benchmark and relations between them, which can be carried out within a single analysis procedure using all signals of all available faulty cases.

### *C. Data preparation*

The authors of [35] stated that the faults $f8$ and $f12$ have theoretical behaviour, i.e. the methods applied by the authors were not able to distinguish these faults from the healthy state. Results obtained by the authors of [34] show that the small type fault $f16$ and incipient type fault $f18$ are not detectable, while the authors of [33] stated that using their method the faults $f9$ and $f16$ are not detectable. The authors of [5] did not detected some types of faults: $f5$, $f8$, $f9$, $f12$ and $f14$ using the GMDH neural network-based approach. In the presented study all of the faulty cases were considered in the analysis. Considering initial analysis of signals generated from the DAMADICS benchmark, the signal $P2$ was removed from every set of tested cases due to the fact that this signal remained insensitive to the faults of a system. Moreover, following the previous studies in the fault detection problem on the DAMADICS benchmark data [4], [35], it is difficult to detect some of the faults due to their weak intensities (i.e. very small changes in the diagnostic signals) and/or very slow development. However, these faults cases were considered in further studies in order to investigate their detectability using the proposed methodology.

The sets of signals of each case listed in Table II were generated using the DABLib library. In each case the fault occurred after the 900th second of the simulation. Following this, the duration of the most of the generated sets of signals was limited to 2000 seconds. In some cases of slowly propagating incipient faults the duration of simulation was extended to 4600 seconds ($f17$) and to 86000 seconds ($f2$, $f3$, $f5$, $f6$, $f9$, $f11$ and $f18$) in order to achieve the full propagation history of a given fault until the fault index reaches the unity value.

During the performed researches four classes of faults: small, medium and large of abrupt and incipient faults, were considered. For the clarity of results presentation the numbers from 1 to 4 were assigned to the classes as the subscripts in the form: $fault_{class}$, e.g. the notation $(f2, f14)_4$ denotes that the faults $f2$ and $f14$ of incipient type, whereas $f3_{2,3}$ denotes that the fault $f3$ of medium and large abrupt types is considered.

### D. Simple statistics-based direct diagnostics

In the diagnostics of processes, the often applied approach for detection and isolation of faults is the direct diagnostics, which is based on simple signal processing and control of constraints. For this purpose the current values of the features of process variables such as mean value, root mean square, maximal and minimal values, variance and others, are determined. In the case of diagnosing of a system basing on the alarm thresholds the analysis is performed on the basis of a comparison of current signal value of a process with the assumed minimal and maximal values, between which the signal can change in the case of non-faulty condition. In the case of statistical features the estimators in the specified window are calculated and then compared with the appropriate values determined for the non-faulty condition.

In the investigated problem both approaches were applied. In the case of control of constraints the evaluation was performed basing on comparison of current values of signals of processes with maximal and minimal values of the signals representing non-faulty condition of a system. In the case of the analysis of statistical features the mean value and a variance were calculated for comparative procedures. For the comparison purpose of current signals with the signals for the non-fault condition the threshold of 5% of variability range was assumed. When the values of signals exceed the threshold the change is recognized as significant one, which indicates the damage. The results of damage detection using the above-described approaches were stored in Tables III, IV, V and VI. Each table refers to the intensity of a damage and the presented values denote the ability of detection of a given fault using the described approaches: 0 for the case of undetectable fault and 1 for the case of detectable fault.

Obtained results show that the detection of the following faults was not possible: small faults $(f10, f14, f16, f8)_1$, medium faults $(f10, f14, f16, f8)_3$, large faults $(f14, f8)_3$ and incipient faults $(f11, f2, f3, f4, f5, f6, f9)_4$. As can be noticed, the application of the described approaches for faults detection is possible only for a limited group of cases. The isolation of faults using these approaches is not possible in direct manner, but could be possible only after application of the algorithm, which allows for distinguish the faults. One of such extensions, which allows for distinguishing the faults, is the binary diagnostic matrix, however in numerous cases distinguishing of the faults was not possible. Therefore, it is necessary to develop appropriate approaches that would be sensitive to changes in the diagnostic signals and allow for their distinguishing and diagnostic inference about faults occurrence.

### E. Motivation

Considering the results of faults detection presented above it can be noticed that the simple statistical methods often remain insensitive to changes in signals and thus the faults could not be detected. In order to detect and identify the faults occurred during the investigated industrial process the more advanced techniques should be applied. The SOM seems to be an effective tool both for the faults detection and identification and has an adequate sensitivity to recognize even the tiny changes in the analysed signals. Moreover, the application of SOM allows for the analysis and comparison of sets of multiple signals, i.e. it is not necessary to preprocess the measurement data in order to find the signal containing information about the occurred fault.

## III. SOM AND METHODOLOGY

### A. Fundamentals on the self-organizing maps

The Kohonen SOM are a type of ANN with unsupervised learning. Two layers, input and output, can be distinguished in the structure of this network (see Fig. 2). The input layer is a vector of neurons, while the output layer is a multidimensional representation of neurons, commonly the two-dimensional (2D) or three-dimensional (3D) one. In the case of 2D representation a rectangular or hexagonal map can be used. A toroidal or cylindrical maps could be used in case of 3D representation. The number of neurons in the input layer is equal to the number of attributes that describe each input data. In case of output layer $5\sqrt{N}$ neurons, where $N$ is the number of input samples, was usually used [36]. But the practical size of this layer depends on nature of input samples. For diversified cases of input data bigger maps should be applied. Each neuron from input layer is connected with each neuron

TABLE III: Results of statistics-based diagnostics for abrupt small faults

| type | $f10_1$ | $f12_1$ | $f13_1$ | $f14_1$ | $f16_1$ | $f18_1$ | $f19_1$ | $f1_1$ | $f7_1$ | $f8_1$ |
|---|---|---|---|---|---|---|---|---|---|---|
| min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| max | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| mean | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| var | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |

TABLE IV: Results of statistics-based diagnostics for abrupt medium faults

| type | $f10_2$ | $f12_2$ | $f13_2$ | $f14_2$ | $f16_2$ | $f18_2$ | $f19_2$ | $f1_2$ | $f7_2$ | $f8_2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| max | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| mean | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| var | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |

TABLE V: Results of statistics-based diagnostics for abrupt large faults

| type | $f10_3$ | $f11_3$ | $f12_3$ | $f13_3$ | $f14_3$ | $f15_3$ | $f16_3$ |
|---|---|---|---|---|---|---|---|
|  | $f17_3$ | $f18_3$ | $f19_3$ | $f1_3$ | $f2_3$ | $f7_3$ | $f8_3$ |
| min | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
|  | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| max | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
|  | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| mean | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
|  | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| var | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
|  | 1 | 1 | 1 | 1 | 1 | 1 | 0 |

TABLE VI: Results of statistics-based diagnostics for incipient faults

| type | $f11_4$ | $f13_4$ | $f17_4$ | $f18_4$ | $f2_4$ | $f3_4$ | $f4_4$ | $f5_4$ | $f6_4$ | $f9_4$ |
|---|---|---|---|---|---|---|---|---|---|---|
| min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| max | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| mean | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| var | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

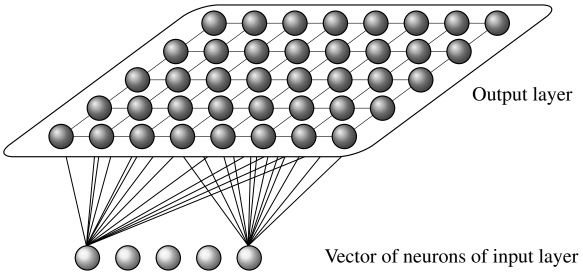


Fig. 2: The structure of the self-organizing map

of output layer [37]. For an arbitrary neuron $n_i$ from output layer it is possible to define some neighbourhood. For this purpose the radius $r$ is used. Based on this radius some area (called a neighbourhood) around considered neuron can be selected. The neighbourhood of the neuron $n_i$ is described by a function $h_{ci}[\bullet] \in [0; 1]_R$ called a neighbourhood's function. The values of this function are distributed from neuron $n_i$ along the radius $r$.

During the learning process of SOM for particular input data the best neuron from output layer is selected. In the literature this neuron is called a winner neuron or best matching unit (BMU). For the purpose of such selection some of the distance metrics e.g. Euclidean, Chebyshev, Manhattan, etc. can be used. In a learning process the weights of chosen neurons are updated according to the formula:

$$w_i(k+1) = w_i(k) + \eta(k) h_{ci}[x(k) - w_i(k)], \tag{1}$$

where $k$ is the number of iterations of the learning process, $\eta(k)$ is the learning coefficient, $h_{ci}$ is the value of the neighbourhood function of the neuron $n_i$ with respect to the neighbourhood of the winner neuron $n_c$ and $x(k)$ is the input vector inputted to the network during $k$-th iteration. Generally the learning process consists of two stages. The first of them is called Winner Takes Most (WTM). During this stage the weights of a winner neuron and its neighbours neurons are updated. Second stage is called Winner Takes All (WTA). In this case the radius $r = 0$, therefore the value of coefficient $h_{ci} = 1$, and only the weights of the winner neuron are updated.

### *B. The methodology*

For the purpose of carrying out the research described in this paper, the SOM Toolbox for Matlab® environment [38] was used. The input data was prepared basing on values of particular signals measures during occurrence of some faults. From each of considered signals (five signals without "$P2$, see the list in Sec. II-A) $N$ consecutive values were considered (e.g. from 100 s to $(100 + N)$ s). Based on these values the vectors of $5N$ elements that describe particular faults were created. Finally, the input matrix has a size of $M \times 5N$, where

$M$ is a number of considered fault cases plus one non-faulty exemplary case. During the conducted study different numbers of considered cases were being taken into consideration. The main settings of the SOM algorithm were as follows:

- type of output layer – 2D hexagonal maps of neurons,
- size of maps – from $10 \times 10$ for the simplest example to $100 \times 100$ for the most complicated examples,
- data normalization – variance was normalized to one,
- training algorithm – batch algorithm,
- number of training (WTM,WTA) iterations – from $(10, 10)$ to $(100, 80)$, depending on complexity of data,
- neighbourhood function – Gaussian function,
- distance measure – Euclidean metric.

For particular cases of input data the $U$-matrix maps were determined. An inspection of the obtained maps and quantization error were used for verification and evaluation of the quality of obtained results. For selected neurons of these maps the labels of considered faults or non-faulty case were assigned. Based on determined in this way maps the relevant conclusions about the possibility of detection, isolation and identification of particular faults were formulated.

In accordance with [39] the fault is detected when the difference in the set of signals with respect to an initial (healthy) state can be observed. As isolation stage we assume that the faults can be distinguished from each other and the identification stage allows us to identify a class and a type of a given fault. In this context we assume that basing on obtained SOM maps the fault is detected, isolated and identified when is located in separate cluster of map according to clusters related to the adequate non-faulty state or other fault cases. According to this assumption in Fig. 3 (subscript $s$ and $b$ in labels of faults denotes the small and big cases of fault, respectively) one can see that all faults are detectable. Faults $f2_s$ and $f2_b$ are detectable, but because they are located in the same cluster there are not isolated. Faults $f8_b$ and $f11_s$ are detectable and isolable as well since they are located in separate clusters. Finally, faults $f1_b$ and $f1_s$ are both detectable, isolable and identifiable. Additionally, based on obtained maps it is possible to conclude that some faults e.g. $f3_b$ could be weakly detectable.

In the most cases 1100 samples of each signal, starting from 901-th second (the onset of fault), were taken into account during the study. For several cases the number of considered samples was enlarged due to slow propagation of some of incipient faults. The analysis was realized following the standard two-step approach used during application of ANN. Firstly, the SOM was trained on standard data obtained from the generator following the description above. Then, three additional sets of data with modified character of noise distribution in signals were generated in order to validate the proposed approach during the testing stage.

## IV. Results

### A. Diagnosing process

At the beginning, all of the cases of faults were considered in order to evaluate their distinguishability. The resulted exemplary maps for this study are shown in Fig. 4. Two groups of faults: weakly and very well detected can be distinguished. The set of weakly detectable faults (the same cluster on the obtained map) contains the following faults: $(f8, f14, f16)_1$, $(f8, f14, f16)_2$, $(f8, f14)_3$, $(f2, f3, f5, f6, f9, f11, f18)_4$. On the other hand, the subset of very well detected faults contains the subsequent items: $f7_1$, $(f1, f7)_2$, $(f1, f2, f7, f15, f16)_3$, $(f13, f17)_4$.

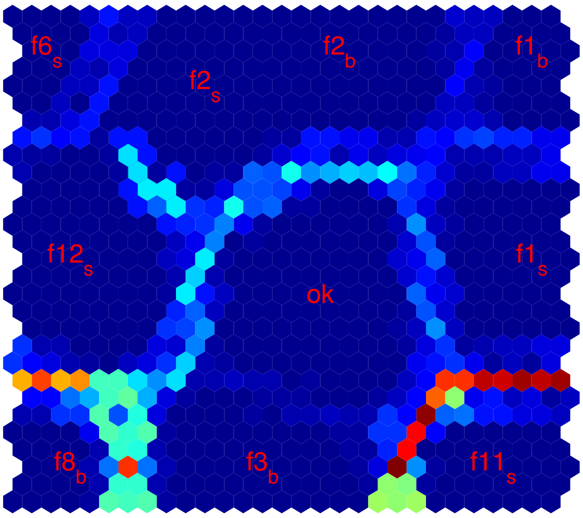


Fig. 3: Exemplary maps with associated faults for selected clusters

In the next stage of the analysis the faults in the context of their particular classes (three types of abrupt and one type of incipient faults) were considered. Examples of obtained maps are shown in Fig. 5. Based on the obtained results two groups of faults can be distinguished. In the first of them the weakly detectable faults such as $(f8, f14)_1$, $(f8, f14)_2$, $(f8, f14)_3$, $(f5, f6, f18)_4$ are assigned. A special case of assignment was obtained for $f14$, for which the same neuron as for non-faulty condition was assigned. Further analysis (search for the differences between the non-faulty and $f14$ signals) allows to conclude that the process signals of $f14$ were identical to the signals from the non-faulty condition of a system. The second group of faults contains the faults which are well detectable while taking into account the individual classes. The faults $(f1, f7, f13, f18)_1$, $(f1, f7, f13, f18)_2$, $(f2, f7, f15)_3$ and $(f4, f13, 17)_4$ can be assigned to this subset. Furthermore, it can be seen that some faults are very similar to each other, e.g. $f13_2$ and $f18_2$ as well as $f13_3$ and $f18_3$.

The consideration of the weakly detectable faults only, such as

$(f2, f3, f5, f6, f9, f11, f18)_4$, $f8_{1,2,3}$, (the faults $f14_{1,2,3}$ are not further considered), which were excluded during the previous analysis, allows to conclude that all of them are very well isolated with respect to the non-faulty case (see Fig. 6). Only the fault $f5_4$ is located in the same cluster as non-faulty case, thus this fault is very similar to this state.

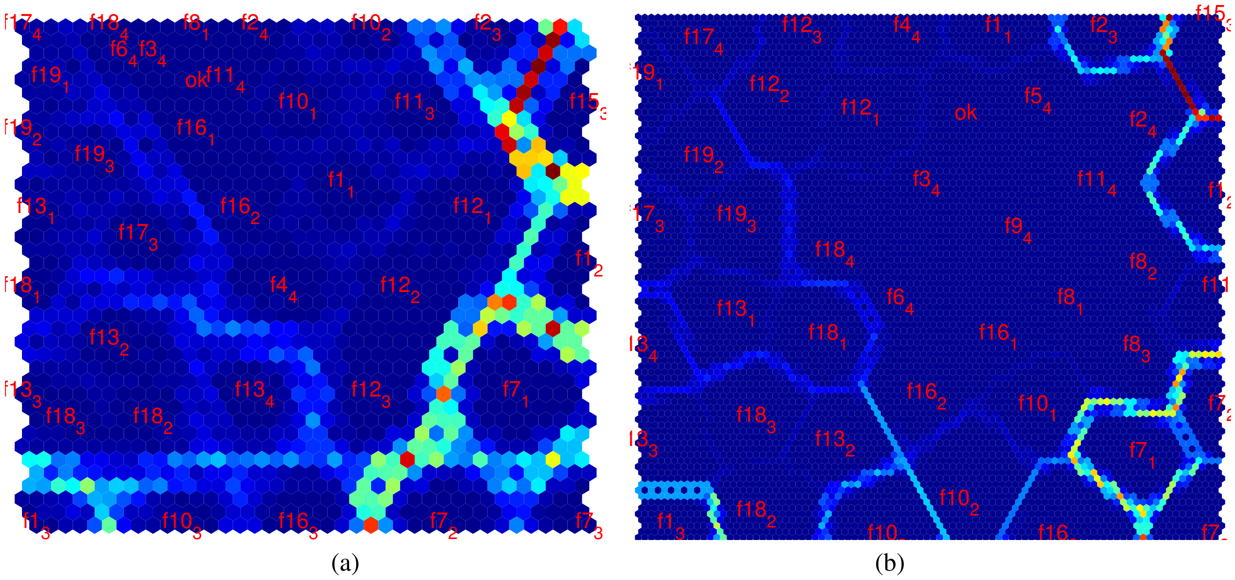


Fig. 4: Results obtained for all considered faults: (a) map size: 20 × 20, learning steps (WTM, WTA): (10, 10); (b) map size: 45 × 45, learning steps (WTM, WTA): (50, 30)

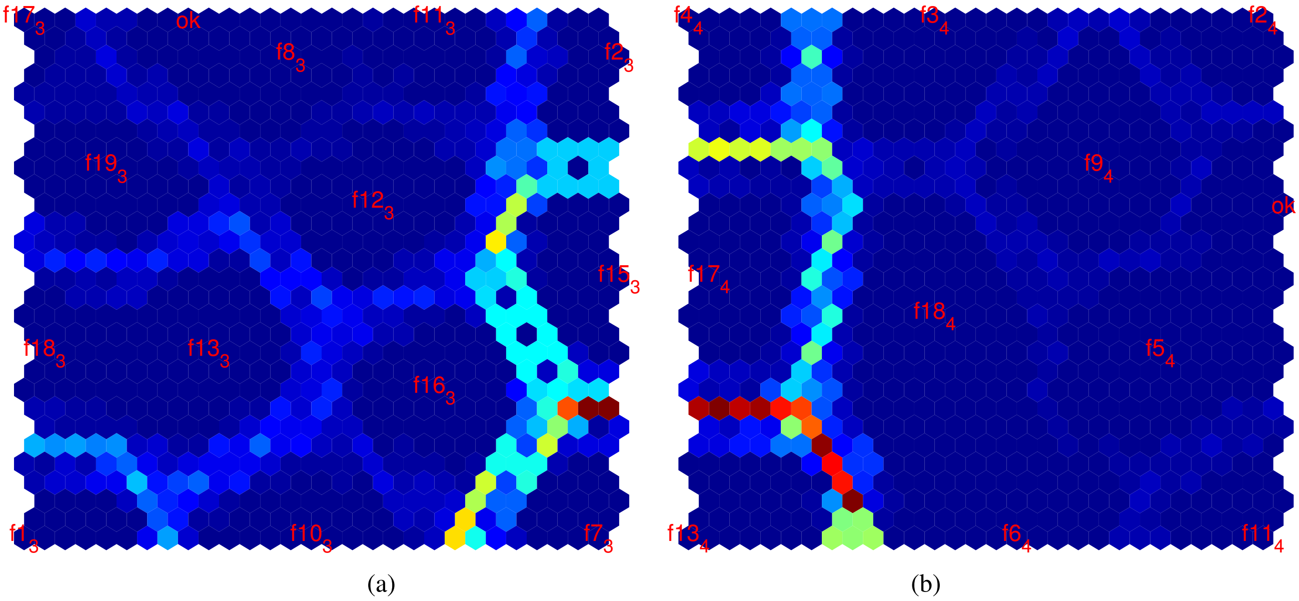


Fig. 5: Maps obtained for single classes of faults, map size: 15x15, learning steps (WTM, WTA): (30, 10): (a) large abrupt faults; (b) incipient faults

Further studies were concentrated on the sensitivity analysis with taking into consideration the shorter realizations of signals for the particular groups of faults. Two main groups: weakly detectable (listed above) and the remaining faults were considered. Exemplary maps are shown in Fig. 7. Based on the obtained results for these analyses, it is possible to conclude that the reduction of a length of signals influences on the quality of the results. The minimal period of time that makes possible the detection of selected faults equals 1 s. A vector of five elements (five signals were considered) describes each fault in these cases. Selected groups of faults (usually the intensities of particular faults) became indistinguishable from each other, e.g. $f7_{1,2,3}$, $f10_{2,3}$, $f18_{1,2,3}$.

Final summary of faults detection with division to the possible types is presented in Table VII.

### B. *Validation of a methodology*

Obtained SOM maps can be considered as a diagnostic pattern, based on which it is possible to conclude about technical

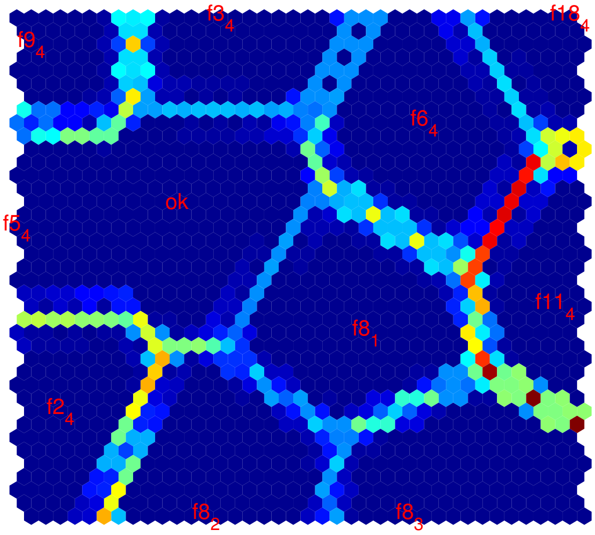


Fig. 6: Obtained maps for weakly detectable faults, map size: 20 × 20, learning steps (WTM, WTA): (20, 15)

TABLE VII: Results assigned to the possible classes of faults

| Type | Faults 1100 s |
|---|---|
| undetectable | $f14_{1,2,3}$ |
| weakly detectable (inside the classes of faults) | $f8_{1,2,3}$ $(f5, f6, f18)_4$ |
| weakly detectable (overall) | $f5_4$ |
| pair of not distinguishable faults | $(f13, f18)_2$ $(f13, f18)_3$ |
| very well detectable | $f1, f7, f13, f18)_{1,2}$<br>$(f2, f7, f15)_3$<br>$(f4, f13, f17)_4$ |

condition of the considered system. In order to validate this approach the new three sets of data for considered 44 faults were prepared (small modification of standard DAMADICS benchmark simulator was made). Obtained data was divided into five groups. The faults, according to the classes: small, medium and large abrupt and incipient, are included in the first four of them. The fifth class includes faults, which were obtained during previous analysis and assigned as the weakly detectable (see Figure 6). From the collection of numerous maps, which were generated previously, five maps were selected according to the particular groups of faults. Basing on these maps, for subsequent cases of new data the best neuron (BMU) from suitable map was determined. The considered fault was assigned as recognizable if the assigned neuron coincided with a neuron corresponding to the same fault from the basic map, or the neuron was located in the same cluster of a map. An example of assigned labels of large abrupt faults is shown in Figure 8, while a full summary of the results is posted in Table VIII. In the last three columns, the numbers of recognized cases of faults for individual groups are presented. It can be observed that the faults $f8_{1,2,3}$ i $f14_{1,2,3}$ are weakly recognizable. The best results are achieved for large abrupt faults, where for the most cases of considered faults the same neurons are assigned similarly as in the basic map.

TABLE VIII: Results of validation of methodology

| Groups of faults | Faults | 1st set | 2nd set | 3rd set |
|---|---|---|---|---|
| Small abrupt | 10 | 8 | 8 | 8 |
| Medium abrupt | 10 | 8 | 8 | 8 |
| Large abrupt | 14 | 12 | 12 | 12 |
| Incipient | 10 | 2 | 2 | 3 |
| Weakly detected | 10 | 4 | 5 | 7 |

## V. Conclusions

In the presented paper the novel approach to the process diagnosis problem based on self-organizing maps was presented. The authors used the DAMADICS benchmark for simulation of faults of various types and intensities, which was the basis of the investigated fault diagnosis problem. Using SOM-based approach and the simulation data it was shown that the faults can be not only precisely detected, but also isolated and identified basing on resulted BMU maps. As it was noticed, applied approach is very sensitive to changes between the signals, which allow to detect, isolate and identify almost all unique cases of faults and provides better results of faults detection and isolation than the previously applied methods for the DAMADICS benchmark problem. Moreover, using SOM it is possible to compare multiple to multiple signals, which improves much the methodology of fault diagnosis due to the lack of necessity to carry out preprocessing procedures of the process signals. In order to validate the presented approach additional datasets were generated from the DAMADICS benchmark simulator, which were slightly different than those generated for the main analysis. The validation procedures show that in the most cases the faults were appropriately classified to the particular clusters.

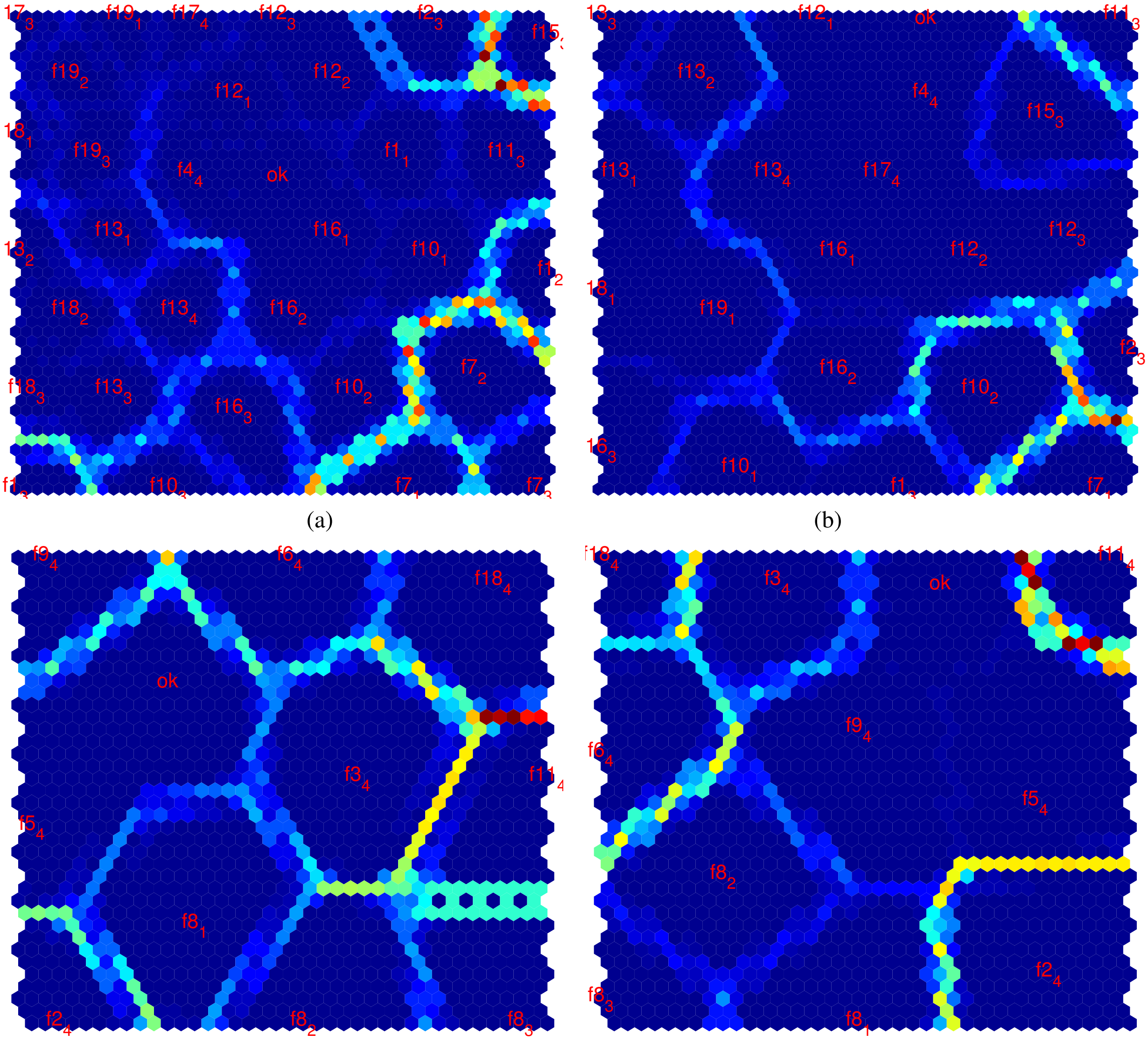
Presented studies is a part of on-going research. Several open problems, which were planned to solve in future, should be highlighted. The improvement of the distinguishability of faults is expected during the additional preprocessing of input diagnostic signals, e.g. their normalization. Additionally, the qualitative measure (e.g. a threshold), based on which the value of a neuron on the BMU map can be used for the identification of a fault, should be developed.

The application of SOM-based algorithm in process diagnostic systems has a great potential due to the very high sensitivity to the differences in signals and datasets and could be automated and successfully adapted for the industrial installations. Since the presented algorithm is time-consuming, it cannot be applied for real-time diagnosis, however further attempts will be made to optimize the SOM-based fault diagnosis algorithm in order to increase its effectiveness.

## References

[1] M. Basseville and I. Nikiforov, “Detection of Abrupt Changes: Theory and Applications”, Prentice Hall Information and Systems Science Series, 1993.

[2] M. Witczak, R.J. Patton and J. Korbicz, “Fault Detection with Observers and Genetic Programming: Application to the DAMADICS Benchmark Problem”, *Proc. 5th IFAC Symp. on Fault Detection, Supervision and Safety of Technical Processes – SAFEPROCESS*, Washington, 2003, pp. 1203-1208.



(a)

(b)

(c)

(d)

Fig. 7: Maps for different times of analyzed signals, (a) well detectable faults; period of time: 1000 s, map size: 25x25, learning steps (WTM, WTA): (35, 15); (b) well detectable faults; period of time: 1 s, map size: 25 × 25, learning steps (WTM, WTA): (30, 15); (c) weakly detectable faults; period of time: 750 s, map size: 20x20, learning steps (WTM, WTA): (25, 15); (d) weakly detectable faults; period of time: 1 s, map size: 20 × 20, learning steps (WTM, WTA): (25, 10)

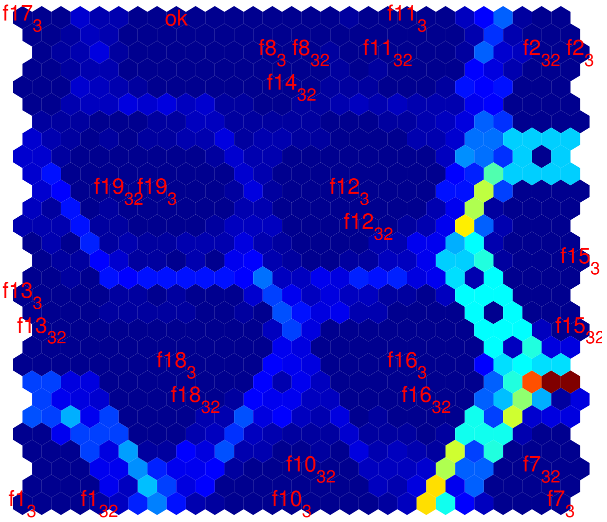


Fig. 8: Exemplary map of validation of proposed approach

[3] A. Lipnickas and J. Korbicz, “Evolutionary Learning in Identification of Fuzzy Models: Application to DAMADICS Benchmark”, *Proc. 6th Domestic Conf. “Diagnostics of Industrial Processes” DPP’03*, Władysławowo, 2003.

[4] J.M.F. Calado, F.P.N.F. Carreira, M.J.G.C. Mendes, J.M.G. Sá da Costa

and M. Bartyś, "Fault Detection Approach Based on Fuzzy Qualitative Reasoning Applied to the DAMADICS Benchmark Problem", *Proc. 5th IFAC Symp. on Fault Detection, Supervision and Safety of Technical Processes – SAFEPROCESS*, Washington, 2003, pp. 1179-1184.

[5] M. Witczak, J. Korbicz, M. Mrugalski and R.J. Patton, "A GMDH Neural Network-Based Approach to Robust Fault Diagnosis: Application to the DAMADICS Benchmark Problem", *Control Eng. Pract.,* vol. 14, 2006, pp. 671-683, http://dx.doi.org/10.1016/j.conengprac.2005.04.007.

[6] Y. Kourd, N. Guersi and D. Lefebvre, "Neuro-Fuzzy Approach for Default Diagnosis: Application to the DAMADICS", *Proc. 4th IEEE Conf. on Digital Ecosystems and Technologies*, Dubai, 2010, pp. 107-111, http://dx.doi.org/10.1109/DEST.2010.5610663.

[7] Y. Kourd, D. Lefebvre and N. Guersi, "Fault Diagnosis Based on Neural Networks and Decision Trees: Application to DAMADICS", *Int. J. Innov. Comput. I.,* vol. 9, 2013, pp. 3185-3195.

[8] G.M. de Almeida and S.W. Park, "Fault Detection and Diagnosis in the DAMADICS Benchmark Actuator System – a Hidden Markov Model Approach", *Proc. 17th World Congress of the IFAC*, Seoul, 2008, pp. 12419-12424, http://dx.doi.org/10.3182/20080706-5-KR-1001.2573.

[9] S. Openshaw and I. Turton, "A Parallel Kohonen Algorithm for the Classification of Large Spatial Datasets", *Comput. Geosci.,* vol. 22, 1996, pp. 1019-1026, http://dx.doi.org/10.1016/S0098-3004(96)00040-4.

[10] W. Melssen, R. Wehrens and L. Buydens, "Supervised Kohonen Networks for Classification Problems", *Chemometr. Intell. Lab.,* vol. 83, 2006, pp. 99-113, http://dx.doi.org/10.1016/j.chemolab.2006.02.003.

[11] D. Bianchi, R.Calogero and B. Tirozzi, "Kohonen Neural Networks and Genetic Classification", *Math. Comput. Model.,* vol. 45, 2007, pp. 34-60, http://dx.doi.org/10.1016/j.mcm.2006.04.004.

[12] M. Amarowicz and A. Katunin, "Clustering of Delaminations in Composite Rotors Using Self-Organizing Maps", *Intelligent Systems in Technical and Medical Diagnostics,* J. Korbicz and M. Kowal, Eds., *Advances in Intelligent Systems and Computing,* vol. 230, Berlin-Heidelberg, Springer, 2014, pp. 149–159, http://dx.doi.org/10.1007/978-3-642-39881-0_12.

[13] S.M. Bhandarkar, J. Koh and M. Suk, "Multiscale Image Segmentation Using a Hierarchical Self-Organizing Map", *Neurocomputing,* vol. 14, 1997, pp. 241-272, http://dx.doi.org/10.1016/S0925-2312(96)00048-3.

[14] C. Amerijckx, J.D. Legat and M. Verleysen, "Image Compression by Self-Organized Kohonen Map", *Syst. Anal. Model. Sim.,* vol. 43, 2003, pp. 1529-1543, http://dx.doi.org/10.1080/0232929032000115182.

[15] W.G. Teng and P.L. Chang, "Identifying Regions of Interest in Medical Images Using Self-Organizing Maps", *J. Med. Syst.,* vol. 36, 2012, pp. 2761-2768, http://dx.doi.org/10.1007/s10916-011-9752-8.

[16] A.G. de Barreto, A.F.R. Araújo and H.J. Ritter, "Self-Organizing Feature Maps for Modeling and Control of Robotic Manipulators", *J. Intell. Robot. Syst.,* vol. 36, 2003, pp. 407-450, http://dx.doi.org/10.1023/A:1023641801514.

[17] M. Johnsson and C. Balkenius, "Sense of Touch in Robots with Self-Organizing Maps", *IEEE T. Robot.,* vol. 27, 2011, pp. 498-507, http://dx.doi.org/10.1109/TRO.2011.2130090.

[18] A. Lendasse, J. Lee, V. Wertz and M. Verleysen, "Forecasting Electricity Consumption Using Nonlinear Projection and Self-Organizing Maps", *Neurocomputing,* vol. 48, 2002, pp. 299-311, http://dx.doi.org/10.1016/S0925-2312(01)00646-4.

[19] G. Simon, A. Lendasse, M. Cottrell, J.C. Fort and M. Verleysen, "Time Series Forecasting: Obtaining Long Term Trends with Self-Organizing Maps", *Pattern Recogn. Lett.,* vol. 26, 2005, pp. 1795-1808, http://dx.doi.org/10.1016/j.patrec.2005.03.002.

[20] C.M. Hsu, "A Hybrid Procedure for Stock Price Prediction by Integrating Self-Organizing Map and Genetic Programming", *Expert Syst. Appl.,* vol. 38, 2011, pp. 14026-14036, http://dx.doi.org/10.1016/j.eswa.2011.04.210.

[21] C.W. Chan, H. Jin, K.C. Cheung and H.Y. Zhang, "Fault Detection of Systems with Redundant Sensors Using Constrained Kohonen Networks", *Automatica,* vol. 37, 2001, pp. 1671-1676, http://dx.doi.org/10.1016/S0005-1098(01)00126-1.

[22] S.L. Jämsä-Jounela, M. Vermasvuori, P. Endén and S. Haavisto, "A Process Monitoring System Based on the Kohonen Self-Organizing Maps", *Control Eng. Pract.,* vol. 11, 2003, pp. 83-92, http://dx.doi.org/10.1016/S0967-0661(02)00141-7.

[23] M. Seera, C.P. Lim, D. Ishak and H. Singh, "Offline and Online Fault Detection and Diagnosis of Induction Motors Using a Hybrid Soft Computing Model", *Appl. Soft Comput.,* vol. 13, 2013, pp. 4493-4507, http://dx.doi.org/10.1016/j.asoc.2013.08.002.

[24] T. Chopra and J. Vajpai, "Classification of Faults in DAMADICS Benchmark Process Control System Using Self Organizing Maps", *Int. J. Soft Comput. Eng.,* vol. 1, 2011, pp. 85-90.

[25] M. Syfert, R. Patton, M. Bartyś and J. Quevedo, "Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems (Damadics): A Benchmark Study", *Proc. 5th IFAC Symp. Fault Detection, Supervision and Safety of Technical Processes – SAFEPROCESS*, Washington, 2003, pp. 939-950.

[26] M. Bartyś, R. Patton, M. Syfert, S. de las Herras and J. Quevedo, "Introduction to the DAMADICS Actuator FDI Benchmark Study", *Control Eng. Pract.,* vol. 14, 2006, pp. 577-596, http://dx.doi.org/10.1016/j.conengprac.2005.06.015.

[27] DAMADICS, "Website of the Research Training Network on Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems" [online], Institute of Automatic Control and Robotics, Warsaw University of Technology, 2004 [viewed: 2015-04-11]. Available from: http://diag.mchtr.pw.edu.pl/damadics.

[28] M. Kościelny, M. Bartyś, P. Rzepiejewski and J. Sá da Costa, "Actuator Fault Distinguishability Study for the DAMADICS Benchmark Problem", *Control Eng. Pract.,* vol. 14, 2006, pp. 645-652, http://dx.doi.org/10.1016/j.conengprac.2005.06.014.

[29] P. Supavatanakul, J. Lunze, V. Puig and J. Quevedo, "Diagnosis of Timed Automata: Theory and Application to the DAMADICS Actuator Benchmark Problem", *Control Eng. Pract.,* vol. 14, 2006, pp. 609-619, http://dx.doi.org/10.1016/j.conengprac.2005.03.028.

[30] F. Previdi and T. Parisini, "Model-Free Actuator Fault Detection Using a Spectral Estimation Approach: the Case of the DAMADICS Benchmark Problem", *Control Eng. Pract.,* vol. 14, 2006, pp. 635-644, http://dx.doi.org/10.1016/j.conengprac.2005.04.001.

[31] V. Puig, A. Stancu, T. Escobet, F. Nejjari, J. Quevedo and R.J. Patton, "Passive Robust Fault Detection Using Interval Observers: Application to the DAMADICS Benchmark Problem", *Control Eng. Pract.,* vol. 14, 2006, pp. 621–633, http://dx.doi.org/10.1016/j.conengprac.2005.03.016.

[32] C.M. Bocaniala and J.M.G. Sá da Costa, "Application of a Novel Fuzzy Classifier to Fault Detection and Isolation of the DAMADICS Benchmark Problem", *Control Eng. Pract.,* vol. 14, 2006, pp. 653-669, http://dx.doi.org/10.1016/j.conengprac.2005.06.008.

[33] D. Düştegör, E. Frisk, V. Cocquempot, M. Krysander and M. Staroświecki, "Structural Analysis of Fault Isolability in the DAMADICS Benchmark", *Control Eng. Pract.,* vol. 14, 2006, pp. 597-608, http://dx.doi.org/10.1016/j.conengprac.2005.04.008.

[34] J.M.F. Calado, J.M.G. Sá de Costa, M. Bartyś and J. Korbicz, "FDI Approach to the DAMADICS Benchmark Problem Based on Qualitative Reasoning Coupled with Fuzzy Neural Networks", *Control Eng. Pract.,* vol. 14, 2006, pp. 685-698, http://dx.doi.org/10.1016/j.conengprac.2005.03.025.

[35] A.R.C. Oliveira and J.M.G. Sá da Costa, "Hierarchic Fault Diagnosis by Pattern-Recognition Approaches Applied to DAMADICS Benchmark", *Proc. 18th IFAC World Congress*, vol. 18, Milano, 2011, pp. 7737-7742, http://dx.doi.org/10.3182/20110828-6-IT-1002.03638.

[36] J. Vesanto, "SOM Implementation in SOM Toolbox, SOM Toolbox Online Help" [Online], Laboratory of Computer and Information Science, 2005 [viewed: 2015-04-11]. Available from: http://www.cis.hut.fi/projects/somtoolbox/documentation/somalg.shtml.

[37] T. Kohonen, "Self-Organizing Maps", Springer Series in Information Sciences, vol. 30, Berlin-Heielberg, Springer, 2001.

[38] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas, "SOM Toolbox for Matlab 5" [Online], Helsinki University of Technology, 2000 [viewed: 2015-04-11], Research report. Available from: http://www.cis.hut.fi/somtoolbox/package/papers/techrep.pdf.

[39] R. Isemann and P. Ballé, "Trends in the Application of Model-Based Fault Detection and Diagnosis of Technical Processes", *Control Eng. Pract.,* vol. 5, 1997, pp. 709–719, http://dx.doi.org/10.1016/S0967-0661(97)00053-1.

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1683–1689

DOI: 10.15439/2015F353
ACSIS, Vol. 5

# Support of Contextual Classifier Ensembles Design

Janina A. Jakubczyc
Wroclaw University of Economics
ul. Komandorska 118/120,
53-345 Wrocław, Poland
Email:
janina.jakubczyc@ue.wroc.pl

Mieczysław L. Owoc
Wroclaw University of Economics
ul. Komandorska 118/120,
53-345 Wrocław, Poland
Email:
mieczyslaw.owoc@ue.wroc.pl

□

***Abstract*— An idea of contextual classifier ensembles extends the application possibility of additional measures of quality of base and ensemble classifiers in the process of contextual ensembles design. These measures besides the obvious classifier accuracy and diversity/similarity take under consideration the complexity, interpretability and significance. The complexity (the number of used measures and multi level measure structure), the diversity of the scales of used measures and the necessity of the fusion of different measures to one assessment value are the reasons for user support in contextual classifier ensembles design using fuzzy logic and multi criteria analysis. The aim for this paper is an idea of the framework of the process of contextual ensemble design.**

## I. Introduction

The contextual classifier ensemble other than, name it classical ensemble, gives the possibility to look more profoundly at problem under consideration and classifier ensemble design. The reason is context that is the criterion for diverse classifier creation. The contexts of the classification problem give the possibility to view it from different perspectives, to get more familiar with it, to choose the more appropriate contexts for each case.

This approach of classifier ensembles design extends the application possibility of additional measures of quality of the base and ensemble classifiers in the process of contextual ensemble design. These measures besides the obvious classifier accuracy and diversity/similarity take under consideration the complexity, interpretability and significance. The additional measures appear in the stages of the process of contextual ensemble design. This process and the idea of contextual classifier ensemble are introduced the section II.

The framework of contextual classifier ensembles design is included in the section III.

The complexity (the number of used measures and multi level measure structure that), the diversity of the scales of used measures and the necessity of the fusion of different measures in one assessment value for the given stage of process design are the reasons for applying fuzzy logic and multi criteria analysis. This methodology is introduced in section IV.

The first results of experiment that employs this idea contextual classifier design support is included in the section V. And finally the section VI gives some summary of the introduced idea.

## II. The Essentials of Contextual Classifier Ensemble

Classifier ensembles allow the different needs of a difficult problem to be handled by classifiers suited to those particular needs. Classifier ensembles provide an extra degree of freedom in the classical bias/variance tradeoff, allowing solutions that would be difficult, if not possible to reach with only a single classifier. Because of these advantages, classifier ensembles have been applied to many difficult real-world problems see: Oza and Tumer, 2008 [5], Patel and Nawathe 2013 [11].

The classifier ensemble is a set of base classifiers that together are solving the problem of the discrimination. There are three basic mechanisms of creation the classifier ensembles. The first is a mechanism of creation the base classifiers see: Dietterich, 2000 [2], Zhang and Ma, 2012 [9]. There are the following approaches:

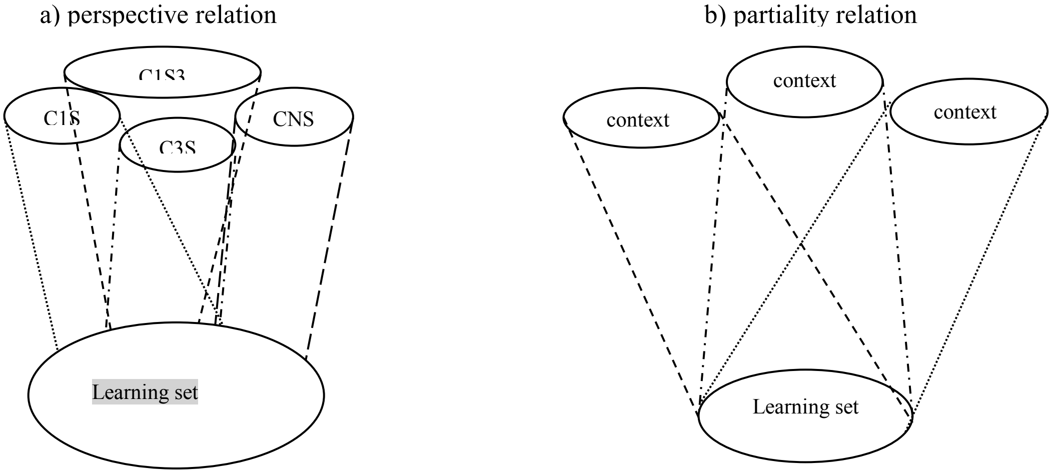
– data manipulation - the models for different data subsets of the learning set or for different attribute subsets
– different techniques for modeling one learning data set
– the one kind of model with different parameter sets for one learning set.

Contextual classifier ensemble is a variant of classifier ensembles that is based on data

□ This work was not supported by any organization

978-83-60810-66-8/$25.00©2015, IEEE

manipulation. Instead of random samples generation, or weighing examples according to correct classification in consecutive modeling step, contextual classifier ensemble generates attribute sets according known or discovered context from data that describe problem under consideration – compare: Jakubczyc, 2007 [3]. The most interesting situation is when the contexts are discovered in learning set. This case is a subject of this study. The context can be a single category i.e. localization with the urban and rural values that determine two contextual situations (localization=urban, localization=rural) or complex concept that is described by a set of attributes in the form of some kind of model for example decision tree. Each base classifier represents description of classification problem for single context or for single contextual situation. In the decision tree model each path from the root to the leaf indicates the contextual situation and the whole decision tree means the context. Whether it is context or contextual situation is determined by the type of relation between contexts and learning examples that represent problem under consideration are perspective and partiality (Figure 1). The perspective relation stands for possible contextual views on a problem. The partiality relation describes only subsets of more appropriate models of the problem for contextual situations. The last possible relation means partiality and perspective relation simultaneously (not presented in the Figure 1)


a) perspective relation

b) partiality relation

Figure 1. The relations between contexts and learning set of example

The type of relation that is used in creation contextual classifier ensembles is determined by the quality of contextual models. If the level of classification accuracy is acceptable for classifiers of identified contexts there is perspective relation. In this case the contextual classifier ensemble consists of models for each discovered contexts. On the contrary, unacceptable accuracy forces the one to take under consideration only some models for contextual situations that may guarantee the requested quality for each context. In this case there is partiality relation that may not cover the whole learning set. In this case the contextual classifier ensemble consists of models for the contextual situations not for whole context (for example in Figure1 C1S1 states for contextual situation 1 from context 1 model) with acceptable accuracy. It seems acceptable because the learning set may not be representative sample according to statistic theory.

There is also possible the mix relation that joins the first two. Our interest for now is in the first one that means the base classifiers are models for whole contexts (perspective relation). For now this focus allows us to exempt from the attendance of uncovered example by contextual classifier. It will be the subject of future research.

The selection of classifiers to the ensemble that solve the classification problem is the second mechanism. There are two problems: how many classifiers and which classifiers should be chosen.

The general answer for the first question is: “the more classifiers the better” according to the jury theorem of J.A. Marquis Condorcet - originally: Condorcet Marquis J.A.: Sur les elections for scrutiny, [in:] Histoire de l’Academie Royale des Sciences, 31-34, 1781 see: Cunnigham, 2007 [1] - but the empirical studies shows that even few classifiers may improve classification accuracy described compare: Abreu and Canuto, 2007 [10] Schiele, 2002 [7]. They have showed that ensemble of three or of five classifiers may bring much improvement too. In the case of contextual classifier ensemble the number of base classifiers in ensemble is limited to the number of identified contexts or the number of contextual situations.

The choice of the most appropriate base classifiers is generally based on two measures – applied by Kuncheva and Whitaker, 2003 [4]. The first measure is classification accuracy that should be above 50%. The second are the measures of diversity. Intuitively it means that chosen base classifiers should be mistaken on the different cases thus they can complement one another. Because the results of empirical studies are ambiguous, i.e. there is no linear relation between the diversity and accuracy, so the increase of diversity not necessary causes the increase of classification accuracy. So there is a need for additional measures that can support the choice of base classifiers. The contextual approach to creation classifier ensembles gives such possibility. The additional perspectives may be user evaluation of identified contexts and contextual base classifiers.

There is a need for additional measures also in the next step of combining single decisions of base classifiers into one final decision (it is the third mechanism). Applied techniques are determined by continuity or discreteness of class values. Since decision tree is chosen as the classification algorithm, we are dealing with nominal values of classes. In this case, the voting schemata are accessible techniques of the combination of the decision of base classifiers in this case. In simple voting schemata each classifier has one vote. But looking at different base contextual classifiers according to their quality, significance, comprehensibility, the equally assignment of votes is difficult to accept. The none-equivalence of base classifiers should be taken into account in global assessment of contextual ensemble.

The context as a criterion for creation classifier ensembles gives the possibility to look more profoundly at generated classifiers trough different contexts and at identified contexts alone. The profound assessment of the base contextual classifier and contextual ensemble needs the framework to account for the following aspects: identification of possible assessment criteria for single contextual classifier and for contextual classifier ensemble with different level of detail and aggregation, the possibility of evaluating simultaneously qualitative and quantitative assessments, contextual classifier ranking trough fuzzy multi-criteria analysis. The general framework with methodological approach is presented in the next section. The working example is shown in further section.

## III. The Framework for Contextual Classifier Ensemble

The proposition is aimed at more reasonable creation of classifier ensemble through the different, and not random criterion to build classifier ensembles. The contextual classifiers give the possibility of more profoundly evaluation of base and ensemble classifiers. The user active participation in this process of assessment is not to overestimate. The general framework of creation contextual ensembles is introduced at Figure 2. There is one global goal to choose the most appropriate contextual classifier ensemble and one intermediate goal to choose the best base contextual classifiers.

The classification problem is difficult so there is not possible to built the one model with acceptable level of accuracy. The possible solution is classifier ensemble, in this case with the contexts as criterion for building base classifiers.

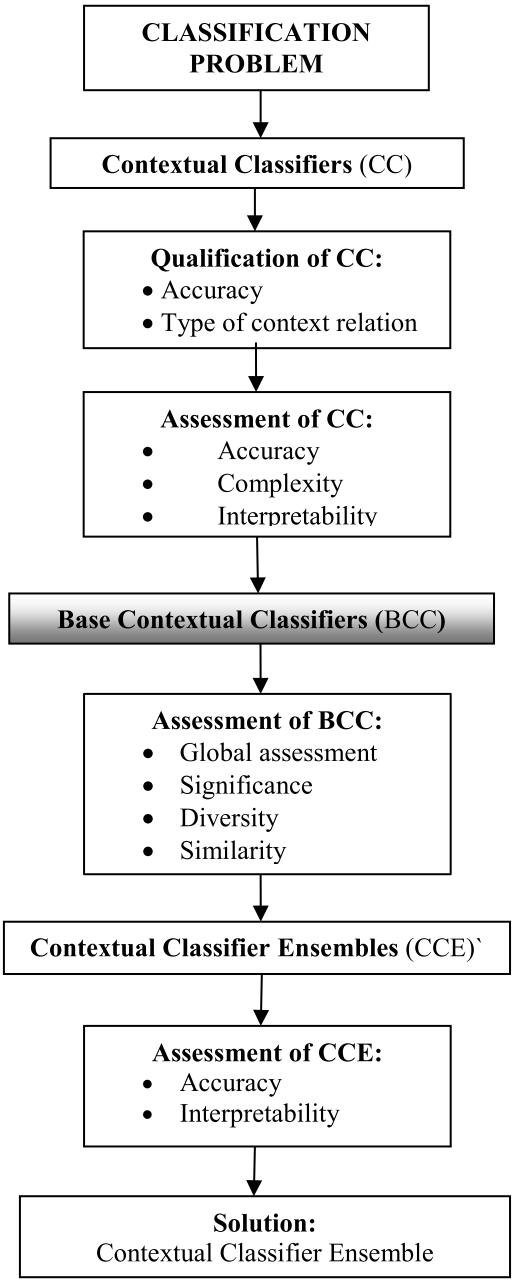


Figure 2. The general framework of creation contextual ensemble

The contextual classifiers to became the base contextual classifiers have to achieve the level of accuracy above 60% for each contextual situation (it takes place when the relation between contexts and learning set of examples is so called 'perspective relation') - presented in Figure 1 - and for each class. The contextual classifiers that have passed the qualification step are evaluated objectively according to accuracy and subjectively with support of detailed measures for complexity and interpretability (see Table 1.). To the ranking of base contextual classifiers are used two measures for each pair: similarity and diversity and user evaluation of classifier significance.

The similarity and diversity do not play a key role in process of evaluation since the number of base contextual classifiers are finite, so there can be used all possible combination of classifiers. The measures of similarity and diversity can be used to examine their quality influence.

The possible combinations of base contextual classifiers then are evaluated using objective measure of accuracy and subjective measure of interpretability. The ranking list of contextual classifier ensembles is the result of this process. The choice the most appropriate contextual classifier ensemble is up to the user. The detailed methodology used in the evaluation process is presented in the next section.

## IV. The Used Evaluation Methodology

The method for evaluation of base contextual classifiers and contextual classifier ensembles depends on character of the criteria and their properties (see Table 1). As we can see the criteria have the hierarchical structure, the different types of value (qualitative, quantitative), and different value scale. This imposes the use of fuzzy logic system introduced by L. Zadeh [8] as representing criteria values at all levels. The non-equivalence of identified criteria determines the use of pair wise matrices for the hierarchical structure of evaluation criteria means the need for gradual criteria aggregation at each higher levels.

The first step (the first column in Table 1) is the candidate qualification. To pass, the contextual classifier should have the level of accuracy above 60% and the context with perspective relation. The evaluation of contextual base classifier includes the three layers hierarchy of criteria and gradual way of evaluation.

The level 1 encompass four criteria. Each of them is represented as fuzzy membership function. The shape of these functions and key points are determined as result of analysis and requirement. The value of interpretability increase s when the value of 'novelty' decreases. There is assumed the 20% level of 'novelty' that does not influence the interpretability value.

Table  1. The criteria for assessment of base contextual classifier and classifier ensemble

| Contextual classifier | Base contextual classifier | | | Contextual classifier ensemble |
|---|---|---|---|---|
| | Level 1 | Level 2 | Level 3 | Level 4 |
| Accuracy<br>Type of context relation | Comprehensibility | Model interpretability | Base contextual classifier assessment | Accuracy<br>Interpretability |
| | Novelty | | | |
| | Number of descriptors | Model complexity | | |
| | Number of branches | | | |
| | | Model accuracy | | |
| | | | Similarity | |
| | | | Diversity | |
| | | | Significance | |

The value of ‘comprehensibility’ influences ‘interpretability the another direction: the higher ‘comprehensibility’ value the higher the value of ‘interpretability’. The range with full understanding is assumed in the interval 85-100%.

The fuzzy function of number of model descriptors has the two key points that determine the optimal range between 2 and 10 descriptors. The more descriptors indicate the more complexity of the classification model. Similarly is the matter with the ‘number of branches’ but with duplicated values.

The synthetic measure for base contextual classifier takes under consideration the non-equivalence of identified criteria. The pair wise comparison matrices (see: Saaty, 1977) are applied to derive implicit weights for a given set of criteria. The user has the possibility to introduce the significance weights for pair comparison matrix.

The model accuracy (Figure 3) has three key points. The first one indicate the minimal level of accuracy that is required for classifier ensemble i.e. 60%. The next two key points indicate the most desired range of accuracy i.e. 80%-95%. The values close to 100% are unacceptable.

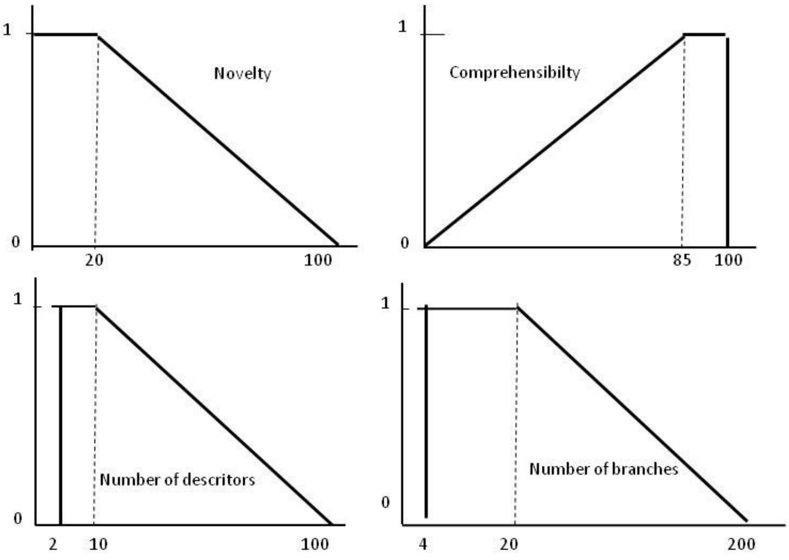


Figure 3. The shapes of fuzzy membership functions of the level 1 criteria

The complexity of contextual classifier is determined by detailed parameters of the number of descriptors and tree branches (Figure 4). The higher value of model complexity indicate the worst classifier quality. The first and second key point indicates the range of minimal level of complexity that may differ according to number of possible descriptors. After that , the quality of model decreases when the complexity increases. For different classification problem the user recommendation of key point may differ.

The interpretability indicates to what extent user understand and unexpected the classifier according to his knowledge. In this case the fuzzy membership function has the shape of monotonically increasing – the higher the value of interpretability the higher the quality of classifier.

The next level 3 takes into account synthetic evaluation of base classifier ensemble (interpretability, complexity and accuracy) and similarity and diversity measures and significance of base contextual classifiers.

The similarity measure is used as a means to eliminate the most similar classifiers from one ensemble. On the contrary the measure of diversity gives the possibility to include in one ensemble the most different contextual classifiers.

The weights assignment to particular base contextual classifier is the task for the user with his knowledge and experiences. So the assigned weights may differ from one user to another. The pair wise matrices are applied to this operation.

The calculation of synthetic measures takes place at level 2 for interpretability, complexity, and accuracy, at level 3 for base contextual classifier evaluation and at level 4 for contextual classifier ensemble.

All these calculation take into account non-equivalence of used criteria and identified contexts . This not-equivalence is measured by the general assessment of the base contextual classifier and the pair-relation of significance between contextual classifiers determined by the domain expert. Such approach allows for more precise choice of the most appropriate classifier ensemble for the problem under consideration. There are not a priori settled weights. The user each time introduces the values of the pair comparison between the base contextual classifiers. The system interface gives the possibility to view the all possible values with their interpretation.

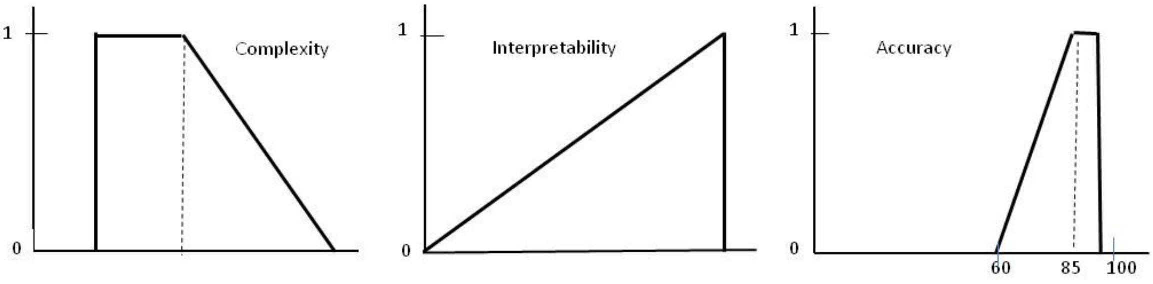


Figure 4. The shapes of fuzzy membership functions of the level 2 of criteria

The means for criteria aggregation are three following operators :

- maximum pessimism $D1 = \min(\mu_1(x_1)^{\alpha1}, \mu_2(x_2)^{\alpha2}, \dots, \mu_n(x_n)^{\alpha n})$
- multiplicative $D2 = \prod_{i=1}^{n} \mu_i (x_i)^{\alpha i}$
- additive $D3 = \sum_{i=1}^{n} \frac{\alpha_i * \mu_i(x_i)}{n}$

where:

$\mu_1(x_1),(\mu_2(x_2), \dots ,(\mu_n(x_n)$ - membership function,

$\{x_i\}$ – numerical and nominal criteria of classifier quality

$\alpha_1, \alpha_2, \dots, \alpha_n$ – the relative significance of the criteria and in the case of global criteria assessment the relative significance of base contextual classifiers.

The global assessment of contextual classifier and contextual classifier ensemble is determined by the three criteria: maximum pessimism (D1) multiplicative (D2), additive (D3) to observe the relation between them. D2 and D3 have the propriety of little value compensation of one criterion by the increasing the other criteria. The conformity of the obtained results for D1, D2, D3 increases the confidence of study results. All of them gives the global value in the range of 0 and 1 (the higher the better).

## V. WORKING EXAMPLE

The experiment was conducted on the problem of the banks client behavior and the estimation of clients behavior. The classification task was to predict whether the client is active or non-active. This problem is very important for each bank management If the client is non-active there is an indication to take pro-active action to the client to keep him as an stable bank client for the future and do not let him switch the bank for example. The bank provided 24 000 examples in learning set. The Bank is providing 24 000 representation data. There were nine discovered context that were the basis for contextual classifier design.

Our experiment aimed at two issues. The first is to verify the usability of the proposal of the framework and applied methodology. It proved to be appropriate and flexible. The flexibility concerns the automatic or manual design of contextual classifier ensemble. Because of the subjectivity of some used measures, as for example interpretability and comprehension, assumption in automatic process for these measures default values was not the best idea.

The second aspect refers to view the classification problem according to discovered contexts and knowledge/experience of the user. Generally the designed contextual ensemble by the various users were different but their quality quite similar. They differ from the weights assignment to significance of contexts and contextual models. Each of users built own fuzzy rule set and used the possibility to design few contextual ensemble to choose the best one. It results of knowledge and experience of the users.

## VI. Summary

The context as criterion for creation classifier ensembles gives the possibility to consider more profoundly classification problem under consideration. The user activation seems to be important too, although the interactive problem solving do not forces the user to join the process. But from our current experiments with the user the solutions were more reasonable. The applied approach proved to be adequate. The applied methodology showed to be appropriate too. Additionally the users fast get familiar with it.

We observed how the level of diversity changes the quality of classifier ensemble, how the knowledge and experiences influence the solutions. But it is too few experiments to conclude more general conclusions. Our work is going on.

## References

[1] Cunnigham P., Ensaml;e Techniques Technical Report UCD-CSI-2007-5 April 2, 2007

[2] Dietterich T. G.: Ensemble methods in Machine Learning. [in:] The Proceedings of 1th International Workshop on Multiple Classifier Systems, s. 1-15, 2000

[3] Jakubczyc J. A., Contextual Classifier Ensembles. [in:] Abramowicz W. (ed.): Business Information Systems, LNCS 4439, Springer 2007.

[4] Kuncheva L.I., Whitaker C.J.: Measures of diversity in classifier ensembles, Machine Learning 51, s. 181-207, 2003.

[5] Oza N. C. and Tumer K., Classifier Ensamles: Select Real World Applications. Information Fusion Vol 9 Issue 1 2008, Elseviewer

[6] Saaty T. Scaling Method for Priorities in Hierarchical Structures //J. of Mathematical Psychology. 1977. Vol. 15. 3. p. 234 – 281.

[7] Schiele B.: How many Classifiers Do I Need? IEEE Pattern Recognition, vol.2, 2002

[8] Zadeh L. A.: Is there a need for fuzzy logic?, Information Sciences an International Journal 178 (2008) 2751-2779, www.elsevier.com/locatelinks

[9] Zhang C., Ma Y.(eds.): Ensemble Machine Learning: Methods and Application, Springer Science+Business Media LLC 2012

[10] Abreu M. C. C., Canuto A. M. P.: Using Fuzzy, Neural and Fuzzy-Neural Combination Methods in Ensembles with different Levels of Diversity, in: Proceeding ICANN'07 Proceedings of the 17th international conference on Artificial neural networks, Springer-Verlag Berlin, Heidelberg 2007, p. 349-359

[11] Patel S. P., Patel M. P., Nawathe A. N.: A review on Ensemble of Classifier Using Artificial Neural Networks as Base Classifier, International Journal of Computer Science and Mobile Applications, vol.1 Issue 4, October 2013, p 7-16

DOI: 10.15439/2015F342
ACSIS, Vol. 5

# Benefits of Knowledge Acquisition Systems for Management. An Empirical Study

Moh'd Alsqour
Wroclaw University of Economics
ul. Komandorska118/120,
53-345 Wroclaw, Poland
Email: mohsqour@wp.pl

Mieczysław L. Owoc
Wroclaw University of Economics
ul. Komandorska 118/20,
53-345 Wroclaw, Poland
E-mail:mieczyslaw.owoc@ue.wroc.pl

*Abstract—* **The main objective of conducting this study is to shed light on the benefits and role of data warehouse (DW) as a source of knowledge acquisition in enhancing the process of decision-making. It is believed and assumed that meaningful and significant information can be acquired from DW which provides valuable knowledge to support business process and decision-making.**

**The theoretical assumptions are supported by the results of a questionnaire survey, which was conducted on top management of Jordanian firms. The questionnaire was developed based on the findings from related literature and other related research questionnaires. Over 250 firms, which were listed on Amman Stock Exchange (ASE) at the time of data collection, were involved in the survey. The researchers arrived at scores of significant and remarkable results regarding DW and its benefits and role in enhancing the process of decision-making. In general, the respondents had a positive attitude towards the use of DW.**

## I. Introduction

Today's organizations face a very hard time, largely as a result of competition, globalization, automation and scarcity of resources. As the business environment is changing. At the same time, those companies are likewise evolving. In this changing environment, companies are much more eager in getting immediate and accurate information to make better decisions. Successfully supporting managerial decision-making has become critically dependent upon the availability of integrated and high quality information organized and presented to managers in a timely and easily understood manner ([1]; [2]; [3]).

In such environments it is important to assure decision-makers of the quality of data they use ([13; [12]). [16] claim that Data warehouse (DW) hastens the process of retrieving information needed for decision-making. DW technology has emerged as a key source and powerful tool for delivering and accessing information for decision-makers ([4]; [1]; [5]; [6]; [7]; [3]). Since the 1990s ([8]; [9]), DWs have been an essential information technology (IT) strategy component for large and medium-sized global organizations [8].

Despite the recognition of data warehousing as an important area of practice and research, there is little empirical research ([3]; [17]) about implementation of DW in general ([18]; [14]; [17]; [3]).

Timely and informed decision-making is becoming crucial for the long-term success of businesses [20]. [21] claims that business decisions must be made with speed and accuracy if organizations are to remain competitive. There is quasi-consent that DW provides more detailed and accurate information for decision-makers to improve their decisions. Considering the usefulness of DW there has been little research in Jordan on DW, i.e. it has been comparatively less investigated in Jordan (exceptionally our last paper during KAM'14 – see [76]. Therefore, the main focus of this study is on the advantages and benefits of DW as a provider of information to the process of decision-making. It investigates mainly the relation between decision-making, the need for information and the employment of DW in Jordanian firms.

In addition, this study investigates how top management of Jordanian firms perceives the effectiveness (usefulness) of DW as a source of reliable and accurate information for decision-making.

## II. Aim of the Study

In this paper, a field study of DW and its benefits and role in easing and enhancing the process of strategic decision-making among top managers in Jordanian firms were investigated. Therefore, the main aim of this study is to investigate the benefits, which are provided by DW, in enhancing the process of decision-making.

It has been ascertained that DW is superior to traditional database and improves the process of decision-making. Therefore, the study's aim is to investigate empirically whether or not the DW provides better and more accurate information.

A study of a large number of data warehousing practitioners and experts by [24] showed that the implementation of DW was motivated more by internal pressures than external. A majority of the respondents said that the need was information related, including the need for better access to information, more accurate information and a single source of data. As the authors point out, organizations appear to initiate DW projects to provide decision-makers with accurate and effective information. [25] are of the opinion that improving access to information and delivering better and more accurate information are motivations for using DW.

978-83-60810-66-8/$25.00©2015, IEEE

## III. THE NOVELTY VALUE OF THE STUDY

This study was applied in Jordan, which is one of the developing countries in the Middle East. As Jordan's firms are not in isolation from the rest of the world, they are also influenced by the current competitive environment. The implementation of advanced and recent innovations, such as DW, is essential for the firms in developing economies, such as Jordan.

Despite the exaltation and adulation of DW, there is a need for evidences that the implementation of DW improves the quality and accessibility to information. Therefore, this study aims to practically investigate whether or not the implementation of DW improves the quality and accessibility to information and enhances the process of decision-making. Doherty and Doig in [26] drew attention to the role and importance of information accessibility. The authors claim that the information accessibility is a precursor of information quality-it has a significant impact on the information's usage, and consequently is an indicator of the DW's success in storing and processing information.

In addition to information's quality, previous literature and studies on DW have emphasized the importance of accessing information. The accessibility to information and their quality are crucial to the success of their use. It has been claimed that the use/ implementation of DW improves the quality and accessibility to information, and consequently leads to sound decisions. In other words, it leads to more fact-based decisions. DW has, according to DW literature, the ability to store a vast amount of data in a usable and appropriate form for the decision-makers' needs and uses. Although a wide range of primary and secondary sources has emphasized the importance and role of information quality and accessibility in enhancing the process of decision-making, little empirical research has been conducted so far. Such claims need to be tested empirically. It is essential, therefore, that the researchers investigate whether or not DW provides easy access to data and information, frequent, accessible and timely reports and more accurate, useful, reliable, complete and relevant information to decision-makers.

Based on extensive literature review, the researchers have identified that firms are often unsuccessful due to a lack of appropriate information or more precisely their inability to get the right information to the right person at the right time. The availability of apposite information to decision-making helps mangers in taking reliable decisions, which improve the firm's performance.

Additionally, the researchers have not found and are completely unaware of any empirical studies regarding the implementation of DW in Jordan. Therefore, it is hoped that the findings of this study give the readers and those who are interested in these issues practical insights into DW's field in Jordan. To some extent, it is one of the academic contributions. It contributes to our understanding of the DW in general and in Jordan in particular, and may form a basis and motivation for future research in the important fields. It is also believed that the final outcome of this paper adds up to the improvement and development in DSS, such as DW, by helping their users and developers to be more aware of the data and information's quality.

## IV. RESEARCH METHODOLOGY

The writing of this paper is passed in different phases. In the initial phases of the study, renowned journals, publications, conferences proceedings and books were reviewed. In addition to these sources, the findings of numerous empirical studies were researched and analyzed. As the researchers previously pointed out, the sample of the study comprises all the 277 firms, which are listed on ASE at the time of the data collection. Thus, a questionnaire was found to be the best instrument for collecting the data in this study. During the next phases, therefore, a survey questionnaire was conducted with the top managers of Jordanian firms. The questionnaire, which is used in this study, is based on previous studies and the researchers' assessments and discretion and adapted to suit the objectives and requirements of the study.

The data, which were collected, is very quantitative in nature. Therefore, in the final phases of the study, the data were statistically analyzed by employing Statistical Package for the Social Sciences (SPSS) in order that proper descriptive and inferential statistics to analyze the results and draw conclusions can be reached, including means, frequencies, standard deviation,.

## V. LITERATURE REVIEW

The concept of data warehousing has evolved out of the need for easy access to a structured store of quality data that can be used for decision-making [27 p. 5]. Organizations have vast amounts of data but have found it increasingly difficult to access it and make use of it [27 p. 5].

As an attempt to solve the problem, DWs were introduced. DWs have become the focal point for decision support in organizations today [28]. [27 p. 5] claim that the data warehousing offers a better approach. Data warehousing implements the process to access heterogeneous data sources; clean, filter, and transform the data; and store the data in a structure that is easy to access, understand, and use. The data is then used for query, reporting, and data analysis [27 p. 5]. [29] also claim that the data warehousing has emerged as an effective mechanism for converting data into useful information.

DW systems offer efficient access to integrated and historical data from heterogeneous sources to support managers in their planning and decision-making [30]. [31] also claim that data warehousing provides an infrastructure that enables businesses to extract, cleanse, and store vast amounts of data. Most medium to large organizations, according to [32], operate DWs.

It has been claimed that the main DW's role is to support decision-making. However, the role of DWs has been broadened. [33] state that DW provides information from external data sources for decision-making. DW has the potential to create radical changes to existing business processes and is often viewed within the context of business process reengineering [18]. Accordingly, [34] claims that DW gives business' users the ability to analyze data. [35] also claim that DWs enable organizations to exploit decision-making.

According to [36], DWs provide the basis for management reports and decision support. In support of the above mentioned claims participants to a study by [6] agreed that the Return on Investment (ROI) for the DW was well justified through considerable gains in productivity and enhanced quality of customer service. Moreover, in an independent detailed study of 62 organizations worldwide ([34]; [37]), the major findings of International Data Corporation (IDC) based upon 62 case studies of organizations that have successful DWs in use are an average three-year ROI of 401percent was realized by organizations building DWs. Although this study is primarily focused on quantitative information, there are several qualitative benefits ([34]; [37]), such as providing standardized, clean and value-added data to create information from disparate sources. In addition, the DW makes the data available across corporate organizations and provides the needed information quickly.

The DW is developed in order to support the integration of external data sources ([38]; [39]) for the purpose of advanced data analysis. [40 p. 35] argues that a DW produces tangible impacts to the quality of day-to-day business transactions. Previous research on DW has produced some encouraging findings about its benefits and indicated that a DW can offer several benefits to an organization [18], such as enabling effective decision support; ensuring data integrity, accuracy, security, and availability; easing the setting and enforcing of standards, facilitating data sharing, and improving customer service [35]. [41] presented time savings for data suppliers and for users, more and better information, better decisions, improvement of business processes, and support for the accomplishment of strategic business objectives as benefits from data warehousing.

Furthermore, [5], who examined data warehousing at the Housing and Development Board (HDB) in Singapore, found that the main benefits of the DW, which were developed by HDB, are enabling the users to have access to consistent and reliable data in a timely fashion which facilitated forecasting and planning efforts and improved decision-making. In addition, a study by [42] revealed that DW appears to be used more to improve the flow of information in an organization than to change the way the organization does business. The authors found that more and better data is the greatest realized benefit from DW. Moreover, a study by [43] identified time savings, new and better information, and improved decision-making as benefits of DW.

[44] conducted an explanatory case study at a financial services organization to investigate how DW provides decision support to individual decision-makers. The results showed that the organizations successfully automated the retrieval and input of data for front-end users. [40 p. 33-34], who interviewed people from seven companies, found that the benefits of implementing DW were improving asset management, reducing customer support costs, auditing billing practices, terminating unprofitable product, reducing staff requirements and running the business. [28], who described the DW implementation at Blue Cross and Blue Shield of North Carolina (BCBSNC), claim that the DW had resulted in many organizational benefits, including better data analysis and time savings for users. [45], who looked at the DW of Egypt's Cabinet Information and Decision Support Center, found that the DW provides a lot of benefits to the users, including ease of access to the information, fast and more consistent reports, support the decision-makers and integrating the data from various sources.

Additionally, [7], who conducted a laboratory experiment in 2006, found that the implementation and use of DW improves the DSS users' decision performance, by which he means improving the quality of the DSS by adding a DW can improve information availability and quality and enhance DSS users' decision performance. In conclusion, the study showed that DW can have a positive impact on decision-making.

[46], who described an example of implementing DW in medical institutions, found that the DWs provide the users access to important information. [47], who conducted a survey to find out how DW assists decision-making process in healthcare, found that the DW provides better accessibility to data, integrated disparate data sources and improved decision-making. [48] found that all companies, which are studied, recognized some benefits such as cost reduction, reach-out to other markets, increase in sales, time saving in amount and preparation of reports and more effective decision-making based on the obtained information. [49], who conducted two case studies on American Airlines and Hallmark Cards, found the easy to use, speedy information retrieval, more information, better quality information, improved productivity, and better decisions as benefits of DW. [50], who examined the implementation of DW in public security, found that the DW is very important in improving the comprehensive ability of leadership and decision-making. In addition, it quickly and efficiently integrates heterogeneous data sources.

Previous literature on DW, such as [42] and [30], claims that the DW does not create value by itself; the value comes from the use of the data in the DW. [30] claim that improved decision-making results from the better information available in the DW. By making the right information available at the right time to the right decision-makers in the right manner, DWs empower the users with the ability to make the right decisions [51]. [42] also claim that this use can result in numerous benefits, including more and better information, improved user ability to produce information and reduced effort by developers to produce information. [52] also maintain that DWs have tremendous potential to present information. The greatest potential benefits of the DW occur when it is used to redesign business processes and to support strategic business objectives [41].

[53] also identified many different measures of success; these include benefits such as data accuracy, useful information, accurate information, ease of use, user satisfaction, time to make decision and increased revenue. However, [54] claims that despite clear evidences that many DW projects have resulted in interesting business benefits, there are also many examples of cost and schedule overruns and dissatisfaction regarding the results from these projects. [55] argue that DW is one of the key developments in the IS field and has plentiful benefits. In addition, [56] indicates that the introduction of a new IS into an organization should deliver multiple benefits.

Since the early 1990s, DWs have become the technology of choice for building data management infrastructures [8] and been investigated and implemented around the world in many areas and by many researchers, authors, and scholars [57]. According to [58] p. 13, the early successful implementation of DW dates back to mid-1980s at ABN AMRO Bank (Netherlands). The author claims that the end-user's needs were the key feature behind the implementation. As a result, those requirements were modeled rather broadly, and all available data was stored in the DW. In fact, in the first few years of general use, its usage had grown at an annual rate of 50%, and by 1995 the DW had supported some 3,000 end-users. [58 p. 17] also mentioned that a study of 62 DW projects, which was conducted in 1996, showed an average return on investment (ROI) of 321% for these enterprise-wide implementation in an average payback period of 2.73 years.

In addition, [59], who investigated whether lodging companies are involved with DW technology through a sample of twelve large lodging corporations, found that the most of hotel corporations in the study were using their DWs to support market analysis. However, [13], who conducted a survey on a large Australian public organization, found that 60% of the users were with limited or no usage (or were anticipating the use in the future) of the DW. The data also helped the users to make informed decisions and the data, which was retrieved from the DW, was also presented to the senior management and other strategically oriented sections in the form of reports i.e. annual and quarterly reports.

Similarly, [18], who surveyed DW's managers and data suppliers from 111 organizations in different regions of the United States (US), also found that all companies had operational DW and nearly all of them considered that their initiative is successful. In other words, 26% of the respondents considered it runaway success. Another survey by Forrester showed that 878 IT decision-makers in the US enterprises were somewhat satisfied with the accessibility and quality of customer information; 82% of these respondents are satisfied or very satisfied [60]. [35] also found that about 55% of 196 respondents firms (107) in two major states in the US had already adopted and used DWs.

Similarly, [23] p.192-196 found that 60% of the respondents consider the functionality of their DW below expectations. 40% of the dissatisfied group was actually still using it. This means that 80% of the respondents were still using their internal DW, which is a good indication of the overall degree of satisfaction.

In addition to those studies, [61], who reported the results of a survey which was conducted at The Data Warehousing Institute (TDWI) World Conference in New Orleans 2003, found that 45% of the respondents had been already in production with their current DW or implementing a second or third release.

The 2007's IBM Data Warehousing Satisfaction Survey showed that 56% of those questioned were very successful with their DW (200 end-user enterprises were participated), however, 43% of the respondents acknowledged the need for improvement, as there are still a number of business and technical challenges confronting the enterprises which make use of DWs according to the survey [64]. The success of DW's implementation, according to [64], is growing as 56% of the respondents were very successful and satisfied. [65], who conducted a survey on 84 users of DW, found that the majority of respondents (73%) in the surveyed firm were successful in obtaining and accessing the needed data and information from the DW, only two respondents indicated that they were not at all successful, while 56% indicated somewhat successful and 13% indicated very successful. In addition to these results, 67% and 33% of the respondents rated the importance of the obtained information in performing their job better as vital and somewhat important respectively.

Numerous studies have also shown successful implementation of DW. For example, [66] used a case study and conducted a series of interviews at Continental Airlines. The results showed that the organization has realized an enviable level of DW maturity and significant cumulative benefits. [67], who surveyed 244 members of TDWI, found that 51.2% of the respondents had at least one DW application and 30.3% were still in development stage. 17.2% were still in planning stage and 1.2% of respondents had no any efforts made in their organizations to implement the DW. This result showed that more than 80% of respondents had implemented DW.

According to [14], there is a scarcity of empirical studies that examine the DW success. Therefore, this paper investigates the extent to which Jordanian firms implemented DW and the benefits for implementing DW. It is aimed at providing empirical evidence, thereby extending the body of research regarding the implementation of DSSs in general and DW in particular.

## VI. Results, Findings and Discussion

As mentioned earlier, the study's sample consists of the Jordanian firms' top management. The results of the study's sample analysis are shown in the figure below (Figure 1). In this figure, the sample results are broken down by responses.

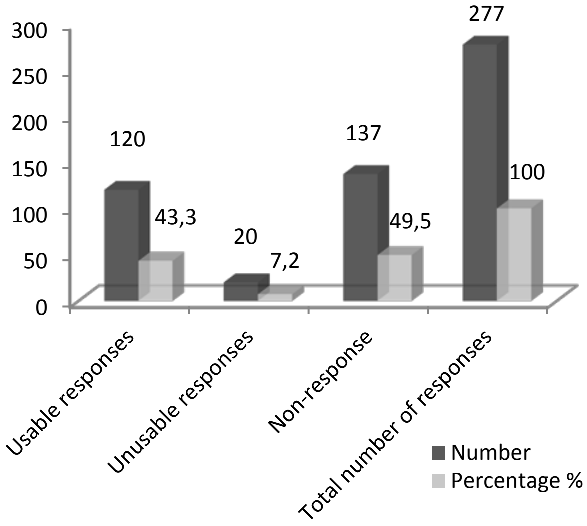


**Fig. 1.** The analysis of the study's sample (Source: the figures are based on the responses to the questionnaire).

As can be seen from these results, 140 completed questionnaires were returned to the researcher's address with a response rate of 50.5%. According to these figures, usable questionnaires accounts for 43.3% of the total sample. 20 of the questionnaires were discarded as unreliable, i.e. there were many essential questions missing from the questionnaires. To sum up, all the firms (277), which were listed on ASE at the time of data collection, were selected. 140 filled questionnaires were returned generating 50.5 %. This response rate somehow on average comparison to many similar studies such as ([69]; [70]).

The reliability, internal consistency and validity of the Likert scale questions are assessed by using Cronbach's alpha. It was found that Cronbach's alpha is the most popular method for assessing the reliability of scales.

It has been used by many researchers, including [71]. Cronbach's alpha determines the internal consistency of the items in a survey instrument (questionnaire) to assess its reliability. Figure 2 shows Cronbach's alpha for the Likert-scale question. In addition, the figure demonstrates the mean, SD, sum of item variances (V) and standard error (SE) for the rating scale question. As can be seen from these figures, Cronbach's alpha coefficient is more than 0.9.

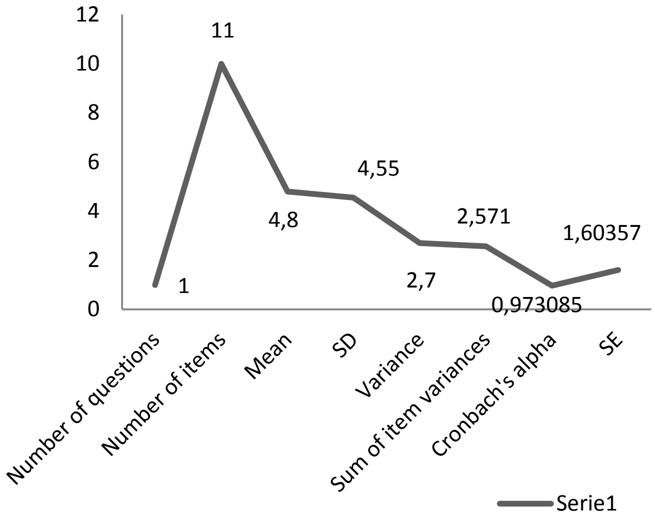


**Fig. 2.** The statistical analysis of reliability and validity (Source: the figures are based on the responses to the questionnaire).

The respondents, who their firms implemented DW (42 respondents), were solicited for their opinions regarding the actual benefits (which their firms have realized) of implementing the DW. The likely and expected benefits, which were measured on a seven-point scale ranging from 1 (strongly disagree) to 7 (strongly agree), are shown in table (1).

The table shows the mean and SD for all the select benefits. As can be seen from these results, the role of DW's information in enhancing and facilitating the process of decision-making (mean=6.38) was the most important benefit of implementing the DW in the Jordanian firms involved. From the data in the above table, it is apparent that the improvement to decision-making process is the second most widely mentioned benefit of implementing DW.

Other important benefits of implementing DW include more accurate, useful, reliable, complete and relevant information to decision-makers (mean= 6.28), helping decision-makers in taking fact-based decisions (mean= 6.26) and more efficient and successful decisions (mean= 6.23). Moreover, the figures show clearly that the DW has provided many benefits, including frequent, accessible and timely reports and information to decision-makers (mean=6.11), improved the quality of decisions (mean=6.04), easy access to data and information (mean=6.04) and DW's information is used as a basis for decision-making. According to these figures, Increase in competitive capability (mean=5.57) and better performance measurement (mean= 5.54) were the least important benefits of implementing DW.

TABLE I.
THE BENEFITS OF IMPLEMENTING DW

| Benefits of implementing DW | No | Min. | Max. | Mean | SD |
|---|---|---|---|---|---|
| DW has improved the quality of the decisions | 42 | 4 | 7 | 6.04 | 1.01 |
| DW's information has helped decision-makers in taking fact-based decisions | 42 | 4 | 7 | 6.26 | 0.83 |
| Better performance measurement | 42 | 3 | 7 | 5.54 | 1.26 |
| Easy access to data and information | 42 | 5 | 7 | 6.04 | 0.79 |
| DW has provided more accurate, useful, reliable, complete and relevant information to decision-makers | 42 | 5 | 7 | 6.28 | 0.81 |
| DW has provided frequent, accessible and timely reports and information to decision-makers | 42 | 4 | 7 | 6.11 | 0.99 |
| DW's information is used as a basis for decision-making in your firm | 42 | 4 | 7 | 6.02 | 0.92 |
| Increase in competitive capability | 42 | 1 | 7 | 5.57 | 1.67 |
| Improvement to decision-making process | 42 | 5 | 7 | 6.31 | 0.71 |
| DW has led to more efficient and successful decisions | 42 | 4 | 7 | 6.23 | 0.93 |
| DW's information enhances and facilitates the process of decision-making | 42 | 5 | 7 | 6.38 | 0.76 |

Source: The figures are based on the responses to questionnaire.

In conclusion, the figures clearly lead to the conclusion that the select benefits are highly rated. These results are entirely consistent with other earlier research on DW. For example, [10], who investigated the adoption of DW in Small and Medium Enterprises (SME) In Zimbabwe, found that DW provides a variety of benefits, including the reduction in the overall effort concerning data analysis and reporting, improvements to the reports' quality, more flexible reaction to new information needs and improvement to business decisions through more precise as well as more current data analyses.

Park and Kim [11], who presented a DSS for the management of sewer infrastructure using DW technology, found that the managers, by using DW, could significantly reduce the burden of collecting unnecessary information, easily retrieve managerial information and achieve effective data management. Consequently, the proposed DSS could provide flexible storage of valuable information, reduce subjectivity in decision-making processes, accelerate the pace of data flow, improve the consistency in decision-making processes (consistent decisions), and better environment for strategic operation of infrastructure.

## VII. CONCLUSION

The evidences, which are obtained by analyzing the data from the questionnaires, reveal exceptionally remarkable facts, first of all and to some extent, the success of implementing DW in Jordanian firms was a direct result of the benefits which were reaped from using the DW by the managers of Jordanian firms. One consequence of implementing DW was the great role of DW's information in enhancing and facilitating the process of decision-making. The results also showed that the Jordanian firms benefited greatly from implementing DW: These benefits reflected well on the process of decision-making. The results also revealed that the DW is a fruitful source of information. Moreover, the implementation of DW proved to be a success through helping decision-makers in taking fact-based decisions. Based on some of the study's findings, it can be concluded that the implementation of DW did provide frequent, accessible and timely reports and information to decision-makers and easy access to data and information. Furthermore, the DW was lauded by the users for the successful use of its information as a basis for decision-making.

This study has humbly contributed to the field of scientific research in general and the field of decision support systems (DSS) in particular in many ways, first of all, the studies on the implementation of DW were nearly all in developed countries. This study was applied in Jordan, therefore, the results of this study made a humble contribution to the existing knowledge in the field of implementing DW worldwide in general and in Jordan in particular.

Second, there is a need for evidences that the DW improves the quality and accessibility to information. Therefore, this study practically investigated whether or not the implementation of DW improves the quality and accessibility to information and facilitates the decision-makers' tasks. Lastly, Based on extensive literature review, the researchers have identified that firms are often unsuccessful due to a lack of appropriate information. For this reason, this study is one of the few empirical studies which have attempted to examine the effect of DW on decision effectiveness. In addition, previous research has not empirically tested its effectiveness in DSS contexts in Jordanian firms "to the researchers' knowledge".

Despite the usefulness and positive contributions of the study's results, these results should be treated and interpreted with caution. In fact, the study's sample included only the Jordanian firms which are listed on ASE. As a consequence, this might severely restrict the generalization of the results. It is believed that the results of this study might have been dissimilar, if all Jordanian firms have been surveyed. Therefore, prospective researchers are recommended to broaden the scope of their investigation to include all Jordanian firms.

## REFERENCES

[1] P. Lehmann, and J. Jaszewski, Business terms as a critical success factor for data warehousing, Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'99) Heidelberg, Germany, 1999.

[2] K.W. Chau, Y. Cao, M. Anson, and J. Zhang,, Application of data warehouse and Decision Support System in construction management, Automation in Construction, 12, 2002, pp. 213– 224.

[3] R. L. Hayen, C. D. Rutashobya, and D. E. Vetter, An investigation of the factors affecting data warehousing success, Issues in Information Systems, Vol. 8, No. 2, 2007, pp. 547-53.

[4] M. Mohania, S. Samtani, J. F. Roddick, and Y. Kambayashi, Advances and Research Directions in Data-Warehousing Technology, AJIS, Vol. 7, No. 1, September 1999, pp. 41-59.

[5] J. Ang, and T. S. H. Teo, Management issues in data warehousing: insights from the Housing and Development Board, Decision Support Systems, 29, 2000, pp. 11–20.

[6] B. Shin, An exploratory investigation of system success factors in data warehousing, Journal of the Association for Information Systems, Vol. 4, 2003, pp. 141–170.

[7] Y.-T. Park, An empirical investigation of the effects of data warehousing on decision performance, Information & Management, 43, 2006, pp. 51–61.

[8] D. Mukherjee, and D. D'Souza, Think phased implementation for successful data warehousing, information systems management, Spring 2003, pp. 82-90.

[9] S. Nilakanta, K. Scheibe, and A. Rai, Dimensional issues in agricultural data warehouse designs, computers and electronics in agriculture, 60, 2008, pp. 263–278.

[10] W. Mtembo, , F. Madzikanda, and T. Musiiwa, An Examination of the Benefits and Challenges of Data Warehouses Adoption in SMEs of Zimbabwe, International Journal of Management & Business Studiem, Vol. 3, Issue 2, April – June, 2013, pp. 99-100.

[11] T. Park, and H. Kim, A data warehouse-based decision support system for sewer infrastructure management, Automation in Construction, 30, 2013, pp. 37–49.

[12] G. Shankaranarayanan, and Y. Cai, Supporting data quality management in decision-making, Decision Support Systems 42, 2006, pp. 302–317.

[13] A. Rudra, and E. Yeo, Issues in User Perceptions of Data Quality and Satisfaction in Using a Data Warehouse - An Australian Experience, Proceedings of the 33rd Hawaii International Conference on System Sciences, IEEE 2000, pp. 1-7.

[14] F. Hegazy, and K. Ghorab, The impact of system support on adoption & diffusion of data warehousing success, 2003, http://www.hicbusiness.org/biz2003proceedings, accessed 31/08/2011.

[15] M. D. Solomon, Ensuring a successful data warehouse initiative, information systems management, Vol. 22, Issue 1, December 2005, pp. 26-36.

[16] A. Aljanabi, A. Alhamami, and B. Alhadidi, Query Dispatching Tool Supporting Fast Access to Data Warehouse, The International Arab Journal of Information Technology, Vol. 10, No. 3, May 2013, pp. 269-275.

[17] M. I. Hwang, and H. Xu, The Effect of Implementation Factors on Data Warehousing Success: An Exploratory Study, Journal of Information, Information Technology, and Organizations, Vol. 2, special section 2007, pp. 1-14.

[18] B. H. Wixom, and H. J. Watson, An empirical investigation of the factors affecting data warehousing success, MIS Quarterly, Vol. 25, No. 1, 2001, pp. 17–41.

[19] E. Gimzauskiene, and L. Valanciene, Efficiency of Performance Measurement System: The Perspective of Decision Making, economics and management, 15, 2010, pp. 917-923.

[20] S. A. Mansouri, D. Gallear, and M. H. Askariazad, Decision support for build-to-order supply chain management through multiobjective optimization, International Journal of Production Economics, 135, 2012, pp. 24–36.

[21] J. P. McKenna, Moving Toward Real-Time Data Warehousing, business intelligence Journal, Vol. 16, No. 3, 2011, pp. 14-19.

[22] N. Au, E. W. T. Ngai, and T. C. E. Cheng Extending the Understanding of End User Information Systems Satisfaction Formation: An Equitable Needs Fulfillment Model Approach, MIS Quarterly, Vol. 32, Issue 1, 2008, pp. 43-66.

[23] N. Rasmussen, P. S. Goldy, and P. O. Solli Financial Business Intelligence Trends, Technology, Software Selection, and Implementation, John Wiley and Sons, Inc., New York, 2002.

[24] H. J. Watson, and B. J. Haley, Data warehousing: A framework and survey of current practices, Journal of Data Warehousing, Vol. 2, No. 1, 1997, pp.10-17.

[25] S. Gatziu, and A. Vavouras, Data Warehousing: Concepts and Mechanisms, Informatik, Informatique 1, 1999.

[26] N. F. Doherty, and G. Doig, The role of enhanced information accessibility in realizing the benefits from data warehousing investments, Journal of Organizational Transformation and Social Change, Vol. 8, No, 2, 2011, pp. 163-182.

[27] C. Ballard, D. Herreman, D. Schau, R. Bell, E. Kim, and A. Valencic, Data Modeling Techniques for Data Warehousing, International Business Machines Corporation(IBM Corp), 1st edition,1998.

[28] H. J. Watson, C. Fuller and T. Ariyachandra, Data warehouse governance: best practices at blue cross and blue shield of North Carolina, Decision Support Systems archive, Vol. 38, Issue 3, December 2004, pp. 435 – 450.

[29] I. Ahmad, and S. Azhar, "Data Warehousing in Construction: From Conception to Application," Proceedings of the First International Conference on Construction in the Twenty First Century, Miami, Florida, USA, April 2002.

[30] B. List, R. Bruckner, K. Machaczek, and J. Schiefer, A Comparison of Data Warehouse Development Methodologies - Case Study of the Process Warehouse, DEXA, Munich, 2002.

[31] H. R. Nemati, D. M. Steiger, L. S. Iyer, and R. T. Herschel, Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing, Decision Support Systems, Vol. 33, Issue 2, June 2002, pp. 143–161.

[32] M. V .Mannino, S. N. Hong, and I. J. Choi, Efficiency evaluation of data warehouse operations, Decision Support Systems, Vol. 44, No. 4, 2008, pp. 883-898.

[33] B. Bębel, J. Eder, C. Koncilia, T. Morzy, and R. Wrembel, Creation and management of versions in multiversion data warehouse, Proceedings of the 2004 ACM symposium on Applied computing, SAC 2004, March 14-17, Nicosia, Cyprus, pp. 717-723.

[34] T. Brown, Data Warehouse Implementation with the SAS System, SAS Institute Inc., Dallas, TX, 1996, http://www2.sas.com/proceedings/sugi22/DATAWARE/PAPER132.PDF.

[35] K. R Ramamurthy, A. Sen, and A. P. Sinha, An empirical investigation of the key determinants of data warehouse adoption, Decision Support Systems, 44, 2008, pp. 817–841.

[36] S. Nilakanta, K. Scheibe, and A. Rai, Dimensional issues in agricultural data warehouse designs, computers and electronics in agriculture, 60, 2008, pp. 263–278.

[37] S. Graham, The Foundations of Wisdom: A Study of the Financial Impact of Data Warehousing, International Data Corporation (Canada) Ltd, 1996.

[38] P. Vassiliadis, C. Quix, Y. Vassiliou, and M. Jarke, Data Warehouse Process Management, Information Systems, Vol. 26, No. 3, June 2001, pp. 205-236.

[39] J. Chmiel T. Morzy, and R. Wrembel, Multiversion join index for multiversion data warehouse, Information and Software Technology archive, Vol. 51, Issue 1, January 2009, pp. 98-108.

[40] R. Hackathorn, Current Practices in Active Data Warehousing, Bolder Technology, Inc., 2002.

[41] H. Watson, and B. Haley, Managerial Considerations, In Communications of the ACM, Vol. 41, No. 9, September 1998, pp. 32-37.

[42] H. J. Watson, D. Goodhue, and B. H. Wixom, The benefits of data warehousing: why some organizations realize exceptional payoffs, Information & Management, 2001 (a), pp. 1–12.

[43] H. Watson, T. Ariyachandra, and Jr, R. J. Matyska, Data Warehousing Stages of Growth, Information Systems Management, Vol. 18, Issue 3, June 2001 (b), pp. 42 – 50.

[44] J. D. Wells, and T. J. Hess, Understanding decision-making in data warehousing and related decision support systems: An Explanatory

Study of Customer Relationship Management Application, Information Resources Management Journal, Vol. 15, No. 4, October-December 2002, pp. 16-32.

[45] H. A. Abdel Hafez, and S. Kamel, Web Based Data Warehouse in the Egyptian Cabinet Information and Decision Support Center, Decision Support in an Uncertain and Complex World: The IFIP TC8/WG8.3 International Conference, 2004, pp. 402-409.

[46] D. L. Rubin, and T. S. Desser, A Data Warehouse for Integrating Radiologic and Pathologic Data, Journal of the American College of Radiology, Vol. 5, No. 3, March 2008, pp. 210-217.

[47] P. K. Mawilmada, Impact of a data warehouse model for improved decision-making process in healthcare. Masters by Research thesis, Queensland University of Technology, October 2011.

[48] Á. Ojeda-Castro, M. Ramaswamy, A. Rivera-Collazo, and A. Jumah, Critical Factors For Successful Implementation Of Data Warehouses, Issues in Information Systems, Vol. 12, No. 1, 2011, pp. 88-96.

[49] R. Alshboul, Data Warehouse Explorative Study, Applied Mathematical Sciences, Vol. 6, No. 61, 2012, pp. 3015– 3024.

[50] L. Shen, S. Liu, S. Chen, and X. Wang, The Application Research of OLAP in Police Intelligence Decision System, Procedia Engineering 29, 2012, pp. 397 – 402.

[51] K. Shams, and M. Farishta, Data warehousing: toward knowledge management, Topics in Health Information Management, Vol. 21, No 3, February 2001, pp. 24-32.

[52] T. Chenoweth, K. Corral, and H. Demirkan, Seven Key Interventions for data warehouse success, Communications of the ACM, Vol. 49, No. 1, January 2006, pp. 115-119.

[53] W. H. DeLone, and E. R. McLean, Information systems success: the quest for the dependent variable, Information Systems Research, Vol. 3, No. 1, 1992, pp. 60–95.

[54] R. L. Kumar, Justifying Data Warehousing Investments, in Data Warehousing and Web Engineering, Shirley Becke (Ed.), 2002, pp. 100-102.

[55] H. J. Watson, J. G. Gerard, L. E. Gonzalez, M. E. Haywood, and D. Fenton, Data warehousing failures: case studies and findings, Journal of Data Warehousing, Vol. 4, No. 1, Spring 1999, pp. 44– 55.

[56] D. Sammon, F. Adam, and F. Carton, Benefit Realisation through ERP: The Re-Emergence of Data Warehousing, Electronic Journal of Information Systems Evaluation, Vol. 6, Issue 2, 2003, pp. 155-16.

[57] M. D. Aguila, and E. Felber, Data Warehouses and Evidence-Based Dental Insurance Benefits, Journal of Evidence Based Dental Practice, Vol. 4, Issue 1, 2004, pp. 113-119.

[58] Devlin, B. Data Warehouse from Architecture to Implementation, Addison Wesley Longman, Inc., 1997.

[59] R. K. Griffin, Data warehousing, Cornell Hotel and Restaurant Administration Quarterly, Vol. 39, No. 4, 1998, pp. 28–40.

[60] N. Wilkoff, T. Pohlmann, R. Hudson, and N. Lambert, The State Of Technology Adoption, Business Technographics North America, May 5 2004, Forrester Research, Inc.

[61] L. Agosta, Hub-and- Spoke Architecture Favored, DM Review, Vol. 15, Issue 3, March 2005, pp. 14-63.

[62] H.-G. Hwang, C.-Y. Ku, D. C. Yen, and C.-C. Cheng, Critical factors influencing the adoption of data warehouse technology: a study of the banking industry in Taiwan, Decision Support Systems, 37, 2004, pp. 1–21.

[63] S. Hong, P. Katerattanakul, S.-K. Hong, and Q. Cao, Usage and perceived impact of data warehouses: a study in Korean financial companies, International Journal of Information Technology & Decision Making, Vol. 5, No. 2, 2006, pp. 297–315.

[64] L. Agosta, M. Andrews, and M. Ritzmann, The Data Warehouse Satisfaction Survey, Part 1: The Number One Complaint About Data Warehousing, Information Management Special Reports, October 2 2007.

[65] K. L. Merritt, User Satisfaction In Data warehousing: An Empirical Investigation Of Salient Variables, Issues in Information Systems, Vol. 9, No. 2, 2008, pp. 500-508.

[66] B. H. Wixom, H. J. Watson, A. M. Reynolds, and J. A. Hoffer, Continental Airlines Continues to Soar with Business Intelligence, Information Systems Management, 25, 2008, pp. 102–112.

[67] A. Almabhouh, A. R. Saleh, and A. Azizah, Examining the Influence of Relationship Quality on Data Warehouse Success, International Journal of Modeling and Optimization, Vol. 1, No. 5, December 2011, pp. 402-409.

[68] M. G. Lodico, D. T. Spaulding, and K. H. Voegtle, Methods In Educational Research From Theory to Practice, John Wiley & Sons, Inc., San Francisco, CA, 2006.

[69] B. AL-allak, Evaluating the Adoption and Use of Internet-based Marketing Information Systems to Improve Marketing Intelligence (The Case of Tourism SMEs in Jordan), International Journal of Marketing Studies, Vol. 2, No. 2, November 2010, pp. 87- 101.

[70] A. Al Khattab, The Role of Corporate Risk Managers in Country Risk Management: A Survey of Jordanian Multinational Enterprises, International Journal of Business and Management, Vol. 6, No. 1, January 2011, pp. 274-282.

[71] W. Chongruksut, the adoption of activity-based costing in Thailand, doctoral thesis, Faculty of Business and Law, Victoria University, 2002.

[72] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, AIMQ: a methodology for information quality assessment, Information & Management 40, 2002, pp.133–146.

[73] A. S. Hardan, and T. M. Shatnawi, Impact of Applying the ABC on Improving the Financial Performance in Telecom Companies, International Journal of Business and Management, Vol. 8, No. 12, 2013, pp.48-61.

[74] S. B. Gerber, and K. V. Finn, Using SPSS For Windows Data Analysis and Graphics, 2nd Edition, Springer Science Business Media, Inc., 2005.

[75] D. George, and P. Mallery, SPSS for Windows Step-by-Step: A Simple Guide and Reference, 14.0 update, 7th Edition 2006, Allyn & Bacon.

[76] Owoc M. L., Alsqour M., Abdulrhman S. A.: Data Warehouse as a Source of Knowledge Acquisition. An Empirical Study. Proc. of the 2014 Federated Conference on Computer Science and Information Systems, Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki (eds.), pp. 1421-1430, ISSN 2300-5963

DOI: 10.15439/2015F343
ACSIS, Vol. 5

# A concept of enterprise Big Data and BI workflow driven platform.

Maciej Pondel
Wroclaw University of Economics
ul. Komandorska 118/120, 53-345
Wrocław, Poland
Email: Maciej.Pondel@ue.wroc.pl

*Abstract*—**This paper describes the author's concept for universal enterprise platform that allows to benefit from modern approaches to data analysis based on Business Intelligence and Big Data technologies. This idea is based on SOA architecture that enables workflow to coordinate the business processes and communicate domain applications with bespoke Business Intelligence and Big Data Solutions. Whole platform is designed the way that can be provided in a cloud environment.**

## I. Introduction

IN current times business requires scalable IT platforms supporting whole business processes from the beginning till the successful finish. Due to a huge complexity of the processes they require very often several domain applications like ERP, CRM, DMS and many more. In most cases those processes are handled manually and need human interaction between various IT Systems. Such approach leads to mistakes or inconsistencies in processes. Business Intelligence systems and Big Data solutions in current world provide essential information and knowledge for managers on the strategic or tactical level. The most important challenge defined in this paper is to take advantage of them also on the operational level of a business.

## II. Big data and Business Intelligence

There are several definitions of Big Data and Business Intelligence solutions. Some of them emphasize that Big Data is a modern concept replacing older approach defined as Business Intelligence [1], [2].

We can also find the approach stating that Big Data resolves new problems, brings new opportunities and meets new challenges that are different or complementary to the known Business Intelligence and Data Warehouses areas of usage [3], [4].

Author of this paper claims that both Big Data and Business Intelligence solutions meets similar problems when we look at it from the very general perspective. In details they have slightly different foundations and some problems could be better satisfied with multidimensional approach based on data warehouse and Business Intelligence solutions (OLAP, reporting, querying) and some problems are better satisfied with Big Data approach (NoSQL data structures, map reduce parallel processing).

We have to remember that Business Intelligence is no only a technology or a set of IT tools. It is considered a management strategy used to create a more structured and effective approach to decision making [5]. Of course, to fulfill this strategy, we need IT tools providing an access to data and capabilities allowing analysis and presentation of information. We can consider traditional BI based on ETL Process, data warehouses, data marts, OLAP, dashboards, scorecards and analytics. The Business Intelligence approach is still evolving. We can mention Business Intelligence 2.0 that includes following Features [5]:

- Proactive alerts and notifications.
- Event driven/ real time/ instant access to information.
- Advanced analytics.
- Enterprise Integration.
- Mashups and portal integration.
- Mobile/ Ubiquitous access.
- Improved visualization, Rich Interfaces (RIA).
- BI as a service (SOA and SaaS)
- In-memory analytics
- Open Source BI.

We can also meet (mainly on the BI Vendors websites) the term Business Intelligence 3.0. Author of this paper couldn't find a consistent scientific definition of the term. Different sources define Business Intelligence 3.0 as another evolution of BI systems that focus on the user interface, that is so intuitive to the regular user, that little or no training is necessary to explore data. We can list the following terms that appear when we consider this new approach: Data Discovery, Advanced Visualization, Visual Analytics, Business Discovery, Self-Service Business Intelligence. Example tools delivered by the top vendors are [7]:

- Visual Insight by MicroStrategy,

978-83-60810-66-8/$25.00©2015, IEEE

- Visual Intelligence by SAP,
- Visual Analytics by SAS,
- PowerPivot, Power View Power BI by Microsoft,
- Cognos Insight by IBM,
- Endeca by Oracle.

Also, it's important to highlight that there are two kinds of self-serve BI user:

- Analytics Power Users who create visual apps from multiple data sources – both internal and external.
- Regular Users that can fully explore the visual apps created by power users or IT.

Literature also mentions that BI 3.0 focuses on mobile devices. Mobile devices are here treated as the interface of data visualization but also as a source of the data together with other sensor-based Internet-enabled devices equipped with RFID, barcodes, and radio tags (the "Internet of Things") [8].

According to the quoted key factors of Business Intelligence 3.0 we can state that it is not a completely new concept but more development of Business Intelligence 2.0 with emphasized selected features. In some cases BI 3.0 concentrates on the same data sources as Big Data approach.

From the author's point of view the main difference between Business Intelligence and Big Data concept regards the following:

- Business area: Business Intelligence tools and Warehouses are concentrated on business transactions (sales, costs, money transfers) [13] and Big Data stores the records describing all the activities leading to the transaction or leading to the decision about transaction (tracking the customer path in on-line or off-line shop, social activity)
- Data quality – in most cases data in warehouses come from transactional databases that are consistent, stable and there is no room for any errors. In case of Big Data systems that are loaded very often with social content, by devices (Internet of Things) or server logs, the quality of data can be lower. In any case the lower data quality doesn't contradict the value of Big Data solutions. If there is an error in the data describing transaction in data warehouse it can result with the serious problems in financial / tax reports. The outcome of the missing or incorrect items the Big Data system usually do not affect the business so much (of course it the wrong data do not represent the majority of data).
- Data structure – the structure of data warehouses is based on relational structures and usually are denormalized multidimensional star or snowflake structure. Basing on such structure we can easily build the OLAP cubes. Big Data system store data in non-relational databases.

## III. Workflow Based platform

Workflow systems are considered mainly as tools supporting business processes. A workflow application implements a business process model. The model describes the process steps to be performed to achieve a specific business goal, business rules for coordination of those steps and responsibilities of process participants[9]. The steps include tasks that should be performed by agents that can be human, computer systems or combination of both [10]. Workflow systems, with the benefits of efficient and flexible process modelling and process automation, have been widely used for managing business processes [11]. The main advantages of using workflow platforms are:

- Ability to model and deploy processes by a business user without engagement of software developers and deployment procedures.
- Ability to measure the performance of the processes and the whole business area related to the process.
- Full control over processes. Every step must be undertaken during the process and whole process (there is no possibility to forget some activity) has to be performed according to the business rules (no risky shortcuts allowed). Workflow makes processes less dependent on users mistakes and faults.
- Workflow can automate both business critical processes and also those back-office and automation makes processes to consume less time and resources.

Workflow activities execute programs that consume and produce data (parameters values and files). An output data produced by an activity can be consumed as input data to another activity, establishing a dependency relation between those activities [12].

## IV. Concept of BI and Big Data platform that is workflow driven

Being aware that steps in workflow produce some data it seems natural that they could be connected directly to the data storage such us data warehouse or big data repositories. Such approach can change the current well known and stable ETL approach (Extract Transform and Load).

Regarding to the consumption data by workflow activities we can assume that beside the data produced on the earlier stages of the process, the activity can be also fed with the data or information coming from data warehouses or even knowledge produced by a knowledge discovery modules of both BI and Big Data Systems.

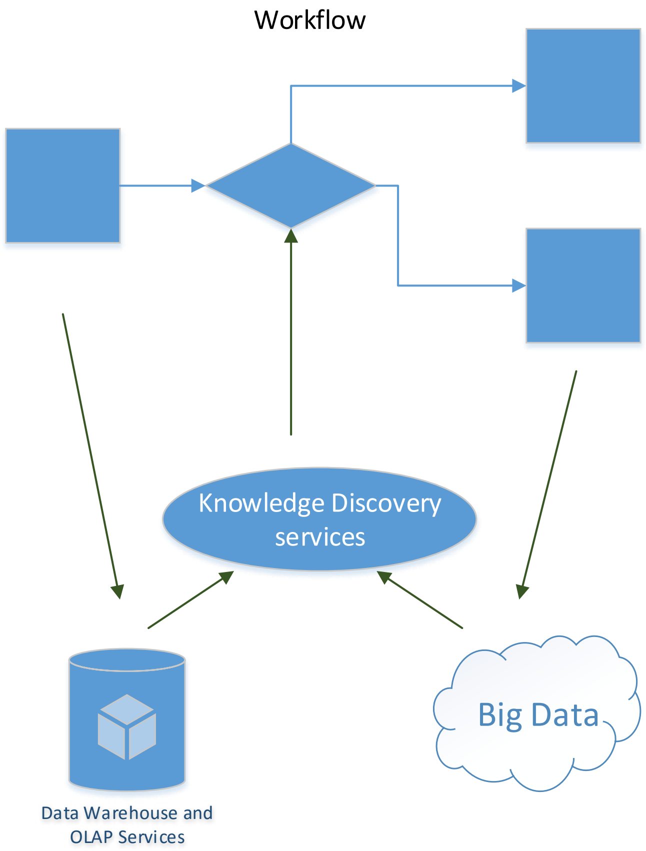

Workflow
Knowledge Discovery services
Big Data
Data Warehouse and OLAP Services

**Figure 1. Concept of workflow and Big Data and Warehouse communication**

Some key items of the system's architecture are visualized on the Figure 1. There we can business process supported by workflow tool. Some steps (decision blocks) can use the Knowledge acquired in Knowledge Discovery module. Other steps (actions) can load the data into:

- Data Warehouse
- Big Data repository

The compound layer called Knowledge Discover services is working on the collected data and in this layer the Data Mining models are prepared. We can use well the known methods of data discovery like:

- Artificial neural networks
- Machine learning
- Naïve Bayes networks
- Collaborative filtering
- Clustering
- Association rules

and any other that can produce the valuable conclusions in the business environment.

The most important assumption regarding this layer is accessibility to its actions through SOA protocols like SOAP and REST. Regarding the actions it is meant mostly executing the models on the real object to gather the classification or assign the object to the cluster. Author assumes that through the service model we could also launch the training, testing or validation processes.

In SOA architecture we can efficiently send data for the technology different:

- Workflow platform
- Data loading services
- Knowledge discovery services

Example of a workflow driven combination of such activities can be the sales process in an example company. During the process, the customer goes the path (live or virtual) among various products and stops to watch some of them. The information about user focusing on the product and making decision regarding putting it to the basket or not is a valuable item to store in big data repository. After the user makes some particular activity, we can ask the conclusion model (bulid on previous customers behaviors) if it is efficient to present the customer directed proposal (or any other marketing message). If the recommendation of conclusion model is positive we can communicate with customer by:

- sending him message on his mobile (using sms, mms or push notification in the bespoke mobile app),
- presenting him the message on the screen visible in the market,
- make the shop assistant to approach the customer and ask some more information and present the dedicated offer.

We can imagine many more ways the store is communicating the customer to make the offer more suitable for him and optimize the income and turnover from the customer service.

Of course the areas of usage of such platform are not limited only to sales and marketing processes, although it is probably the most efficient and most obvious to use data mining based on big data and Business Intelligence modules there.

The important remark regarding the proposal of workflow using Big data and BI states that such workflow can automate nearly every process in the company (both Front Offices and Back Office as well). Workflow is able to communicate not only with Knowledge Discovery services but also with any application working in the enterprise environment that is build using Service Oriented Architecture approach.

## V. Project of Platform Implementation in Enterprise Environment

Due to various business processes and the logic of existing IT environment in enterprises we cannot assume that the implementation of proposed platform may be done in a standard way and be processed easily. Of course whole platform will contain a set of coupled services, ready to cooperate, but they all will need to be integrated with enterprise data sources and configured to fulfill the business requirements.

The project of the platform implementation will contain the following stages:

1. Identification of business goals. The processes or business areas of efficiency improvement must be defined. We can imagine a whole variety of platform usage like:
    a. Sales processes
    b. Customer relation and loyalty management
    c. Logistic processes
    d. Back office processes

    And many more
2. Information and data sources identification. We should investigate the possible sources of data required to acquire sufficient knowledge useful in chosen processes improvement. In current approach the data may come from:
    a. The classic electronic sources like:
        i. IT systems databases.
        ii. Server logs describing usage of the IT tools.
        iii. Search engines' queries corresponding to the knowledge flows and users' needs.
        iv. Social media activities of internal users and external individuals (customers, suppliers, fans, competitors). Social systems are now full of interesting information. We can distinguish 2 ways of social systems:
            1. Open – like Facebook, Twitter, Youtube, LinkedIn and many more
            2. Internal – available only for internal users or intentionally invited externals – like Yammer, HighQ, eXo Platform and many more.
    b. The electronic sources emerging especially to feed the database of implemented platform. They can register activities of individuals that have not been registered yet eg.:
        i. Sensors identifying paths of customers movement around market or employees in a warehouse.
        ii. Barcode scanners, RFID tags / readers.
        iii. Mobile devices (tablet/smartphones) equipped with applications to register some facts that have not been registered previously (Customer assistant may register the conclusion of conversation with customer eg. Customer has just bought a house and soon he will need pieces of furnishings or the logistics employee may calculate the time every activity takes and register that)
        iv. Stationary machines allowing customers to give a feedback or provide a satisfaction survey in a simple way.
3. Verification of individuals acceptance for registering the data describing individuals and their behavior. If customers feel surveilled they can avoid the services or the company that implemented platform. Internal employees who are aware of continuous tracking may feel strange and not work fully efficient. Another important aspect is validation of law and regulatory compliance. Personal data is protected by the law and we should verify if the data we gather can be processed without breaking the law.
4. Data load processes design. We can distinguish in here:
    a. ETL (Extract, Transform and Load) processes – mainly for data coming from traditional electronic sources.
    b. Services design (SOAP/REST) – mainly for data describing events that were when event occurred. In such approach the database is filled immediately

5. Data structures design that will include:
    a. Data warehouse structure in the form of multidimensional structure (star or snowflake). Basing on the structure we can build OLAP cubes structure or we can process a data mining models.
    b. Non – relational database that can be much more efficient while storing the behavioral data. We can distinguish in here:
        i. Column database like Cassandra or HBase
        ii. Document database like MongoDB or HyperDex
        iii. Key-value like FoundationDB or CouchDB

    NoSQLdatabases are more efficient for huge data sets processing and the performance of data loading is much higher.
6. Validation of the data designed structure especially it's adaptation to the business goals that were discovered on the earlier stage. Verification of the data quality must also be made.
7. To build the knowledge discovery services it is crucial to choose the correct data exploration

methods that fits best to business problem solutions. The huge topic is extraction of the algorithms parameters and their adjustment for models to generate the best possible results.

8. Since the main concept is based on the fact, that business processes using knowledge services and feeding them with data should be supported by workflow tool – it is important on the design stage of the project to model the business processes and automate them in the workflow software.

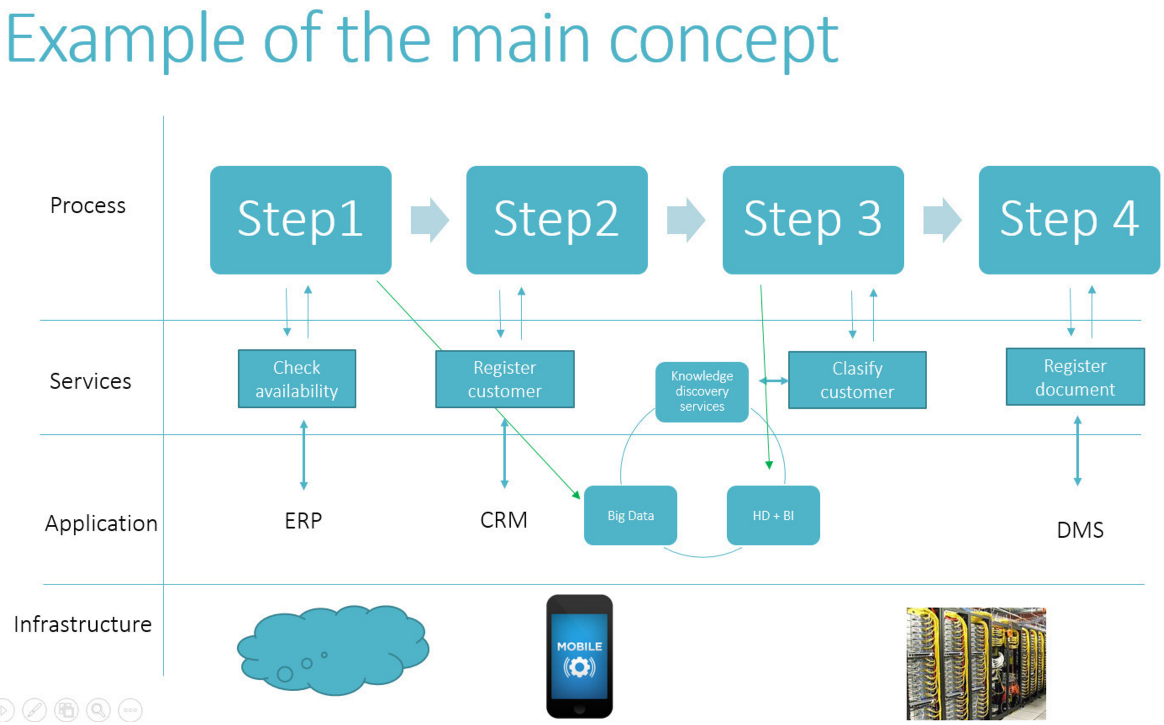
## VI. Platform Structure Assumptions

In the current state the platform is technology independent. The necessary components like:

- Data warehouse
- NoSQL databases
- Services (SOAP/REST) protocols
- ETL tools
- Workflow Engines
- Data exploration algorithms

have implementation in various operating systems and software development platforms, that is why from today's perspective it is not crucial to select the final technology.

Figure 2 . presents the example of the concept platform architecture basing on real life examples of use cases.



Example of the main concept

**Figure 2. Example of concept platform**

## VII. Challenges

Without a doubt, implementation of presented platform is a risky project and includes a lot of challenges. Among them we can distinguish:

- Performance – the knowledge discovery services and intended to respond on-line for process requests. If there is any significant delay – it may harm the efficiency of core business processes.
- Data mining models' relevancy – it is crucial to acquire valuable knowledge form data and provide it to the business scenarios supported by presented platform.
- Business efficiency – implementation of the platform is high cost affair. If we choose wrong processes or wrong methods of knowledge discovery and usage - the return of investment bay be not rewarding
- Stability of the platform - some of the technologies / solutions intended to be used in the platform are modern and in some cases may be unstable. It may influence the stability of the whole platform.

## VIII. Conclusions

The article presents modern approach for building enterprise platform, that could be used in majority of enterprises. The advantages of such approach are the expected high performance and most efficient usage of analytical platform (BI and Big Data) not only on a strategic level as management support systems but also on an operational level where business processes supported by knowledge can be most efficient from a business perspective. The future researches will concentrate on choosing example technologies and implementation of pilot atomic use cases to prove the correctness of proposed approach.

## References

[1] Tabakow M., Korczak J., Franczyk B., "Big Data – Definicje, Wyzwania I Technologie Informatyczne" In Informatyka Ekonomiczna Business Informatics 1 (31) 2014, Publishing House of Wroclaw University of Economics, Wrocław 2014

[2] D. Che, M. Safran, and Z. Peng, "From Big Data to Big Data Mining: challenges, issues, and opportunities," in Database Systems for Advanced Applications, pp. 1–15, Springer, Berlin, Germany, 2013

[3] M. Chen, S. Mao, and Y. Liu, "Big data: a survey," Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[4] J. Manyika, C. Michael, B. Brown et al., "Big data: The next frontier for innovation, competition, and productivity," Tech. Rep., Mc Kinsey, May 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[5] Nelson G., Business Intelligence 2.0: Are we there yet? SAS Global Forum 2010 (http://support.sas.com/resources/papers/proceedings10/040-2010.pdf)

[6] Pondel M.: Business Intelligence as a service in a cloud environment, in: Proceedings of the 2013 Federated Conference on Computer Science and Information Systems / Ganzha Maria, Maciaszek Leszek, Paprzycki Marcin ( red. ), 2013, IEEE, ISBN 978-1-4673-4471-5, ss. 1269-1271

[7] Cabrio B. What Is Business Intelligence 3.0?, Strategic Analytics Blog, http://blog.strat-wise.com/2015/03/what-is-business-intelligence-30_2.html , 2015

[8] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS quarterly 36.4 (2012): 1165-1188.

[9] Schmidt, Marc-Thomas. "Building workflow business objects." Business Object Design and Implementation II. Springer London, 1998. 64-76.

[10] Demeyer, Romain, et al. "Declarative workflows to efficiently manage flexible and advanced business processes." Proceedings of the 12th international ACM SIGPLAN symposium on Principles and practice of declarative programming. ACM, 2010.

[11] Liu, Xiao, et al. "Managing large numbers of business processes with cloud workflow systems." Proceedings of the Tenth Australasian Symposium on Parallel and Distributed Computing-Volume 127. Australian Computer Society, Inc., 2012.

[12] Chirigati, Fernando, et al. "Evaluating parameter sweep workflows in high performance computing." Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies. ACM, 2012.

[13] Alsqour, M., Owoc, M. L., & Ahmed, A. S. (2014, September). Data warehouse as a source of knowledge acquisition. An empirical study. In Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on (pp. 1421-1430). IEEE.

DOI: 10.15439/2015F209
ACSIS, Vol. 5

# Knowledge transfer from agri-food scientific papers to a knowledge base

Rafał Trójczak, Robert Trypuz, Anna Mazurek, and Piotr Kulicki
The John Paul II Catholic University of Lublin
Department of the Foundations of Computer Science
Faculty of Philosophy
Al. Racławickie 14, Lublin, Poland
Email: trypuz@kul.pl

***Abstract*—We address the problem of the access to the results of scientific publications in the agri-food domain. We focus on the description of main contributions of the papers treating them as accepted or rejected beliefs of their authors expressed in the form of scientific laws. We define the structure of different kinds of scientific laws present in the domain in the form of an ontology. The main concern of the paper is a process in which we proceed from the abstracts of papers to the ontological representation of laws. Moreover, we present examples of SPARQL queries which show how the resulting knowledge base can be used. Among the uses we point out discovering new scientific hypotheses and incoherencies among scientific laws.**

***Index Terms*—ontology, scientific law, knowledge extraction, scientific hypotheses, incoherency**

## I. Introduction

In RECENT years the number of scientific publications has increased significantly. It is common knowledge in scientific community that there are so many works published every year that it is not possible for one person to read all of them, even if that person limits themselves to a narrowed specialization. The situation of professionals who are working at the crossroads of science and practice is even worse because they must not only struggle through countless scientific papers but they must also cope with the difficult scientific language in which these papers are written. This problem gets even more serious when it comes to scientific projects where research is being conducted in many loosely connected disciplines.

Some of the journals are already providing structured abstracts within the papers published there. One of the possible structures of such abstracts is to divide the abstract into three sections: background, results, and conclusion. And, additionally, there are keywords at the end[1] [1]. Another way to facilitate access to the content of a paper is using so called *highlights* which are “a short collection of bullet points that convey the core findings and provide readers with a quick textual overview of the article”.[2] There are also semantic solutions addressing this problem. For example, the authors of the Nanopublication and the Micropublications projects have created models to represent statements and argumentation from scientific papers. We shall discuss the projects in the section dedicated to related works.

The research presented in the paper was conducted within ProOptiBeef[3] project, oriented towards increasing the level of innovation in Polish beef sector. One of the tasks of the project is creating a knowledge base (henceforward KB) of recent results published in the top journals relevant to the domain of beef production and consumption. These results were selected by the community of the experts involved in the project. One of the main goals of constructing the KB is to help scientists decide which experiments should be conducted in order for the research to be innovative.

In order to transfer knowledge from the papers to our KB, we have created an ontology called Science. It is designed to represent proved and disproved statements extracted from scientific papers stored in the database of the ProOptiBeef project. We shall informally describe the Science ontology in section II of this paper. In section III we shall present the way in which we fill the KB with the information from scientific papers. The final element of our KB is a reasoning part. This element enables us to perform some reasoning algorithms on our KB in order to discover new scientific hypotheses and to detect incoherency among scientific laws which have already been represented. We shall present an implementation of three of these algorithms in the SPARQL query language in section IV. In section V we shall shortly describe related works. The last section, section VI, contains summary and describes our planned future work related to the project described in this paper.

## II. Ontology of scientific laws

As mentioned above, the project’s KB is structured according to Science ontology (see [2], [3] for the formal structure of the ontology and [4] for the philosophical sources of the ontology). The ontology is expressed in OWL and is composed of two parts: TBox and ABox. TBox (terminological component) consists of a taxonomy of scientific laws (see figure 1) and provides a formal characterization of each class of the

[1] See also http://www.nlm.nih.gov/bsd/policy/structured_abstracts.html. Journal of Biomedical Semantics can serve as an example.

[2] See: http://www.elsevier.com/journal-authors/highlights.

[3] Full name of the project: *ProOptiBeef – Optimizing beef production in Poland according to strategy “from fork to farm”*. Web page: http://prooptibeef.pl.

978-83-60810-66-8/$25.00©2015, IEEE

taxonomy[4]. Laws from different classes differ in their structure and therefore require different representation specified by the ontology. The top distinction in the taxonomy is between *quantitative and qualitative laws*[5]. Quantitative Law concerns the dependency between *qualities* of an entity that can be measured (e.g., time of (beef) aging influences (beef) pH value) while Qualitative Law concerns *perdurants* or *endurants* describing directly their properties or comparing them with respect to their qualities (e.g., meat of bulls housed in groups before slaughter was less tender than meat of bulls individually housed).

The meaning of terms "endurant", "perdurant" and "quality" comes from the DOLCE ontology [5] which is used as a foundational ontology for Science. The terms are to be understood as follows:

- endurants are wholes that endure in time, e.g., beef, pasture, food;
- perdurants are entities that "happen in time"; they can have temporal parts or spatial parts, e.g., feeding, slaughtering, cooking;
- qualities are entities that can be perceived or measured, e.g., height, weight, age, protein content.

Quantitative Law is further divided into Correlation Law and Functional Law. Correlation laws state that there is a correlation between the values of the qualities, but the details of the correlation are unknown, i.e. it is not possible to determine the function (direction) of it (e.g., there is a significant correlation between sarcoplasmic protein solubility and both expressible moisture and color parameters). Functional laws say that there is a dependency between the values of two qualities. One of the qualities is called an independent parameter and the other one is called a dependent parameter. A functional law states that the value of a dependent parameter is a function of the value of an independent parameter (e.g., aging influenced instrumental hardness). If the function describing dependency in a law is monotonic (i.e., it is either increasing or decreasing), this law is an instance of Monotonic Law (e.g., fat reduction results in higher cooking losses).

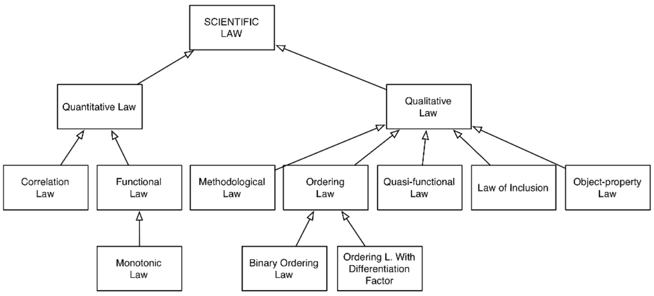


Fig. 1. The backbone of Science ontology.

There are five subcategories of Qualitative Law: Methodological Law, Quasi-functional Law, Law of Inclusion, Object-property Law and Ordering Law. A methodological law states that a method is used to measure a certain property (e.g. Multiple Linear Regression is used to measure beef tenderness). A quasi-functional law—similarly to a functional law—expresses the dependency between two qualities but one of the qualities is qualitative (e.g., the sex of an animal has influence on the tenderness of its meat – "sex" here is a qualitative quality). A law of inclusion states that one class of objects is a subclass of the other (e.g. meat color is a meat quality). An object-property law states that an object possesses a quality (e.g., red meat is an important source of vitamin D). Ordering Law is divided into two subcategories: Binary Ordering Law and Ordering Law with Differentiation Factor. Each instance of Ordering Law has some ordering basis according to which classes of objects are compared. Binary ordering laws compare two classes of objects (e.g., dairy cow lean has a longer display color life than beef cow lean) with respect to some quality playing the role of order basis (e.g., display color life). Each instance of Ordering Law with Differentiation Factor describes a change of some quality of an object after the object has undergone some modification (e.g., magnesium decreased in beef during cooking).

It is also worth noting that the term "scientific law" which is used by us in the paper covers proved and disproved statements described in scientific papers. Proved and disproved statements are instances of classes: Accepted Scientific Law and Rejected Scientific Law, respectively (see figure 2). It suggests that we treat scientific laws similarly to beliefs. Referring to [6] we assume that each scientific law has its "owners", who are the authors of the paper which the scientific law comes from. The paper is also a source in which the law is exemplified. So modelling author role and the paper as the context in which a scientific law appears is very important for our representation ([7], [8]).

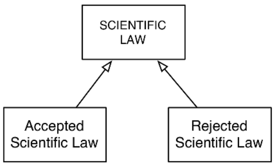


Fig. 2. Each scientific law is either accepted or rejected in a scientific paper.

ABox of our KB is simply a set of individuals (i.e., concrete scientific laws extracted from concrete scientific papers) which are represented in accordance to the TBox just described.

Below we shall describe the process of extraction of scientific laws from the scientific papers and their translation into the language of Science ontology.

[4] The TBox of Science ontology can be browsed here: http://onto.kul.pl/webprotege/.

[5] The way the laws are represented in Science has been widely described in [2], [3].

## III. Knowledge Acquisition

### *A. Extraction of knowledge from scientific papers by domain experts*

ABox contains a significant number of 8k scientific laws adequately representing the content of selected scientific articles. The selection of papers from the top journals—in the opinion of researchers involved in the ProOptiBeef—was prepared in the earlier phase of the project. Thus, the remaining task was to present the information from the selected papers in the form designed to represent scientific laws.

For an efficient realization of the task we decided to use abstracts of the papers as the main source of information. The UNESCO guide for the preparation of scientific papers for publication [9] states that "[t]he abstract should contain the results and conclusions of the paper in brief detail adjusted to the size allowed to the abstract and should, within these limits, refer to any new information which it contains. The abstract should not contain information or claims not contained in the body of the paper, nor should it include inessential details. [...] "*New information* should include observed facts, conclusions of an experiment or argument, the essentials of new methods or apparatus, etc." Moreover, the guide postulates that "[t]he abstract should be *self-contained*". In the light of these principles an abstract should contain enough information about the main results of the papers to reconstruct scientific laws.

Our experience shows that the practice in the journals from the agri-food domain in principle agrees with these postulates. However, authors use there technical vocabulary and discipline's folklore. That makes them difficult to be accessed by non-specialists. Thus, the participation of agri-food science specialists in the process of extracting knowledge from abstracts is inevitable.

On the other hand most specialists are not familiar with the formal and computational tools of knowledge representation. After a short introduction to our schema of the representation of scientific laws the specialists invited to the project were aware of the key conceptual distinctions (object, qualities) and the key elements of the representation of theses – the influences between factors and correlations. However, to encourage them to co-operate we had to allow them to present the results of their analyses in a form quite close to natural language with only some structural restrictions.

As a result we introduced a multi step process whose main steps are:

- agri-food specialists extract information from abstracts and write them down in a semi-structured way,
- information is automatically transferred to a database,
- the database is further elaborated by ontologists and (this is itself a multi step process described below)
- finally the information is automatically converted into OWL ontology.

Let us now trace the process in detail.

*a) STEP 1:* An abstract (earlier divided into sentences that are numbered for easier reference) of a scientific paper goes to the so called *Raiders of The Theses* (most of them are PhD students of broadly understood agri-food science). Their work is controlled and approved by senior members of the project team. The work with the abstract is based on a text form covering the following elements:

- a list of objects
- a list of properties (qualities)
- a thesis formulation from the abstract
- a new elaborated formulation of the thesis

The first two lists constitute a conceptual structure in which the laws presented in the article can be expressed. The description of the list of objects is usually rather complex and is a subject of further ontological elaboration. An example is presented later in the paper when we discuss the work of ontologists (see section III-B). The list of properties is rather simple and is directly transferred into a database where each quality is connected to the article and the thesis it comes from.

*Example 3.1:* Now let us concentrate on the formulation of theses and analyze an example. The original formulation from the abstract is as follows[6]: "W normalnej atmosferze, fotooksydacja była powierzchowna, ponieważ zaobserwowano odwrotną korelację między wagą plastra mięsa i stężenie COP na podłożu lipidowym, w przeciwieństwie do atmosfery bogatej w tlen (32%)." ("In normal atmosphere, photo[o]xidation was a superficial process, since an inverse correlation between meat slice weight and COPs content on a lipid basis was observed, unlike in a highoxygen (32%) atmosphere.")

The specialist disambiguated the formulation and divided it into elementary facts forming laws of one of the type from our ontology. The statement from our example is presented as the two following facts:

- "Wraz ze spadkiem masy plastrów surowej wołowiny, poddanych pakowaniu w atmosferze powietrza, zwiększa się stężenie produktów oksydacji cholesterolu." ("In normal atmosphere, with the decrease of meat slice weight the COPs content increses.")
- "Wraz ze spadkiem masy plastrów surowej wołowiny, poddanych pakowaniu w atmosferze bogatej w tlen (32%), zmniejsza się stężenie produktów oksydacji cholesterolu." ("In a highoxygen (32%) atmosphere, with the decrease of meat slice weight the COPs content decreses.")

### *B. Pre-ontological elaboration of statements, qualities and objects*

*a) STEP 2:* Statements formulated in *STEP 1* were automatically transferred to a database for further elaboration by the ontological staff of the project. Within this process all statements are annotated with unique identifiers. The two above statements from our example received numbers: t_652_3 and t_652_4 respectively. We shall further refer to them using these identifiers.

[6] Specialists worked on Polish translation – we present both Polish and English versions.

*b) STEP 3:* Each statement, whether proved or disproved, has an attributed type of thesis from Science.

For the selected types of scientific laws it is also necessary to determine the monotonic type. There are two main monotonic types: positive and negative. They express the direct or inverse proportion between qualities for the monotonic laws, increase or decrease of quality values for ordering laws. In special cases of ordering laws, additional monotonic types were introduced which are "equal", "different" or "no monotonic type". Exceptions, related to additional monotonic types, refer to these in which values of qualities are equal/stay unchanged or values are said to differ but the exact value differences are unknown. The latter exceptions refer to the ordering laws with differentiation factor, where the value of qualities is unknown.

At this point it is also determined for a particular thesis whether it is proved or disproved, by marking it as "accepted" or "rejected".

*Example 3.2:* Let us describe this on the example of thesis t_652_3 (it is a continuation of example 3.1). The statement is proved, so we mark it as "accepted". It describes dependency between weight and concentration of cholesterol oxidation products. Both qualities are quantitative qualities and quality dependence is known; therefore the thesis should be assigned as an instance of Monotonic Law. Monotonic laws require determining monotonic type. At this case decreasing the weight causes increasing the concentration of cholesterol oxidation products, that indicates an inverse proportion between qualities: "negative" monotonic type.

*c) STEP 4:* The next step of the development refers to relating the qualities to particular scientific laws. The qualities were listed by the domain specialists in the earlier phase (see *STEP 1*), but the list was linked to a particular article not to a thesis (see *STEP 2*). So if there were a few theses extracted from one article (which was almost always the case) there was a need of selecting the qualities for each thesis. A particular thesis needs to have not only proper qualities attributed but also (if possible) their role determined. The latter depends on the type of thesis. To attribute qualities to a thesis, the proper qualities are selected from the list. The proposed names of qualities must precisely describe the qualities mentioned in the thesis. In other case they should be corrected together with their English translation.

For all scientific laws qualities can be attributed; the only exception might be the laws of inclusion, which may refer to two or any quality. If a statement expresses any kind of influence between two or more qualities, then we should characterize the roles of these qualities in the statement. It is so in the case of all functional, monotonic and quasi-functional laws. In such statements dependent and independent qualities are distinguished. For methodological, ordering and object-property laws single qualities are attached to the statement. The special cases are laws of inclusion and correlation laws. As it was said, the laws of inclusion might refer to two or any quality. In the first case two qualities are assigned to a thesis in the following way: the first quality is a specialization of the second one. For correlation laws the direction of influence is unknown, for this reason each quality is assigned a number that corresponds to the order of its occurrence in the thesis.

*Example 3.3:* In our example (it is a continuation of example 3.2) two qualities: stężenie produktów oksydacji cholesterol (concentration of cholesterol oxidation products) and masa (weight) were extracted from the abstract (to which refers thesis t_652_3). Both of them refer to the statement t_652_3. Since the statement belongs to Monotonic Law, qualities dependency must be characterized. Hence, "weight" should be attributed to the statement as independent quality and "concentration of cholesterol oxidation products" as a dependent quality.

*d) STEP 5:* In the next step the elaborated data are automatically translated into an OWL ontology. The result is the Science ontology with classified instances of scientific laws. Each concrete scientific law was assigned qualities in appropriate roles, monotonicity (if applicable), the identifier of the article it comes from, the original sentence from the abstract of the article which expresses the law, its Polish translation and expert's elaboration.

*e) STEP 6:* The last part of the work is related to the description of objects (the experimental materials) the theses are about. The objects for the theses were described by the domain experts. Example 3.4 shows a description of the object for the aforementioned theses t_652_3 and t_652_4.

*Example 3.4 (Objects for theses t_652_3 and t_652_4):*

- mięso wołowe (beef)
- plastry surowej wołowiny poddane działaniu światła fluorescencyjnego (raw beef slices exposed to fluorescent light)
- pakowane w: atmosferze powietrza / bogatej w tlen atmosferze (packed in: normal atmosphere / oxygen-rich atmosphere)
- tlenek cholesterolu 7k (cholesterol oxide 7k)
- tlenek cholesterolu 7-$\beta$-OH (cholesterol oxide 7-$\beta$-OH)
- tlenek cholesterolu 7-$\alpha$-OH (cholesterol oxide 7-$\alpha$-OH)
- tlenek cholesterolu $\beta$-epoksydowy (cholesterol oxide $\beta$-epoxy)

At this level, the concepts of objects are developed according to the top categories of DOLCE (endurants, perdurants, qualities) and established (for this domain) types of relations. At the first step of the formalization, all classes of endurands, perdurants, qualities, their instances and relations between them are extracted from the object description prepared by the domain experts. They are written out with the assigned serial numbers and a shortcut describing the type of category (endurant "e", quality "c", relation "r"). Each object has its English translation in brackets (see example 3.5 below).

*Example 3.5 (Pre-ontological description of objects):*

1.e. mięso (beef)
2.e. bydło (cattle)
3.c. surowe (raw)
4.c. pokrojone w plastry (sliced)
5.e. światło fluorescencyjne (fluorescent light)
6.r. pakowane w (packed in)
7.e. normalna atmosfera (normal atmosphere)
8.e. bogata w tlen atmosfera (oxygen-rich atmosphere)
9.e. tlenek cholesterolu 7k (cholesterol oxide 7k)
10.e. tlenek cholesterolu 7-$\beta$-OH (cholesterol oxide 7-$\beta$-OH)
11.e. tlenek cholesterolu 7-$\alpha$-OH (cholesterol oxide 7-$\alpha$-OH)
12.e. tlenek cholesterolu $\beta$-epoksydowy (cholesterol oxide $\beta$-epoxy)
$\langle 1, o, 2 \rangle$
$\langle 1, q, 3 \rangle$
$\langle 1, q, 4 \rangle$
$\langle 1, tr, 5 \rangle$
$\langle 1, 6, 7 \rangle$
$\langle 1, 6, 8 \rangle$

Next the relations between objects (with use of given shortcuts) are written out in angle brackets. Firstly, the relations might occur between different objects. In that case we use triples: $\langle o_1, r, o_2 \rangle$, where "$o_1$", "$o_2$" describe objects and "$r$" the relation between them (see example 3.5 above). In some cases we also use quadruples. It is the case for relations such as "is aged for", "is stored in temperature", and any other relation which refers to a quality and its value with a proper unit. The pattern for quadruple is the following: $\langle o, r, v, u \rangle$, where "$o$" describes object, "$r$" – relation, "$v$" and "$u$" respectively value and its unit.

For some frequently repeated relations special symbols were established. In our case those symbols are "$o$" and "$tr$", which describe respectively the relation of "obtained from" (otrzymany z) and "is treated with" (potraktowany).

*f) STEP 7:* From the pre-ontological representation we automatically obtain the description of objects in OWL by translating tuples into axioms of the Science ontology.

## IV. Knowledge Management

After the process described in section III is finished, the Science ontology is populated by scientific laws with their formal description. Then by performing SPARQL queries we may obtain many interesting results.

*a) query 1:* For instance by SPARQL queries we are able to find accepted (IRI so:c0000001) monotonic laws which refer to the same qualities in the same roles but having different monotonicity (see query 1).

```
query 1

PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX so:<http://onto.beef.org.pl/science/>

SELECT ?class1 ?class2 ?thesis1 ?thesis2 ?m1 ?m2
WHERE {
?c1 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis1];
rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis2].

?c2 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis1];
rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis2].

?thesis1 rdf:type so:c0000015; so:hasMonotonicType ?m1;
rdf:type  so:c0000001.
?thesis2 rdf:type so:c0000015; so:hasMonotonicType ?m2;
rdf:type so:c0000001.

?c1 rdfs:label ?class1. ?c2 rdfs:label ?class2.
FILTER(lang(?class1) = "en"). FILTER(lang(?class2) = "en").
FILTER(?thesis1 != ?thesis2). FILTER(?m1 != ?m2).
}
```

The query gives 16 results in our KB (so 8 pairs of laws); four of them can be seen in table I. For instance the first row of the table informs that the class with english label "weight"@en is an independent parameter in theses with IRI so:t_652_3 and so:t_652_4 and the class with the English label "concentration of cholesterol oxidation products"@en is a dependent parameter in theses with IRI so:t_652_3 and so:t_652_4. But the thesis so:t_652_3 states that weight has a negative impact on concentration of cholesterol oxidation products, while so:t_652_4 states that the impact is positive. This result makes us curious about the reason for the change of monotonicity of the influence from "negative" to "positive". Having access to the bearers of the mentioned qualities we can learn that weight and cholesterol oxidation products are qualities of meat slices and that in the experiment they were divided into two groups which were packed differently: one in normal atmosphere and the other in highoxygen (32%) atmosphere. So the interesting information is: the way in which meat slices are packed has impact on the direction of influence of weight on the concentration of cholesterol oxidation products.

*b) query 2:* By querying the ontology we are also able to find new laws, i.e. not explicitly expressed in the scientific articles. For instance, the theses which are linked with classes of qualities by so:isIndependentParameterIn and so:isDependentParameterIn roles can be examined in the following way (see query 2): find two functional laws which share the same class of qualities, lets call it a "transitive_class", such that it is a dependent parameter in the first law and independent in the second; if they are found, the other classes appearing in the two laws—i.e. an independent parameter of the first, lets call it "class1", and a dependent parameter of the second, lets call it "class2" — create a new law.

```
query 2

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
```

TABLE I
PARTIAL RESULTS OF QUERY 1

| ?class1 | ?class2 | ?thesis1 | ?thesis2 | ?monoton1 | ?monoton2 |
|---|---|---|---|---|---|
| "weight"@en | "concentration of cholesterol oxidation products"@en | so:t_652_3 | so:t_652_4 | negative | positive |
| "level of cysteine"@en | "flavor"@en | so:t_2303_51 | so:t_2303_38 | positive | negative |
| "pressure"@en | "cooking loss"@en | so:t_1913_8 | so:t_1913_9 | positive | negative |
| "temperature"@en | "cooking loss"@en | so:t_557_5 | so:t_557_6 | positive | negative |

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX so: <http://onto.beef.org.pl/science/>
SELECT ?class1 ?tc_label ?class2
?thesis1 ?m1 ?thesis2 ?m2
WHERE {
?c1 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis1].
?tc rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis1];
rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis2].
?c2 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis2].
?thesis1 rdf:type so:c0000001.
?thesis2 rdf:type so:c0000001.
?c1 rdfs:label ?class1.
?tc rdfs:label ?tc_label.
?c2 rdfs:label ?class2.
FILTER (lang(?class1) = "en").
FILTER ( lang(?tc_label) = "en").
FILTER (lang(?class2) = "en").
OPTIONAL {?thesis1 so:hasMonotonicType ?m1.}
OPTIONAL {?thesis2 so:hasMonotonicType ?m2.}
}
```

Table II presents four out of 704 results the query brings about in our KB. In the first row we find a new scientific law stating that tenderness (column ?class1) influences color value (column ?class2). The law is obtained from the two laws: so:t_837_8 and so:t_278_4. The first states that tenderness has impact on pH ultimate value and the second that color value depends on pH ultimate value.

*c) query 3:* We can also ask about qualities which have an influence on a selected quality, for instance a quality with the label "tenderness"@en (see query 3 below).

```
query 3

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX so: <http://onto.beef.org.pl/science/>
SELECT ?class ?monoton ?article ?thesis
WHERE {
?c rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isIndependentParameterIn;
owl:hasValue ?thesis].

?c2 rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty so:isDependentParameterIn;
owl:hasValue ?thesis];
rdfs:label "tenderness"@en.

?thesis rdf:type so:c0000001;
so:comesFromArticleId ?article.

?c rdfs:label ?class.
FILTER (lang(?class) = "en").
OPTIONAL {?thesis so:hasMonotonicType ?monoton.}
}
```

In table III we see partial results of the query. We get the information that aging time, total collagen content, country origin and the amount of connective tissue influence the tenderness of meat. It is also worth noting that in table III we find three different articles which support the existence of correlation between the amount of connective tissue and tenderness.

## V. Related works

At least a few attempts have been made to address the problem which we have presented in the introduction. Now, we are going to present some of these works which are, in our opinion, interesting and relevant to our work. Firstly, we shall present the SWAN project which “aims to develop a practical, common, semantically structured framework for biomedical discourse initially applied, but not limited, to significant problems in Alzheimer Disease (AD) research” [10, p. 739]. The ontology called SWAN was created within this project. The goals of the SWAN ontology are: 1) to be the schema of a knowledge base in AD research; 2) to link the information from this knowledge base with information from other sources. These goals are intended to be achieved by using the following groups of concepts of the SWAN ontology: people and organizations, discourse elements, bibliographical information, concepts from the domain of interest, and others. The SWAN ontology allows for a representation in which, for example, there is a hypothesis from a scientific publication connected with some journal article by the relation “citesAsEvidence”, with claims by “contains”, and with a person by “authoredBy”. In turn, one of these claims can be connected with another claim by the relation “inconsistentWith”. It is possible, of course, to create these “epistemic” connections among concepts automatically, but in the cited paper the authors seem that they want them to be represented manually. In turn our KB allows for detecting inconsistency between statements automatically.

The second example which we would like to discuss here is the Nanopublications project [11]. The authors of this project argue that the lack of context (i.e. experimental data, citations, argumentation) of a statement (extracted from a scientific paper) which is intended to be represented is a serious problem, because “the statement can only be validated scientifically if you take into consideration its context” [11, p. 51]. And because it is common practice in the Semantic Web to represent a statement without its context (which, traditionally, is implicitly included in a source document), it is important to enable people to add a context to the represented statement.

TABLE II
PARTIAL RESULTS OF QUERY 2

| ?class1 | ?tc_label | ?class2 | ?thesis1 | ?m1 | ?thesis2 | ?m2 |
|---|---|---|---|---|---|---|
| "tenderness"@en | "pH ultimate value"@en | "color value"@en | so:t_837_8 | – | so:t_278_4 | positive |
| "aging time"@en | "pH value"@en | "emulsion capacity"@en | so:t_837_8 | – | so:t_1337_2 | positive |
| "rate of spoilage"@en | "fat content"@en | "cooking loss"@en | so:t_910_2 | – | so:t_936_3 | – |
| "flaxseed flour content"@en | "protein content"@en | "mean growth rate"@en | so:t_398_6 | negative | so:t_2184_10 | positive |

TABLE III
PARTIAL RESULTS OF QUERY 3

| ?class | ?monoton | ?article | ?thesis |
|---|---|---|---|
| "aging time"@en | so:positive | | so:t_1679_2 |
| "total collagen content"@en | so:negative | | so:t_569_21 |
| "country of origin"@en | – | | so:t_72_4 |
| "amount of connective tissue"@en | – | "164" | so:t_164_3 |
| "amount of connective tissue"@en | – | "1348" | so:t_1348_5 |
| "amount of connective tissue"@en | so:negative | "1716" | so:t_1716_1 |

The authors of the project created the nanopublications – a way to connect a statement with its context. A nanopublication is a "set of annotations that refer to the same statement and contains a minimum set of (community) agreed-upon annotations" [11, p. 52]. An annotation is a triple in which the subject is a statement. A statement, in turn, is a triple which can be uniquely identifiable. In such a triple there are three concepts in three different roles (positions): subject, predicate, and object. According to the authors, this model can be realized through the use of Named Graphs which enable adding a URI to a given RDF graph. A statement is a Named Graph and "all annotations belonging to a nanopublication should be part of the same Named Graph" [11, p. 53]. A nanopublication should represent the information from a source publication. In the SWAN ontology the statements were treated as independent components, here the authors of the project laid emphasis on the fact that a publication is a whole, but may still be connected to other publications. The usage of triples for representing statements is a well known practice. In one of the previous sections we have used triples (and, in fact, quadruples) for describing objects. But here we have an additional use of Named Graphs which allows for connecting statements with URIs in triples.

Another example is the Micropublication project within which "a layered metadata model of scientific argumentation and evidence" [12, p. 2] was created in order to "organize, verify, assess, combine and absorb this information [from the world's biomedical literature] in a comprehensive way" [12, p. 2]. The authors designed this model in such a way that it is able to represent both minimal and maximal forms of publications. In [12] the authors present nine use cases of this model. These use cases show that the model is very flexible and enables a wide range of possible usages: from a simple representation of a citable claim with supporting reference to a representation of a claim with a full chain of supporting evidence, citations etc. Also, the micropublication model allows for expressing statements in a natural language due to the fact that it is easier for scientists to annotate their work in that way.

Although the above mentioned approaches are interesting, they fail to take into account our specific need, i.e. the ability to represent information in a very specific manner. It is due to the fact that we would like to conduct automatic reasoning on our knowledge base. The SWAN ontology is intended to represent "lonely" statements in a simple formal manner, more or less the same applies to nanopublications. Micropublications can be represented both in formal and natural languages. We want our statements to be not only composed from a few concepts, but they should also be internally structured. This is provided by the taxonomy of scientific laws. And in our representation it is also possible to connect a law with a statement in natural language (e.g. by data property) and with a sentence from the abstract.

## VI. CONCLUSION AND PERSPECTIVES

We have presented an ontological representation of the main results of scientific articles from agri-food domain. The knowledge base consists of a TBox in which we define the structure of different kinds of scientific laws present in the domain and an ABox containing over 8k of particular laws. We have described in detail a way in which we proceed from abstracts of journal papers to assertions in ontology.

The work on the knowledge base is still in progress. It can be used as a core element of scientific information systems for the domain. We have not developed any such system in full by now but we have presented in the paper several SPARQL queries that show the kind of information that can be obtained from the knowledge base. The development of applications based on the knowledge base is one of the directions of future works.

The information extracted from the abstracts contains the detailed description of objects to which the laws pertain. The objects are connected to these laws in ontology. However, we have not developed algorithms that make extensive use of them. Tackling this task is another direction of future work.

Yet another challenge is concerned with the unification of the present results with the earlier ontological works within ProOptiBeef project. They resulted in the Domain ontology [13], [14] containing concepts that were used as keywords in the scientific papers from the projects database. The set of concepts from Domain and the set of objects and features used in Science ABox overlap. We believe that it is worth to merge both ontologies.

## Acknowledgment

Research was realized within the Project no. WND-POIG.01.03.01-00-204/09 Optimizing of Beef Production in Poland According to "from Fork to Farm" Strategy co-financed by the European Regional Development Fund under the Innovative Economy Operational Programme 2007 – 2013.

## References

[1] T. Nakayama, N. Hirai, S. Yamazaki, and M. Naito, "Adoption of structured abstracts by general medical journals and format for a structured abstract," *Journal of the Medical Library Association: JMLA*, vol. 93, no. 2, pp. 237–242, April 2005, PMID: 15858627. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15858627

[2] P. Kulicki, R. Trypuz, R. Trójczak, J. Wierzbicki, and A. Woźniak, "Ontology-based representation of scientific laws on beef production and consumption," in *MTSR*, E. Garoufallou and J. Greenberg, Eds., 2013. doi: 10.1007/978-3-319-03437-9_42 pp. 430–439.

[3] ——, "Semantic representation of proved and disproved statements extracted from scientific papers. Meat science case study," *Information Processing In Agriculture*, pp. 66–72, 2014. doi: 10.1016/j.inpa.2014.06.002

[4] R. Trójczak, R. Trypuz, and P. Kulicki, "Ontologia praw naukowych w kontekście reprezentacji i udostępniania wyników badań naukowych," *Filozofia Nauki*, vol. [forthcoming], 2015.

[5] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari, "WonderWeb Deliverable D18. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology," 2003. [Online]. Available: http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf

[6] P. Garbacz, M. Lechniak, P. Kulicki, and R. Trypuz, "Do you still want to vote for your favorite politician? Ask Ontobella!" in *In proceeding of: Formal Ontologies Meet Industry, Proceedings of the 4th Workshop FOMI 2009*, 2009. doi: 10.3233/978-1-60750-047-6-102 pp. 102–113.

[7] R. Mizoguchi, K. Kozaki, and Y. Kitamura, "Ontological analyses of roles," in *Federated Conference on Computer Science and Information Systems - FedCSIS 2012, Wroclaw, Poland, 9-12 September 2012, Proceedings*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2012. ISBN 978-83-60810-51-4 pp. 489–496. [Online]. Available: https://fedcsis.org/proceedings/2012/pliks/73.pdf

[8] K. Goczyla, A. Waloszek, and W. Waloszek, "Towards Context-Semantic Knowledge Bases," in *FedCSIS*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2012. ISBN 978-83-60810-51-4 pp. 475–482. [Online]. Available: http://dblp.uni-trier.de/db/conf/fedcsis/fedcsis2012.html#GoczylaWW12

[9] A. Martinson, "Guide for the preparation of scientific papers for publication," UNESCO, Paris, Tech. Rep. PGI-83/WS/10, 1983.

[10] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, and T. Clark, "The SWAN biomedical discourse ontology," *Journal of Biomedical Informatics*, vol. 41, no. 5, pp. 739–751, 2008.

[11] P. Groth, A. Gibson, and J. Velterop, "The Anatomy of a Nanopublication," *Information Services and Use*, vol. 30, no. 1-2, pp. 51–56, 2010. doi: DOI 10.3233/ISU-2010-0613

[12] T. Clark, P. N. Ciccarese, and C. A. Goble, "Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications," *Journal of Biomedical Semantics*, vol. 5, no. 1, p. 28, 2014. doi: 10.1186/2041-1480-5-28. [Online]. Available: http://www.jbiomedsem.com/content/5/1/28

[13] P. Kulicki, R. Trypuz, and J. Wierzbicki, "Towards beef production and consumption ontology and its application," *Federated Conference on Computer Science and Information Systems*, pp. 483–488, 2012. [Online]. Available: http://trypuz.pl/wp-content/papercite-data/pdf/weodia2012.pdf

[14] R. Trójczak, R. Trypuz, P. Grądzki, J. Wierzbicki, and A. Woźniak, "Evaluation of beef production and consumption ontology and presentation of its actual and potential applications," in *FedCSIS*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., 2013. ISBN 9781467344715 pp. 275–278.

[15] M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., *Federated Conference on Computer Science and Information Systems - FedCSIS 2012, Wroclaw, Poland, 9-12 September 2012, Proceedings*, 2012.

# Joint Agent-oriented Workshops in Synergy

JOINT Agent-oriented Workshops in Synergy is a coalition of agent-oriented workshops that come together to build upon synergies of interests and aim at bringing together researchers from the agent community for lively discussions and exchange of ideas. For the first time JAWS was organized during the 2011 FedCSIS Conference. Workshops that constitute JAWS in 2015 are:

- ABC:MI'15 - 10th Workshop on Agent Based Computing: from Model to Implementation
- MAS&S'15 - 9th International Workshop on Multi-Agent Systems and Simulations
- SEN-MAS'15 - 4th International Workshop on Smart Energy Networks & Multi-Agent Systems

# 10th Workshop on Agent Based Computing: from Model to Implementation

THE FIELD of agent technology is rapidly maturing. One of key factors that influence this process is the gathered body of knowledge that allows in-depth reflection on the very nature of designing and implementing agent systems. As a result, there is now significant knowledge on how to design and implement them. There is also a deeper understanding of the most important issues to be addressed in the process. Therefore, on the top-most level a progress in development of methodologies for design of agent-based systems can be seen. Furthermore, these methodologies are usually supported by tools that allow not only top level conceptualization but guide the process towards implementation (e.g. by generating at least some code). Next, it can seen that new languages for agent based systems are created, e.g. AML or API Calculus. Separately, tools/platforms/environments that can be used for design and implementation of agent systems have been through a number of releases, eliminating problems and adding new, important features. Resulting products are becoming truly robust and flexible. Furthermore, open source products (e.g. JADE) are surrounded by user communities, which often generate powerful add-on components, further increasing value of existing solutions.

## Topics

The Workshop primarily focuses on all aspects of the process that leads from the model of the problem domain to the actual agent-based solution. These aspects will cover both principled approaches and established practices of software engineering aimed at producing high quality software. In this context, research into the application of agent-based solutions to key challenges faced by software engineering (e.g. reduction of costs and delivery times, coping with a larger diversity of problems) will be of primary importance. ABC:MI Workshop welcomes submissions of original papers concerning all aspects of software agent engineering.

Topics include but are not limited to:

- Methodologies for design of agent systems
- Multi-agent systems product lines
- Modeling agent systems
- Agent architectures
- Agent-based simulations
- Simulating and verifying agent systems
- Agent benchmarking and performance measurement
- Agent communication, coordination and cooperation
- Agent languages
- Agent learning and planning
- Agent mobility
- Agent modeling, calculi, and logic
- Agent security
- Agents and Service Oriented Computing
- Agents in the Semantic Web
- Applications and Experiences

## Event Chairs

**Badica, Costin,** University of Craiova, Romania

**Ganzha, Maria,** University of Gdańsk and Systems Research Institute Polish Academy of Sciences, Poland

**Paprzycki, Marcin,** Systems Research Institute Polish Academy of Sciences, Poland

**Rahimi, Shahram,** Southern Illinois University, United States

## Program Committee

**Agotnes, Thomas,** University of Bergen, Norway

**Ambroszkiewicz, Stanislaw,** Institute of Computer Science, Polish Academy of Sciences, Poland

**Balke, Tina,** University of Surrey, United Kingdom

**Barseghyan, Artak,** Yerevan, Armenia

**Botía, Juan,** Universidad de Murcia, Spain

**Braubach, Lars,** University of Hamburg, Germany

**Budimac, Zoran,** Faculty of Sciences, Univ. of Novi Sad, Serbia

**Byrski, Aleksander,** AGH University of Science and Technology, Poland

**Cabri, Giacomo,** University of Modena and Reggio Emilia, Italy

**Cervenka, Radovan,** Whitestein Technologies AG, Slovakia

**Cetnarowicz, Krzysztof,** AGH University of Science and Technology, Poland

**Fernández, Alberto,** Universidad Rey Juan Carlos, Spain

**Florea, Adina,** University POLITEHNICA of Bucharest, Romania

**Gams, Matjaz,** Jozef Stefan Institute, Slovenia

**Gomez Sanz,** Jorge, Universidad Complutense de Madrid, Spain

**Goncalves, Ricardo,** Uninova, Portugal

**Hinchey, Mike,** Lero-the Irish Software Engineering Research Centre, Ireland

**Ivanović, Mirjana,** University of Novi Sad, Serbia

**Jamroga, Wojtek,** University of Luxembourg, Luxembourg

**Jedrzejowicz, Piotr,** Gdynia Maritime University, Poland

**Jezic, Gordan,** University of Zagreb, Croatia

**Kaleta, Mariusz,** Warsaw University of Technology, Poland

**Khorasani, Elham,** University of Illinois at Springfield, United States

**Koukam, Abder,** IRTES-SeT Université de Technologie de Belfort Montbéliard, France

**Kruczkiewicz, Zofia,** Wrocław University of Technology, Poland

**Kusek, Mario,** University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia

**Leszczyna, Rafal,** Gdansk University of Technology, Poland

**Letia, Ioan Alfred,** Technical University of Cluj-Napoca, Romania

**Morge, Maxime,** Université Lille 1, France

**Negru, Viorel,** West University of Timisoara, Romania

**Neruda, Roman,** Institute of Computer Science, Academy of Sciences of the Czech Republic, Czech Republic

**Ngoc-Thanh,** Nguyen

**Niazi, Muaz,** COMSATS Institute of IT, Pakistan

**Novak, Peter,** Dept. of Computer Science and Engineering, Czech Technical University in Prague, Czech Republic

**Nowostawski, Mariusz,** University of Otago, Information Science Department, New Zealand

**Oliveira, Eugenio,** Faculty of Engineering, University of Porto, Portugal

**Omicini, Andrea,** Alma Mater Studiorum–Università di Bologna, Italy

**Oren, Nir,** University of Aberdeen, United Kingdom

**Ouedraogo, Moussa,** Public Research Centre Henri Tudor, Luxembourg

**Paik, Incheon,** University of Aizu, Japan

**Poggi, Agostino,** DII - University of Parma, Italy

**Pokahr, Alexander,** University of Hamburg, Germany

**Rimassa, Giovanni,** Whitestein Technologies AG, Switzerland

**Rykowski, Jarogniew,** Poznan University of Economics, Poland

**Sakellariou, Ilias,** Dept. of Applied Informatics, University of Macedonia, Greece

**Santoro, Corrado,** University of Catania, Italy

**Schaefer, Robert,** AGH University of Science and Technology, Poland

**Senatore, Sabrina,** University of Salerno, Italy

**Slavkovik, Marija,** University of Luxembourg, Luxembourg

**Stanek, Stanislaw,** General Tadeusz Kosciuszko Military Academy of Land Forces in Wrocław, Poland

**Tang, Yuqing,** Carnegie Mellon University, United States

**Thimm, Matthias,** University of Koblenz-Landau, Germany

**Trcek, Denis,** University of Ljubljana, Slovenia

**Troquard, Nicolas,** ISTC-CNR, Italy

**Tucci, Salvatore,** University of Rome at Torvergata, Italy

**Venticinque, Salvatore,** Second University of Naples, Italy

**Vouros, George,** University of Piraeus, Greece

**Wahjudi, Paulus,** Marshall Unviersity, United States

**Yazdani, Samaneh,** Islamic Azad University, Science and Research branch, Iran

DOI: 10.15439/2015F300
ACSIS, Vol. 5

# Model Checking of Multi Agent System Architectures Using BigMC

Ahmed Taki Eddine Dib
Faculty of New Information Technologies and Communication
LIRE Laboratory, University of Constantine II.
Nouvelle Ville Ali Mendjeli - BP : 67A,
Constantine – Algeria
Email: dibtaki@gmail.com

Zaidi Sahnoun
Faculty of New Information Technologies and Communication
LIRE Laboratory, University of Constantine II.
Nouvelle Ville Ali Mendjeli - BP : 67A,
Constantine – Algeria
Email: sahnounz@yahoo.fr

*Abstract*— **Formal methods offer a great potential for early integration of verification in the design process. These are based on theories and mathematical notations that allow the formal specification of a program and check its implementation. They offer a global vision and a high-level structure and system organization. In addition, the software architecture plays a key role as a pivot point between the requirements of a system and its implementation. In this paper, we present a formal approach based on Bigraphical Reactive Systems for specifying and verifying the main features of the Multi Agent Systems (MAS) architectures based on the Belief-Desire-Intention (BDI) agent model. The proposed approach supports both the static and dynamic aspects of BDI-MAS architectures at different levels of abstraction. Further, we use automatic proof tool BigMc to analyze the specifications and verify system properties.**

**Keywords**: Multi-Agent Systems, software architecture description language, Bigraphical Reactive System, formal specification, reconfiguration, formal verification, Bigraphical Model Checker.

## I.Introduction

The emergence of large-scale IT networks has given rise to numerous distributed applications. These applications require a strong interaction between different entities distributed on the network that may share the same resources and the same goals. Several distributed development models for these applications have been proposed in the literature. However, the importance of the issue, is to be convinced of the legitimacy and trust granted to IT applications. These concerns have led to methods of development and verification. Lately software systems tend to be more distributed, open and concurrent. This evolution of computing has changed the way of thinking but also the design of such systems. Multi-Agent Systems (MAS) are particularly suitable for developing these kinds of systems. However, the diversity and complexity of the basic concepts that characterize multi-agent systems involve a difficulty in understanding and designing of such systems.

Formal methods offer a great potential for early integration of verification in the design process, these are based on theories and mathematical notations allowing both to formally specify the program, to check and prove that its implantation's compliance with all the properties described in the specification. This is called proper implementation with respect to the specification and formally verified program. Formal methods are recognized by standard references, so that the seventh and final confidence level of the Common Criteria [1] is granted to applications built with them. The awareness of the importance of checking more carefully the programs and the maturity of tools dedicated to this task has generated a considerable growth of programs formal verification in the last decade. Offering a global vision and a high-level structure and organization of a system, the software architecture plays a key role as a pivot point between the requirements of a system and its implementation.

The diversity of design concerns in general and particularly in MAS, request support on formal techniques, which offer enough flexibility and expressiveness to rigorously specify MAS architecture at both the static and dynamic level.

In our previous work [2] we proposed a new approach for modeling and analyzing MAS architectures called BDI MAS architecture in which Bigraphical Reactive Systems (BRS) [3] are adopted as a semantic framework to formalize MAS architectures that are based on the Belief-Desire-Intention (BDI) agent model [4] which is the most commonly used approach for representing agent internal state and it also has been used to build a number of significant real-world applications (i.e. web applications,...). Therefore, Milner's BRS are very suitable to formalize MAS fundamental architectural aspects and their reconfiguration.

Thus, in this work we argue that in addition to their graphical aspect and rigorous basis, BRS are capable of representing both locality and connectivity constituting main concepts of MAS architecture then we propose our a bigraph-based model in order to reason about their properties for that we use the automatic proof tool BigMC to analyze the specifications and verify system properties during configuration. The rest of the paper is organized as follows. In section 2, we introduce Bigraphical Reaction Systems (BRS) and the automatic proof tool BigMc . Section 3 and 4 present the related works and then our bigraphical specification of BDI-MAS architecture. The given

978-83-60810-66-8/$25.00©2015, IEEE

formalization approach is verified and validated by the BigMC tool through examples in section IV. Finally, some concluding remarks and ongoing work finish the paper.

## II. Bigraphs and Bigraphical Model Checker

### A. *Bigraphs*

Bigraphical Reactive Systems were initially introduced by Milner [3] to provide a completely graphical intuitive formal model capable of representing at the same time connectivity and locality of distributed entities which is very close to MAS concepts. The proposal of BRS provides a model for information systems with mobile placing and mobile linking, in which real-world pervasive and distributed systems can be described and analyzed. Further, it provides the unification of existing process calculi for concurrency and mobility (such as π-calculus, Petri nets, λ calculus, and so on) in a simpler way [5].

Structural Aspects: A bigraph is the combination of two independent structures place and link graphs. The place graph represents system entities geographical distribution. The link graph is a hypergraph representing interconnections between these entities. Within a BRS, system entities are represented by nodes and interactions between them are represented by edges (see Fig. 1.). A node can be dotted with ports representing connexion points to edges or inner/outer names.

Each node has a control, which is an identifier belonging to a set that is called a signature (usually denoted as S). Each control indicates how many ports the node has, whose controls are atomic (node empty), and which of the non-atomic controls are active (node permitting reaction inside) or passive. The inner names and outers names of a bigraph indicate connecters to which other bigraphs or roots (i.e. regions) can be connected. Such interconnection is possible only if the outer name of a bigraph or root is equal to the inner name of another bigraph. Sites represent holes into which a root or node can be nested. They are considered as an abstraction indicating the presence of other elements.

Definition [3]: a bigraph is formally defined by $G = (V, E, ctrl, G^P, G^L) : I \rightarrow J$, $I = <m, x>$, $J = <n, y>$, where:

- $V$ and $E$ represent finite sets of nodes and edges respectively.
- $ctrl : V \rightarrow K$ a control map that assigns a control to each node. The signature $K$ is a set of controls.
- $G^P$ and $G^L$ are Place and Link graphs respectively.
- $I$ and $J$ represent inner and outer names (interfaces) respectively of the bigraph $G$.

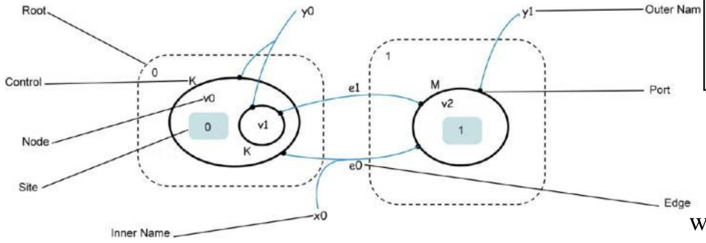


Fig. 1 The anatomy of bigraphs

Bigraph can also be expressed by term language. In [5] Milner axiomatises the structure of bigraphs, to prove that the theory is complete, the algebra of bigraphs structure is surprisingly simple, the primary operations and elements used in this paper are summarized in Table 1.

TABLE 1. TERMS LANGUAGE FOR BIGRAPHS.

| Term | Signification |
|---|---|
| $U \mid\mid V$ | Juxtaposition of roots |
| $U \mid V$ | Juxtaposition of nodes |
| $U \circ V$ | Composition |
| $U . V$ | Nesting( U contains V) |
| $/x . U$ | U with outer name x replaced by an edge |
| $x /y$ | Connection inner names y to outer name x |

Dynamical aspects: Bigraphs structural dynamics is expressed through a BRS (Bigraphical Reactive System) consisting of a category of bigraphs and a set of reaction rules; each one defines a redex bigraph to be transformed to a reactum bigraph. Formally, a reaction rule takes the form$(R,R',n)$ where $R : m \rightarrow J$ is a redex, $R' : m' \rightarrow J$ is a reactum and $n : m' \rightarrow m$ is a map of ordinals [3]. The category of all bigraphs and their reaction rules constitute a BRS.

### B. *Bigraphical Model Checker (BigMC)*

The use of formal methods allows rigorous verification of computer systems. There exist a number of formal verification techniques, model checking [6, 7] is one of many and BigMC (Bigraphical model checker) [8] is one of the few model checking tool devoted to the verification of models encodes as a Bigraphical Reactive System. BigMC ensures that the system's behavior meets the expected properties. This verification is fully automated and consists in exploring all the possible cases. The result of this analysis is to confirm that each property is verified, or not. In the latter case, and this is one of the main interests of this tool, the model checker returns a counter-example. Fig. 2 shows the full BigMC bigraph term:

$$\begin{aligned}
&M ::= E; M \mid E\\
&E ::= \%passive\ k :\ arity\\
&E ::= \%active\ k :\ arity\\
&E ::= \%rule\ n\ T \rightarrow T\\
&E ::= \%property\ n\ P\\
&E ::= T \rightarrow T \mid T\\
&T ::= K.T \mid T \mid T \mid T \mid\mid T \mid \$n \mid K \mid nil\\
&K ::= k[names] \mid k\\
&names ::= n, names \mid n\\
&n ::= [a - zA - Z][a - zA - Z0 - 9] * \mid -\\
&P ::= matches(T) \mid terminal() \mid !P
\end{aligned}$$

Fig. 2 BigMC terms language

Using the grammar in Fig. 2, we can specify a model M which may be a composed of another model or an expression E or both. In turn an expression E can be composed of nodes (being active or passive and assigning an arity to each one of them), reaction rules (whose form is as follows T -> T) and properties denoted by P (which are expressed as a logical formula). In this work, we will use the BigMC grammar to specify our proposed BDI-MAS architecture model in order to check and validate some properties.

## III.MULTI AGENT SYSTEMS BIGRAPH BASED SPECIFICATION

To better understand the multi agent system development, we have reviewed the literature related to Architecture description languages (ADLs) and those of MAS. Therefore, in our previous work [2] we have captured the fundamental concepts to better ensure the specification, evolution and the verification of MAS architecture. This modeling has studied both the structural and dynamic dimension of multi-agent systems architectures. These two dimensions (structural and dynamic) will be developed in this section to show how the proposed framework based on the BRS as a formal notation, express the multi-agent architectures.

At a high level of abstraction, multiagent system is considered as a set of computing entities (a set of agents) that are distributed across multiple sites, and are often referred to as nodes. In Table 2, we summarize fundamental elements intervening in a BDI-MAS architecture.

TABLE 2. CORRESPONDENCE BETWEEN MAS AND BRS CONCEPTS.

| MAS architectural element | Bigraph element |
|---|---|
| Agents, Beliefs module, Desires module, Intention module, plans. | Node |
| Physical or logical location the agents | Root |
| Various type of links between the different elements | Edge/Hyper Edge |
| Abstract elements | Site |

### A. Structural description of the BDI-MAS model

Our BDI-MAS architecture model structure follows core principles, which we organize in two levels of abstraction: (i) internal (or agent) level; (ii) social (or MAS) level. The former describes the internal structure and state of the agent (i.e. the basic construct elements of the MAS) and the second describes the assembly and interaction among agents that compose the MAS architecture. A multi-agent system does not reduce to a centralized computer system; it consists of a set of interconnected agents. Where each agent can initiate communication, generate messages, and respond to other agent's messages, in order for agents participating in these interactions to achieve overall system goals [9].

#### *a) Agent level:*

The Fig. 3 shows our BDI agent and its internal structure. Each agent (denoted by AG) is composed of three principal nodes, which in turn contains other nodes that structure them. In what follows, we will take a closer look on the nodes that compose the agent AG1, for more details see [2].

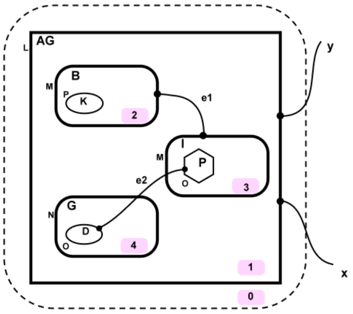


Fig. 3 Bigraphical model of BDI Agent.

The signature associated to a BDI-MAS bigraph is as follows:
K = { L: (2, active), M: (1, active), N :( 0, active), O :( 1, atomic), P :( 0, atomic)}, L, M, N, O and P represent controls associated to different nodes. The different nodes types used in the model and their associated controls are summarized in Table 3.

TABLE 3. NODES TYPES OF BDI-MAS ARCHITECTURE.

| Node | Control | Attribute | Arity | Meaning |
|---|---|---|---|---|
| AG | L | Active | 2 | Agent |
| B | M | Active | 1 | Beliefs Module |
| G | N | Active | 0 | Goal Module |
| I | M | Active | 1 | Intention Module |
| P | O | Atomic | 1 | Plan |
| D | O | Atomic | 1 | Desire |
| K | P | Atomic | 0 | Knowledge |

#### *b) Social level:*

The model presented provides notations for describing the structure of MAS in terms of hierarchical configurations of interacting components. It provides an explicit and common basis for describing MAS architectural configurations (see Fig. 4).

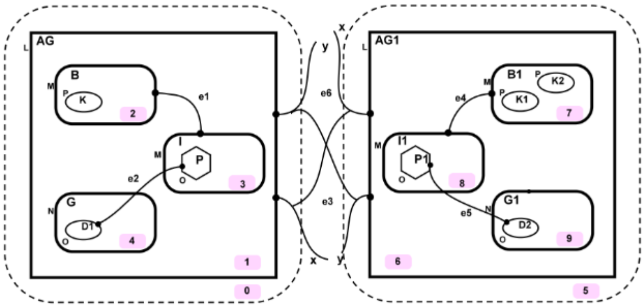


Fig. 4. Bigraphical model of BDI-MAS configuration.

### *B. Modeling BDI-MAS Architectural Reconfiguration*

As defined in our previous work [2] the BDI-MAS architecture dynamics is formalized using reaction rules expressing changes of form in terms of shape shifting while preserving architectural constraints. In this subsection, we give some reaction rules samples defined to model BDI-MAS internal and external behavior and reconfiguration. Table 4 depicts how we defined the behavioral model, based on reaction rules.

**TABLE 4.** MODELLING MULTI AGENT SYSTEM DYNAMICS.

| Multi Agent System | BRS |
|---|---|
| Configuration *MAS*. | Bigraph : $G_{MAS}= (V_{MAS}, E_{MAS}, crtl_{MAS}, G_{MAS}\,p\,, G_{MAS}\,L\,)$ |
| Reconfiguration from *MAS* to *MAS'* | Meta reaction rule: $RL = (MAS, MAS', m' \rightarrow m)$ |

**Example RL1:** Resolution of an internal goal reaction rule

$$AG_{x\,y}\ .(B_{e1}.(K \mid d2) \mid G.(D1 \mid d4) \mid I_{e1}\,.(P \mid d3) \mid d1) \mid d0$$
$$\rightarrow$$
$$AG_{x\,y}\ .(B_{e1}.(K \mid K1 \mid d2) \mid G.(D1_{e2} \mid d4) \mid I_{e1}\,.(P_{e2} \mid d3) \mid d1) \mid d0$$

**Example RL2:** Resolution of an external goal (collaboration) reaction rule

$$AG_{xy}.(B_{e1}.(K|K1|d2)|G.(D1|d4)|I.(d3)|d1)|AG1_{xy}.(B1_{e2}.(K2|d7)|G1.(D3|d9)|I1_{e2}.(d8)\ |d6)$$
$$\rightarrow$$
$$AG_{xy}\ .(B_{e1}.(K \mid K1 \mid K3 \mid d2) \mid G.(D1_{e6} \mid d4) \mid I.(P2_{e6} \mid d3) \mid d1) \mid AG1_{xy}.(B1_{e2}.(K2 \mid K3 \mid d7) \mid G1.(D3 \mid d9) \mid I1_{e2}.(d8) \mid d6)$$

## IV. FORMAL ANALYSIS OF PROPERTIES

Software verification becomes essential to the development of computer systems. Indeed, the use of formal methods allows to prove that a system satisfies a given specification. In fact, these methods appear one of the main solutions for the development of high quality and safe systems at a reasonable costs and time span. Further, the use of these methods in the development process allows the verification and validation of the specification and facilitates the passage to the implementation. These techniques are accompanied by powerful tools that can be used to automate various stages of verification. The use of rather conventional design methods (composition, aggregation, etc.) in the development cycle has paved the way for the smooth introduction of techniques such as model checking.

In our case, we use BigMC a Bigraphical Model Checker to check properties such as deadlock and some violations that the model should not allow to happen during its execution. First, we specify the structural aspect (i.e., nodes and their signature and outer and inner interfaces …) then the dynamic aspect (i.e., reaction rules ex internal resolution of goal …) using the BigMC syntax term language. Then we formulate the properties that we would like to verify on each example. Finally, we will analyze and validate the resulted output given by the BigMC tool.

### *A. Reachability checking*

In this section, we would like to verify the soundness of our model. For that purpose, we decide to start with the building blocks of our model's dynamic, which are no other than the reaction rules for the resolution of an internal, external goal resolution and the reconfiguration example of adding a new agent to the system specified in the section 4.

Fig. 5 describes, how our BDI agent is able to solve an internal goal, as it may be seen in the redex, the presence in the node G of a desire D1 to satisfy, one can also notice the presence into the node B of a knowledge node denoted by K, which is necessary for triggering the rule. In the reactum one can first see the appearance of a node P, which is the best plan among plans that can satisfy the desire D1 (choosing the best possible plan remains tied to heuristics that cannot appear in the architectural level). Secondly, the creation of a link e2 between the node D1 and P specify that the desire D1 could be satisfied by the execution of the plan P. Finally, the execution of P induces two possible cases either: (1) adding / removing knowledge at the node B (2) beliefs of the agent do not change [2].

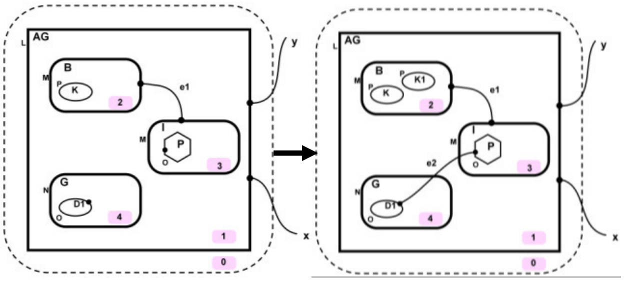
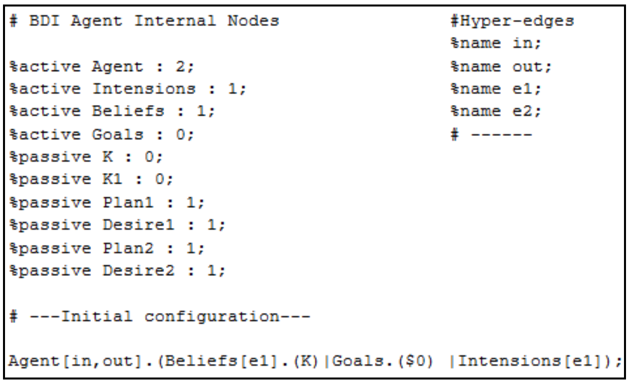


Fig. 5. Internal goal resolution reaction rule.

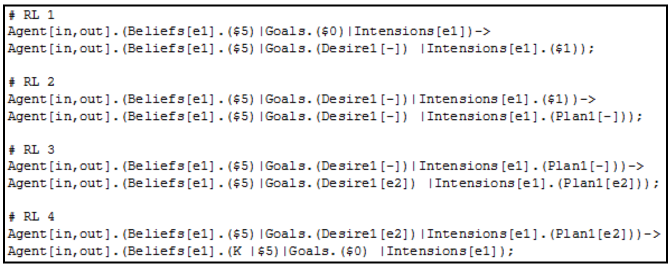
Fig. 6 below represents the structural bigraphical specification of the resolution of an internal goal in BigMC term language.



```
# BDI Agent Internal Nodes                        #Hyper-edges
                                                  %name in;
%active Agent : 2;                                %name out;
%active Intensions : 1;                           %name e1;
%active Beliefs : 1;                              %name e2;
%active Goals : 0;                                # ------
%passive K : 0;
%passive K1 : 0;
%passive Plan1 : 1;
%passive Desire1 : 1;
%passive Plan2 : 1;
%passive Desire2 : 1;

# ---Initial configuration---

Agent[in,out].(Beliefs[e1].(K)|Goals.($0) |Intensions[e1]);
```

Fig. 6 Internal goal resolution in BigMC term language

Fig. 7 shows the dynamics of our example, through the presentation of a sequence of meta-reaction rules written in BigMC, we can also see that there are two types of reactions of rules linking and placing reaction rules. The former is responsible for creating or deleting links between nodes while the second type is responsible for creating, moving or deleting nodes in our BDI-MAS Agent.



```
# RL 1
Agent[in,out].(Beliefs[e1].($5)|Goals.($0)|Intensions[e1])->
Agent[in,out].(Beliefs[e1].($5)|Goals.(Desire1[-]) |Intensions[e1].($1));

# RL 2
Agent[in,out].(Beliefs[e1].($5)|Goals.(Desire1[-])|Intensions[e1].($1))->
Agent[in,out].(Beliefs[e1].($5)|Goals.(Desire1[-]) |Intensions[e1].(Plan1[-]));

# RL 3
Agent[in,out].(Beliefs[e1].($5)|Goals.(Desire1[-])|Intensions[e1].(Plan1[-]))->
Agent[in,out].(Beliefs[e1].($5)|Goals.(Desire1[e2]) |Intensions[e1].(Plan1[e2]));

# RL 4
Agent[in,out].(Beliefs[e1].($5)|Goals.(Desire1[e2])|Intensions[e1].(Plan1[e2]))->
Agent[in,out].(Beliefs[e1].(K |$5)|Goals.($0) |Intensions[e1]);
```

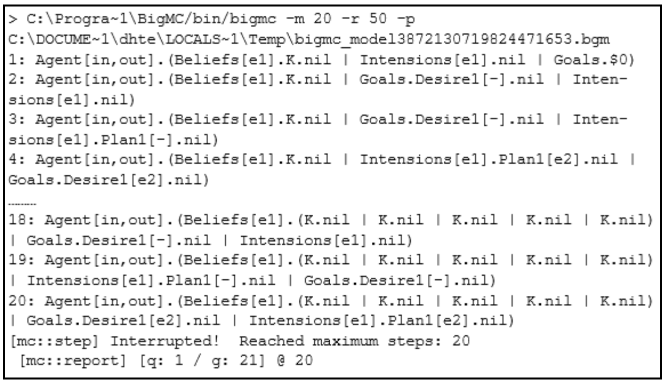
Fig. 7. Internal goal resolution dynamics

Now that we have specified our model structural and dynamical aspects in BigMC the penultimate stage before the checking is to specify or formulate the property to check on our example. For this purpose, BigMC provides a set of predefined predicates using the syntax showing in Fig. 2, in the following example, we will use two of them, the first property that we would like to verify named violation_free uses the predicate !match(T) which states that we must not find a match to the expression between brackets during our system execution. The second property is deadlock_free which uses the predicate !terminal() the predicate as transcribed here states that there will be a possible future state reachable by a step of reaction rule from the current one. For more elaborated properties the common boolean operator such as AND, OR and NOT are used.

%property secure !matches(Agent[in,out].(Beliefs[e1].(K | K1 | $2)|Goals.(Desire1[-] | $0) |Intensions[e1].(Plan1[-] | $1 )));

%property deadlock_free !terminal();

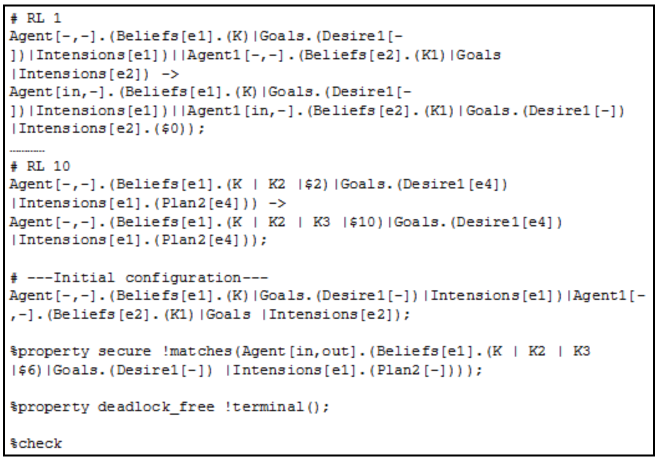
The result of the model checking is shown in the figure after 20 steps the model checker reached successfully the intended state and does not report any property violation (due to the lake of space intermediate rewriting steps are omitted) .



```
> C:\Progra~1\BigMC/bin/bigmc -m 20 -r 50 -p
C:\DOCUME~1\dhte\LOCALS~1\Temp\bigmc_model3872130719824471653.bgm
1: Agent[in,out].(Beliefs[e1].K.nil | Intensions[e1].nil | Goals.$0)
2: Agent[in,out].(Beliefs[e1].K.nil | Goals.Desire1[-].nil | Inten-
sions[e1].nil)
3: Agent[in,out].(Beliefs[e1].K.nil | Goals.Desire1[-].nil | Inten-
sions[e1].Plan1[-].nil)
4: Agent[in,out].(Beliefs[e1].K.nil | Intensions[e1].Plan1[e2].nil |
Goals.Desire1[e2].nil)
………
18: Agent[in,out].(Beliefs[e1].(K.nil | K.nil | K.nil | K.nil | K.nil)
| Goals.Desire1[-].nil | Intensions[e1].nil)
19: Agent[in,out].(Beliefs[e1].(K.nil | K.nil | K.nil | K.nil | K.nil)
| Intensions[e1].Plan1[-].nil | Goals.Desire1[-].nil)
20: Agent[in,out].(Beliefs[e1].(K.nil | K.nil | K.nil | K.nil | K.nil)
| Goals.Desire1[e2].nil | Intensions[e1].Plan1[e2].nil)
[mc::step] Interrupted!  Reached maximum steps: 20
 [mc::report] [q: 1 / g: 21] @ 20
```

Fig. 8. Internal goal resolution checking result.

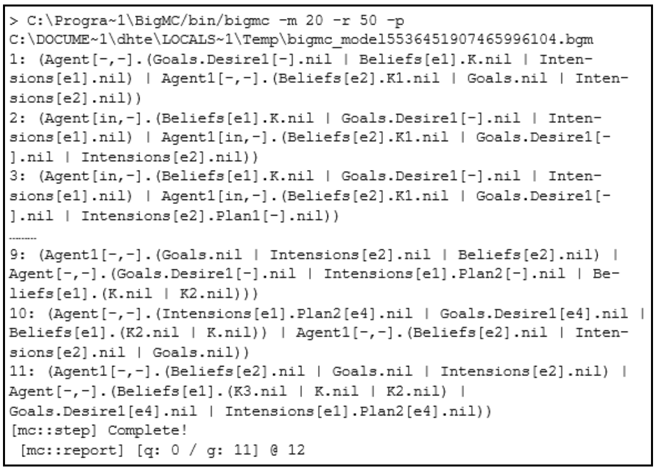
**Example 2** the external goal resolution implies at least two agents as shown in the Fig. 4 of the section 4. The Fig. 9 represents the example transcribed in BigMC term language.



```
# RL 1
Agent[-,-].(Beliefs[e1].(K)|Goals.(Desire1[-
])|Intensions[e1])||Agent1[-,-].(Beliefs[e2].(K1)|Goals
|Intensions[e2]) ->
Agent[in,-].(Beliefs[e1].(K)|Goals.(Desire1[-
])|Intensions[e1])||Agent1[in,-].(Beliefs[e2].(K1)|Goals.(Desire1[-])
|Intensions[e2].($0));
………
# RL 10
Agent[-,-].(Beliefs[e1].(K | K2 |$2)|Goals.(Desire1[e4])
|Intensions[e1].(Plan2[e4])) ->
Agent[-,-].(Beliefs[e1].(K | K2 | K3 |$10)|Goals.(Desire1[e4])
|Intensions[e1].(Plan2[e4]));

# ---Initial configuration---
Agent[-,-].(Beliefs[e1].(K)|Goals.(Desire1[-])|Intensions[e1])|Agent1[-
,-].(Beliefs[e2].(K1)|Goals |Intensions[e2]);

%property secure !matches(Agent[in,out].(Beliefs[e1].(K | K2 | K3
|$6)|Goals.(Desire1[-]) |Intensions[e1].(Plan2[-])));

%property deadlock_free !terminal();

%check
```

Fig. 9 External goal resolution written in BigMC.

The model checker rewriting steps is limited to 50, the result is without call the model is free of violation. As a result, the BigMC tool does not give a counter example see Fig. 10.



```
> C:\Progra~1\BigMC/bin/bigmc -m 20 -r 50 -p
C:\DOCUME~1\dhte\LOCALS~1\Temp\bigmc_model5536451907465996104.bgm
1: (Agent[-,-].(Goals.Desire1[-].nil | Beliefs[e1].K.nil | Inten-
sions[e1].nil) | Agent1[-,-].(Beliefs[e2].K1.nil | Goals.nil | Inten-
sions[e2].nil))
2: (Agent[in,-].(Beliefs[e1].K.nil | Goals.Desire1[-].nil | Inten-
sions[e1].nil) | Agent1[in,-].(Beliefs[e2].K1.nil | Goals.Desire1[-
].nil | Intensions[e2].nil))
3: (Agent[in,-].(Beliefs[e1].K.nil | Goals.Desire1[-].nil | Inten-
sions[e1].nil) | Agent1[in,-].(Beliefs[e2].K1.nil | Goals.Desire1[-
].nil | Intensions[e2].Plan1[-].nil))
………
9: (Agent1[-,-].(Goals.nil | Intensions[e2].nil | Beliefs[e2].nil) |
Agent[-,-].(Goals.Desire1[-].nil | Intensions[e1].Plan2[-].nil | Be-
liefs[e1].(K.nil | K2.nil)))
10: (Agent[-,-].(Intensions[e1].Plan2[e4].nil | Goals.Desire1[e4].nil |
Beliefs[e1].(K2.nil | K.nil)) | Agent1[-,-].(Beliefs[e2].nil | Inten-
sions[e2].nil | Goals.nil))
11: (Agent1[-,-].(Beliefs[e2].nil | Goals.nil | Intensions[e2].nil) |
Agent[-,-].(Beliefs[e1].(K3.nil | K.nil | K2.nil) |
Goals.Desire1[e4].nil | Intensions[e1].Plan2[e4].nil))
[mc::step] Complete!
 [mc::report] [q: 0 / g: 11] @ 12
```

Fig. 10. External goal resolution checking result.

## V.Related work

There is an important core of work regarding to the design and development at the architectural level as mentioned in [9] and [10], several works propose different languages, formal and semi-formal, Architecture Description Language (ADL). Such as Darwin, Rapide, Dynamic-Wright [11] and π-ADL[12] for representing and analyzing software architectures in order to predict architectural qualities before the implementation, and guiding the design and coding process. Nonetheless, these works are labeled by a lack of coverage of concepts related to the definition of a multi-agent system, for example, the representation of the agent is generally limited by a single object devoid of the necessary concepts to express its autonomy and cognitive aspects (such as beliefs, knowledge and competences). As cited in [13] there exist various analysis techniques among the existing ADLs for testing, model checking, and evaluating performance based on architectural models. Bordini in [15, 16] has presented an approach for verifying multi-agent programs. In this approach, the system is written with the logic-based agent-oriented programming language AgentSpeak and automatically translated into either Promela or Java. This is an important work; however, the verification of MAS focuses on the program rather on the architecture. In [17] Walton address the verification of communication between agents participating in multi-agent web service systems, the approach is based on the application of model checking techniques. This approach is too specific, it is used to verify lightweight protocol language and it cannot be applied to a wide range of multi agent systems and neither at the architectural level. In [18] and [19] are abstract formal models for developing formal specifications of multi-agent systems these approachs uses the Z notation as formal foundation. However, the Z language cannot model in an effective way the interaction, distribution and the concurrence in a MAS. Fisher in [20] describes the first steps towards a formal specification and verification of multi-agent systems using Concurrent METATEM and the temporal belief logics. This approach suffers from a low level of abstraction and does not take into account the reconfiguration of the system at the architectural level.

## VI.Conclusion

In this paper, we have described our proposed formal modeling approach of the BDI-MAS architecture. The system has been specified at both individual (agent) and social (MAS) levels. The BDI-MAS bigraph simplifies considerably the MAS architectures readability. A MAS architecture is seen as a hierarchical configuration of interacting nodes. The model emphasizes on both locality and connectivity that can be used to represent the location and interconnection of MAS architectures. On the other hand, reaction rules allow developers to correctly analyze the BDI-MAS architecture features, including modeling the behavior of the BDI agents and describing reconfigurations that could be added to the architecture. Further, the use of bigraphs as formal basis in the development process allows the verification and validation of the specification. Using the BigMC tool we have shown that our BDI-MAS architecture model through its Meta reaction rules are free of violations and deadlock. Our aim is to have a graphical intuitive solid formal foundation for modeling MAS architecture in order to handle the complexity of the systems in general, adopting a high level of abstraction that removes unnecessary details regarding all the expected properties and facilitates the passage to the implementation.

In the perspectives of this work, we plan to:

- Formally analyze and verify some non-functional properties such as security of the BDI-MAS architectures model.
- Provide a tool that generates executable implementation from our BDI-MAS architecture model,
- Develop a methodology around the model in order to guide the development of MAS.

## References

[1] Commoncriteriaportal.org, 'Common Criteria : New CC Portal', 2015. [Online]. Available: http://www.commoncriteriaportal.org/. [Accessed: 16- Feb- 2015].

[2] A. Dib and Z. Sahnoun, 'Formal Specification of Multi-Agent System Architecture', in International Conference on Advanced Aspects of Software Engineering, Constantine, 2014, pp. 65-72.

[3] R. Milner, 'Bigraphs and Their Algebra', Electronic Notes in Theoretical Computer Science, vol. 209, pp. 5-19, 2008.

[4] A. Rao and M. Georgeff, 'Modeling rational agents within a BDI-architecture', Australian Artificial Intelligence Institute, Victoria, Australia, 1991.

[5] R. MILNER, 'Axioms for bigraphical structure', Math. Struct. in Comp. Science, vol. 15, no. 06, p. 1005, 2005.

[6] S. Merz, 'Model Checking: A Tutorial Overview', in Modeling and Verification of Parallel Processes, 2001, pp. 3-38.

[7] S. Merz, 'Model Checking Techniqes for the Analysis of Reactive Systems', Synthese, vol. 133, no. 12, pp. 173-201, 2002

[8] G. Perrone, S. Debois and T. Hildebrandt, 'A model checker for Bigraphs', Proceedings of the 27th Annual ACM Symposium on Applied Computing - SAC '12, 2012.

[9] R. Allen, 'A Formal Approach to Software Architecture', Phd, Carnegie Mellon University, 1997.

[10] I. Malavolta, P. Lago, H. Muccini, P. Pelliccione and A. Tang, 'What Industry Needs from Architectural Languages: A Survey', IIEEE Trans. Software Eng., vol. 39, no. 6, pp. 869-891, 2013.

[11] N. Medvidovic and R. Taylor, 'A classification and comparison framework for software architecture description languages', IIEEE Trans. Software Eng., vol. 26, no. 1, pp. 70-93, 2000.

[12] R. Allen, R. Douence and D. Garlan, 'Specifying and analyzing dynamic software architectures', Fundamental Approaches to Software Engineering, pp. 21-37, 1998.

[13] F. Oquendo, pi-ADL', SIGSOFT Softw. Eng. Notes, vol. 29, no. 3, p. 1, 2004.

[14] P. Zhang, H. Muccini and B. Li, 'A classification and comparison of model checking software architecture techniques', Journal of Systems and Software, vol. 83, no. 5, pp. 723-744, 2010.

[15] R. Bordini, M. Fisher, W. Visser and M. Wooldridge, 'Verifying Multi-agent Programs by Model Checking', Autonomous Agents and Multi-Agent Systems, vol. 12, no. 2, pp. 239-256, 2006.

[16] R. Bordini, M. Fisher, C. Pardavila, W. Visser and M. Wooldridge, 'Model Checking Multi-Agent Programs with CASP', Computer Aided Verification, pp. 110-113, 2003.

[17] W. Wan, J. Bentahar and A. Ben Hamza, 'Modeling and Verifying Agent-Based Communities of Web Services', Trends in Applied Intelligent Systems, pp. 418-427, 2010.

[18] D'Inverno, M., Luck, M., Georgeff, M., Kinny, D. and Wooldridge, M. (2004). The dMARS Architecture: A Specification of the Distributed Multi-Agent Reasoning System. Autonomous Agents and Multi-Agent Systems, 9(1/2), pp.5-53.

[19] Luck, M. and d'Inverno, M. (2006). Formal Methods and Agent-Based Systems. NASA Monographs in Systems and Software Engineering, pp.65-96.

[20] M. Fisher and M. Wooldridge, 'On the Formal Specification and Verification of Multi-Agent Systems', International Journal of Cooperative Information Systems, vol. 06, no. 01, pp. 37-65, 1997.

DOI: 10.15439/2015F302
ACSIS, Vol. 5

# Multisource agent-based healthcare data gathering

Vincenza Carchiolo, Alessandro Longheu, Michele Malgeri and Giuseppe Mangioni

Dip. Ingegneria Elettrica, Elettronica e Informatica - Università degli Studi di Catania - Italy
{vincenza.carchiolo, alessandro.longheu, michele.malgeri, giuseppe.mangioni}@dieei.unict.it

***Abstract*—The number and type of digital sources storing healthcare data is increasing more and more, rising the problem of collecting actually dispersed information about a single patient. In this paper we propose an agent-based system to support integration of health-related data extracted from both structured (HIS) and semi-structured (websites and social networks) sources. Integrated data are exported in HL7 format to finally feed personal health record (PHR).**

## I. Introduction

DIGITAL healthcare data dramatically increased during last years [1]; this is mainly due on one hand to data stored into Health Information Systems (HIS) [2][3] as medical records and clinical exams, while on the other hand a significant contribution comes from specialized websites (e.g. online medical forums) and social networks as FaceBook and Twitter, where doctors and patients post their personal experiences and opinions about various health related topics e.g. illnesses, symptoms, treatments, side effects [4][5].

The number and type of such data sources poses the problem of collecting actually dispersed healthcare information for a single patient, therefore the extraction and integration of such data into a standard repository arise.

Data integration is a well-known and quite old issue [6][7], however it still requires a significant effort, especially when information are extremely sensitive for privacy issues, as it occurs for medical information [8].

In this paper we propose an agent-based systems whose goal is the integration of health related data coming from different sources; as cited before, we consider both HIS (i.e. SQL based data sources) as well as websites (HTML-based data sources) and social networks (in particular, Twitter).

These are used to populate official medical documentation known as Electronic Medical Records (EMR), Electronic Health Records (EHR) and Personal Health Record (PHR). In more detail, EMR/EHR [9][10][11] is the digital collection of a person's health related documents used within HIS to provide an effective, reliable and costs saving health management, contains the standard medical and clinical data and is managed only by health care providers, whereas in the PHR [12] the person can directly manage his personal medical-related information and therefore it can also contain data coming from website and social networks cited previously (i.e. unstructured or semi-structured). In the rest of this paper, we will use the term PHR only, implicitly including also EMR/EHR.

Since our goal is the gathering of *personal* health-related data, we suppose that all information can be accessed on a per-user basis according to authentication mechanisms whenever present, i.e. agents should be granted access to personal profiles to extract data.

Moreover, for an effective integration a common reference terminology is needed; to this purpose, we exploit the SNOMED-CT [13], currently the most comprehensive medical terminology worldwide adopted, to discover medical terms and properly manage, match and integrate synonyms, hyponyms and hypernyms.

Together with standard database terminology, we also include a list of additional informal terms to be searched if nothing is found within SNOMED-CT, indeed it is possible that within unofficial medical data sources (as in the case of social networks and websites) people use words like ”headache” that are not explicitly stored into SNOMED-CT, where ”headache” is indeed referred as ”migraine”; the additional list allows to cope with such real situations.

Finally, a common output format for PHR data is advisable; a widely accepted format is HL7 [14], a standard for information exchange between medical applications and healthcare providers. HL7 includes several recommendations for conceptual representation, documents (included the PHR), applications and messaging standards, and is available as v2.x and v3.0. The v2.x version is a non-XML proposal where data is organized in segments (lines), each containing proper fields and subfields. The v3.0 HL7 messaging protocol leverages XML to provide data structure and also provides support for healthcare workflows.

Other works deal with integration issues, for instance in [15] an OWL ontology is developed to integrate specific medical documents (CCD) authors focus on, whereas in [16] a complete method for ontology based schema and data integration for clinical and genomic databases is presented. Our goal is to provide a tool for extracting and integrating any (neither specific nor structured) medical information without creating a new ontology (rather, exploiting existing ones). Using social forums in healthcare has been investigated e.g. in [17], where answer for medical queries in unresolved posts is provided via similar thread retrieval; to the best of our knowledge, no data gathering from social forum for PHR has been considered so far. Joining virtual social networks and healthcare recently led to neologisms as *Infodemiology* and *Infoveillance* [18]; also Twitter has been exploited in this sense, for instance in [19] the micro-blog is used to detect flu trends, whereas in [5] Ailment Topic Aspect Model is applied to tweets to track public health over time; in [20] authors tracked and

978-83-60810-66-8/$25.00©2015, IEEE

examined disease transmission in particular social contexts via Twitter data, while in [4] social media use improves healthcare delivery by encouraging patient engagement and communication. Apart all these proposals however, no specific use of Twitter data for PHR currently exist. A preliminary and partial study of the work presented in this paper can be found in [21] and [22].

The paper is organized as follows. In section II we describe the overall architecture of our proposal, and how the data extraction and integration are performed for each data source. In section III we show an application to a real case, providing concluding remarks and future works in section IV.

## II. Agent-based gathering system

In fig. 1 our agent-based model is shown. The three main data sources categories we consider are the so-called *database*, that represents standard HIS where official medical personal data are stored (usually, according to a well structured schema), and the *website* and *social*, indicating respectively HTML-based and text-based data sources as described in previous section; in our experiments in particular we looked at online medical forums for the "website" category and Twitter for the "social". Considering agents, the first set (named *wrapper agents*) is devoted to gather personal data from each data source, while *text mining agents* filter previous data to extract medical information. The integration module collect all such data and performs proper integration also exploiting both the SNOMED-CT terminology and an additional dictionary of "common" medical terms (e.g. "headache") that are not stored as official medical terms in the SNOMED-CT. The same references are used by the text mining agents to detect medical information. After this, *feeding agents* are used to populate user's PHR with his relevant medical data. *User agents* collaborate to compare information gathered from different users but semantically related, allowing to build a user network; similarly, *disease agents* cooperate to correlate detected diseases. In the following each component of the proposed architecture is described in more detail.

### A. Wrapper agents

All sources are managed by wrapper agents, that are used to isolate and collect information referring to the same user; in the case of database this is quite trivial and can be accomplished via standard SQL based queries, even if each real HIS has its schema and probably both proprietary solutions as well as authentication issues must be tackled. The question is somehow different for what concern websites and social network. In the case of websites indeed, in particular considering online medical forums we focused on, sometimes all messages are directly available on a per-user basis, therefore the extraction performed by the agent still remains feasible, whereas when forums provide a thread/topic based classification the wrapper agent has to browse all threads/categories and collect all messages for each single user to contribute as much as possible in populating his medical profile; forums where total anonymous messages are allowed are then not considered here (i.e. registered nicknames only). In social networks as Facebook or Twitter, the same approach can be applied since they generally adopt the same organization of traditional websites, i.e. personal messages somehow identifiable even in the case of a topic-based arrangement. Note that the term *person* here refers to a distinguishable user_id or nickname, i.e. the agent stores data together with the *(user_id, source_id)* pair so that each text refers to a single user. Since the nickname should be associated to a real person, this requires he/she should grant the permission for his/her data in order to fulfill privacy issue and related laws; here we suppose that persons do not prevent their personal data to be extracted by agents.

### B. Text mining agents

The next step is the extraction of all medical related data (concerning a given person), discarding other information; this task is performed by text mining agents shown in fig. 1. This agent is not present in the case of HIS since it is likely that such a data source exclusively stores medical related information; conversely, websites and social networks may also include non medical data, also thanks to their semi-structured or unstructured nature, i.e. HTML and text based data posted by common people (not necessarily doctors or medical personnel). Online medical forum and social networks are managed following the same approach, except for a preliminary step in websites, where pure text is extracted from HTML, therefore these sources both provide text data (either webpages or tweets). The first operation we apply is the language recognition, since all further text-mining actions (e.g. stemming) strictly depend on the specific language; in this sense, we discard then all non English text portion. Then, the text mining agent searches for statements containing medical terms; this is accomplished using Natural Language Processing (NLP) techniques [23], in particular first removing irrelevant information (as hyperlink text for web pages, or retweet details and usernames for tweets) and then applying standard text processing operations as tokenization, stopwords removal, stemming and indexing [24]. If the index terms list contains at least a medical term, the statement that term belongs to is preserved, otherwise we discard the statement.

In the last step, we leverage sentiment analysis [25][26] in order to evaluate the remaining statements (those containing medical terms), e.g. if the person suffers a disease cited in a tweet, or if an intolerance to some food specified in a forum question is still present or not. Sentiment analysis or opinion mining [27] leverages NLP, text analysis and computational linguistics to extract subjective information, as the mood of the people regarding a particular product or topic; basically, the sentiment analysis can be viewed as a classification problem of labelling a given text (statement within a tweet or extracted from online forum) as *positive*, *negative* or *neutral*, in our case if the result of classification is positive or negative, the text mining agent passes this information to the integration module to enrich personal medical profile.

To understand how it works, let us consider the text "Last night was too rainy, this morning my headache is stabbing but

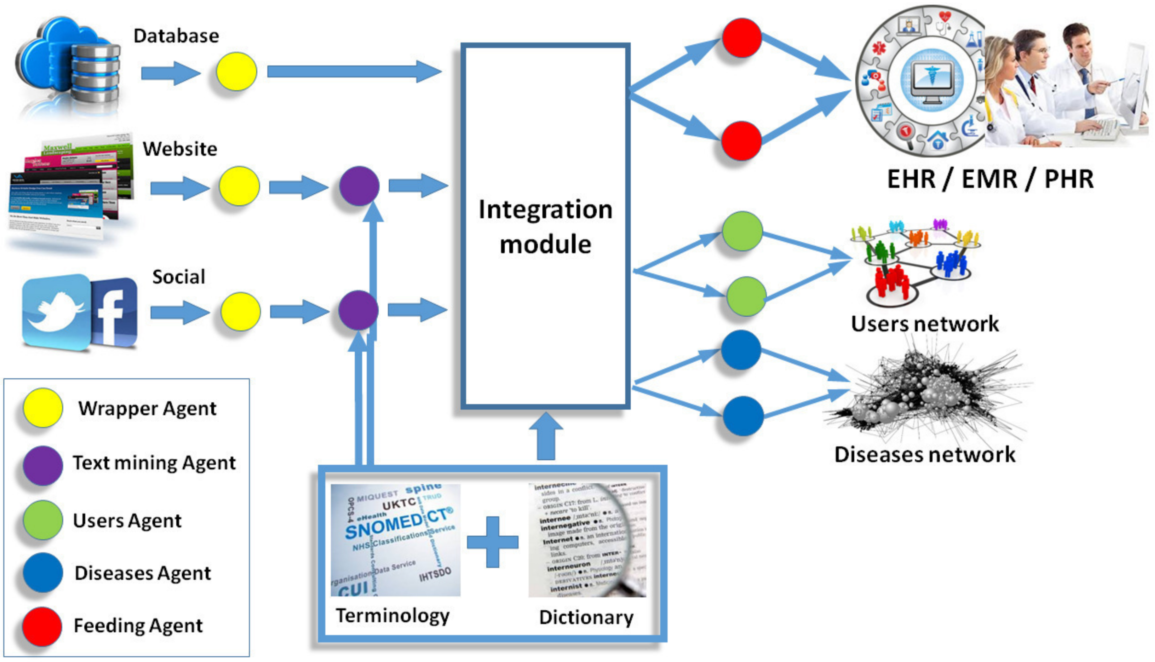


Fig. 1. Application architecture

fortunately I will not go at work" that can come from a social as well as from a website source. Using the NLTK chunking package [28], we first identify short phrases (clusters) like noun phrases (NP) and verb phrases (VP); the text portion cited above in particular produces the following chunks:

”Last night”(NP)
”was” (VP)
”too rainy” (NP)
”this morning” (NP)
”my headache” (NP)
”is stabbing” (VP)
”but fortunately” (NP)
”I” (NP)
”will not go” (VP)
”at work” (NP).

The sentiment analysis exploits the chunking technique first to isolate clusters, then we discard those without medical terms (in the example only ”headache” is present, therefore clusters ”Last night was too rainy” and ”I will not go at work” are discarded), finally trying to assess the meaning of remaining clusters by using a proper list of *positive* and *negative* verbs, so the cluster ”this morning my headache is stabbing” is labelled as *positive* or, in other words, we got the information that person has the headache.

### C. SNOMED-CT and dictionary

In order to discard statements that do not contain any medical term, we search each index term extracted by text mining agent in the SNOMED-CT terminology. To better clarify how this search is performed, we briefly cite the SNOMED-CT core components (details can be found in [29]) that are:

- *concepts*, that represent all entities that characterize health care processes; they are arranged into acyclic taxonomic hierarchies (according to a *is-a* semantics)
- *descriptions*, explaining concepts in terms of various clinical terms or phrases; these can be of three types, Fully Specified Names (FSNs) that is the main (formal) definition, Preferred Terms (PTs), i.e. the most common way of expressing the meaning of the concept, and Synonyms.
- *relationships* between concepts, e.g. the concept (disease) ”Staphylococcal eye infection” has ”Causative agent” relationship with ”Staphylococcus” (different types of relationships exist depending on concepts type)
- *reference sets* used to group concepts e.g. for cross-maps to other standards purposes.

In this work, the first two items are considered, in particular among all concepts hierarchies we focus in the ”disorder/disease” since our goal is to detect tweets about diseases, therefore we do not consider other specific hierarchies (e.g. ”surgical procedures”). Inside the disorder hierarchy, we search each index term extracted from tweets as a FSN, PT or synonym; if found, that tweet is further processed in order to establish whether the specified disorder is present using sentiment analysis (see below).

As indicated in the introduction, to guarantee that all medical terms can be successfully detected, a list of additional informal terms is searched if nothing is found within SNOMED-CT, for instance if the index term is the word ”flu”, this has positive match in the synonym list of ”influenza” disease (the FSN), but the (also quite common) term ”headache” is not explicitly present when browsing SNOMED-CT [30], where this disorder is instead referred as ”migraine” both as FSN and its synonym. Including ”headache” in an additional list (*dictionary* in fig. 1) is the simple solution we adopted; this list is considered just if nothing is found within SNOMED-CT.

Also note that several diseases are defined as a group of words (e.g. ”Viral respiratory infection”), therefore during the indexing phase we also retain N-grams with N=2 and 3; diseases with more than three words can be easily disambiguated even with 3 words since not all words are generally significant (e.g. in ”Disease due to Orthomyxoviridae” the first and the last words are enough for correct matching).

Finally, detected diseases may be hierarchically related, e.g. ”influenza” and ”pneumonia” are both children of ”Viral respiratory infection” according to the ”is-a” semantics. This information could be used for instance by replacing both children with their common parent, in order to build a more generalized, global view of diseases named in the given geographic area during the chosen time period. We choose however to preserve the best level of detail by not using a common ancestor as in the example, while on the other hand we will substitute all terms that represent the same disease with its FSN as indicated in SNOMED-CT, for instance if different tweets refer to ”flu”, ”grippe” and ”influenza” they will be all considered as tweets about ”influenza”.

### *D. Integration module*

The integration of all data concerning the same person into his PHR is performed by the *integration module* after it receives data by wrapper and/or text mining agents. The process is not fully automatic at this stage, so the scenario the user is presented to includes a set of pieces of information that could refer to the same person and have to be somehow integrated.

In particular, the wrapper (or text mining for semi-structured data) agents deliver a set of triples *(user_id, source_id, data)* each referring to a given user, though distinct user_ids, say *UID1* and *UID2* could refer to the same physical person since such triples come from different sources. Whenever *UID1* is literally identical to *UID2*, the system suggests to integrate all related *data* into the same set for further processing, and the user simply can confirm this decision; if however the user believes that although identical those pieces of information do not refer to the same person, for instance when the first *data* is about menopause (female) and the second is about prostate cancer (male), the system stores this decision and rename the second id differently (i.e. *UID2_*) in order to allow disambiguation for further data. In the case *UID1* and *UID2* are different, the user has to decide whether these ids actually represent the same person or not; again, the system stores the decision, therefore if *UID1* is ”Robert Stanton” and *UID2* is ”RBTSTN” and the user establishes that the latter is a portion of ”Robert Stanton”'s fiscal code, all further triples having ”RBTSTN” will automatically incorporated into the same set of ”Robert Stanton” pieces.

After all pieces of health-related information belonging to the same person have been collected into the same set *S*, the integration module tries to integrate *data* present in triples, again through a semi-automatic process. It is important to highlight however that a definitive standard for PHR currently does not exist at all, therefore we do not aim at defining a schema all data should adhere to, rather our goal is to integrate to some extent data concerning the same information.

In particular we first focus on structured data, where the presence of a schema means that *data* in a triple is usually represented as a table, i.e. a query result coming from HIS (databases). Different tables can be integrated first from a structural point of view (e.g. merging ”patient_ID” and ”patient_CODE” columns into a single one) and then from a semantic perspective (e.g. collapsing ”migraine” and ”headache” into the same term); the problem is quite complex but really not new and several solutions already exist [31] [6] [7]. In our first implementation, the user can specify whether he wants a single schema or not; if so, for each table the user has to select which columns are considered (or discarded) and whether columns from different tables must be joined together into a single one; the data type of an output column that mixes two or more existing columns will be the largest data type among those of existing columns being integrated (e.g. float and integer are integrated to float). At the end of the process, a single table comes as output of the integration module and its definition will be XML based, according to HL7 v3.0 format cited in the introduction. Note that data belonging to tables are not integrated in our implementation, rather we simply insert all pieces of information into the single table; if the user do not wish a real integration, all tables are simply preserved and passed as output.

Considering the case of semi-structured data, i.e. website and/or social in fig. 1, *data* stored in a triple is not a relational-like tables, rather it is a labeled statement where the person (identified by the *user_id* element) suffers a given disease (as the headache in the example cited in the text mining module) or has been screened with a given clinical examination etc. In this case the integration we implemented allows to collapse several pieces of information whenever they actually coincide, for instance if we have the statements ”this morning my headache is stabbing” and ”the diagnosis was acute migraine” both labelled as positive and belonging to the same set *S* concerning a given person *user_id*, the system allows the user to discard one of the two statements since they represent the same information. Note that the system recognizes the statements as comparable since it exploits SNOMED-CT and the dictionary of informal terms to map all medical terms (in this case, ”headache” and ”migraine”) into the same term (in this case, the FSN ”migraine”); if such a mapping does not occur, statements are supposed to be different, but the system always allows the user to collapse a set of statements into a single one manually if this is the case.

### *E. Feeding agent*

After the integration process has been performed, a (possibly reduced) set of triples *(user_id, source_id, data)* is provided by the integration module for PHR feeding. The way

PHR is actually stored and accessed has not been definitively standardized [32], however in our architecture PHR should be available as web services to guarantee an easy and uniform access. In this sense, a number of platforms have been proposed, as the Microsoft's HealthVault [33], PatiensLikeMe [34] or other proprietary solutions, in addition to institutional approaches, e.g. [35]. Since each one of them has its features, we specify only general guidelines of the actual implementation of the feeding agent, in particular each agent is devoted to a specific solution and has to manage authentication issues as well as data format and communication protocol; we suppose however that supporting HL7 v3.0 as XML-based data format is the best choice

### *F. User and disease agents*

*Users agents* are devoted to build and manage the network of users whose data have been extracted. The network can be built according to different criteria, for instance two users may be considered as *linked* if they "share" the same disease, or they were admitted at the same hospital. Similarly, a set of *Diseases agents* can build a network of diseases somehow related, e.g. a link between two diseases may represent a co-occurrence of both diseases in a significant number of data concerning the same user, or the fact that they are cited by related users and so on; the establishment of such network leverages the SNOMED-CT to tackle semantic-related issue (e.g. synonyms, homonyms, hypernyms). At this stage these two set of agents are considered for future works and have not been implemented.

## III. Results

In this section we show an example of how the architecture works. In particular, we considered a small group of persons who suffer different diseases and are in treatment at the same medical center; in fig. 2 we show one of them with two so-called *medical problems*, i.e. diabetes and osteoporosis. The Health Information System used in this centre was actually a customized version on the OpenEMR software [36], whose data access were granted to the wrapper agent through an API-based connection to the underlying MySQL database. The wrapper agent submitted an SQL query with users names, in order to get information about their exams; in fig. 3 we show the resulting table for the same person of fig. 2; for the sake of simplicity, we omitted query details, i.e. the foreign keys used to join patient with medical problems tables, and only significant columns are shown in fig. 2 (patient name and related medical problems). As specified in the previous section, the table will be the *data* field in the triple *('Rebecca Greenfield', 'SRC#002', data)* extracted from MySQL database for that person by the wrapper module and delivered to the integration module. Note that to get results, we supposed (as specified in the introduction) that the person in fig. 2 grants the wrapper agent to access her data, in order to overcome authentication issues.

Similarly, other wrapper agents search on websites and/or social networks, in particular we considered [37] (a medical forum) as "website" data source, and, supposed that we are able to associate (even manually) the name stored in the HIS to the nickname "rebgreen46", we allow the wrapper agent to extract from HTML the text contained within posts (fig. 4 shows one of them). The text is further processed by the text mining agent, as discussed in previous section, therefore we get the following triples:

- *('rebgreen46', 'SRC#003', ('back', 'hurts', POSITIVE))*
- *('rebgreen46', 'SRC#003', ('legs and shoulders', 'aches', POSITIVE))*
- *('rebgreen46', 'SRC#003', ('pain', 'is getting worse', POSITIVE))*

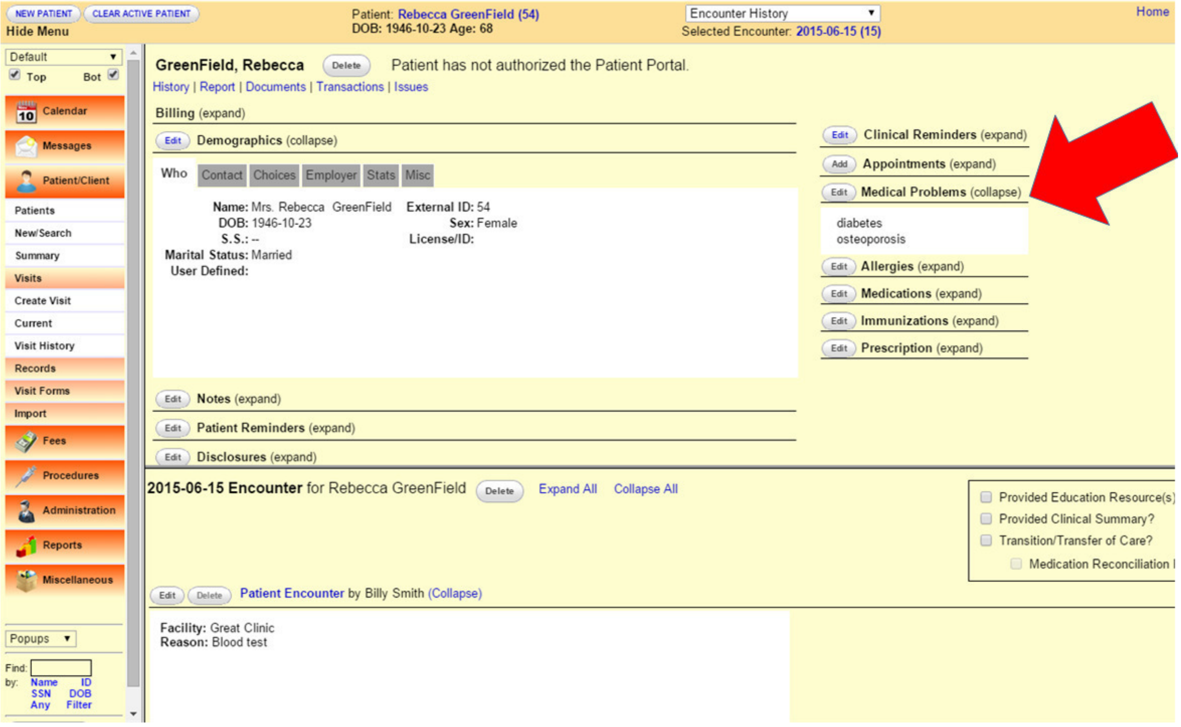
These triples represent therefore additional pieces of information about that person, i.e. from the database emerges that she suffers diabetes and osteoporosis, whereas from this forum the presence of aches for back, legs and shoulders is detected. According to the procedure described in the previous section, the integration module allow to associate these three triples with the table (in this case, manually specifying that 'Rebecca Greenfield' is the same *user_id* as 'rebgreen46'); one or more feeding agents will finally connect to the platform(s) where these pieces of information will be added to the Personal Health Record for that patient. Apart the simple example shown so far, we are currently undergoing on a more significant test with a relevant number of patients (about 200) also considering Twitter as "social" data source.

## IV. Conclusions

The work described in this paper outlines an agent-based architecture for feeding PHR with data extracted from structured and semi-structured (databases and website/social networks, respectively) sources. Our proposal is currently at an early stage, and we believe that the first implementation will provide us with results confirming prosiming expectations. Moreover, the implementation will allow to assess the effectiveness of our approach, expecially when a relevant number of data sources as well as a relevant number of patients to extract data about will be gathered; this validation is also required to evaluate the quality of integrated infomation with respect to that stored into original sources (e.g. redundancy, completeness and so on).

Some future works are planned:

- a useful extension concerning the integration module is the possibility of specifying a mapping function so that the system can be trained a-priori to associate different *user_id* into the same person, for instance the mapping function can be a regular expression acting as a bijective function to convert *UID1* to *UID2* and viceversa. More complex function could be defined to allow automatic integration.
- the integration module presented in this paper is quite elementary; we are considering more effective solutions, e.g. a classifier (supervised or not) that tries to aggregate health data without the user's intervention or reducing this as much as possible. Besides, more effective data


NEW PATIENT
CLEAR ACTIVE PATIENT
Hide Menu
Encounter History
Home
Default
Top
Bot
Calendar
Messages
Patient/Client
Patients
New/Search
Summary
Visits
Create Visit
Current
Visit History
Records
Visit Forms
Import
Fees
Procedures
Administration
Reports
Miscellaneous
Popups
Find:
by: Name ID SSN DOB Any Filter
Delete
History | Report | Documents | Transactions | Issues
Billing (expand)
Edit Demographics (collapse)
Who Contact Choices Employer Stats Misc
Edit Clinical Reminders (expand)
Add Appointments (expand)
Edit Medical Problems (collapse)
Edit Allergies (expand)
Edit Medications (expand)
Edit Immunizations (expand)
Edit Prescription (expand)
Edit Notes (expand)
Edit Patient Reminders (expand)
Edit Disclosures (expand)
Delete
Expand All
Collapse All
Provided Education Resource(s)
Provided Clinical Summary?
Transition/Transfer of Care?
Medication Reconciliation
Edit Delete

Patient: Rebecca GreenField (54)
DOB: 1946-10-23 Age: 68

Selected Encounter: 2015-06-15 (15)

GreenField, Rebecca

Patient has not authorized the Patient Portal.

Name: Mrs. Rebecca GreenField
DOB: 1946-10-23
S.S.: --
Marital Status: Married
User Defined:
External ID: 54
Sex: Female
License/ID:

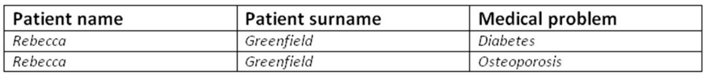
diabetes
osteoporosis

2015-06-15 Encounter for Rebecca GreenField

Patient Encounter by Billy Smith (Collapse)

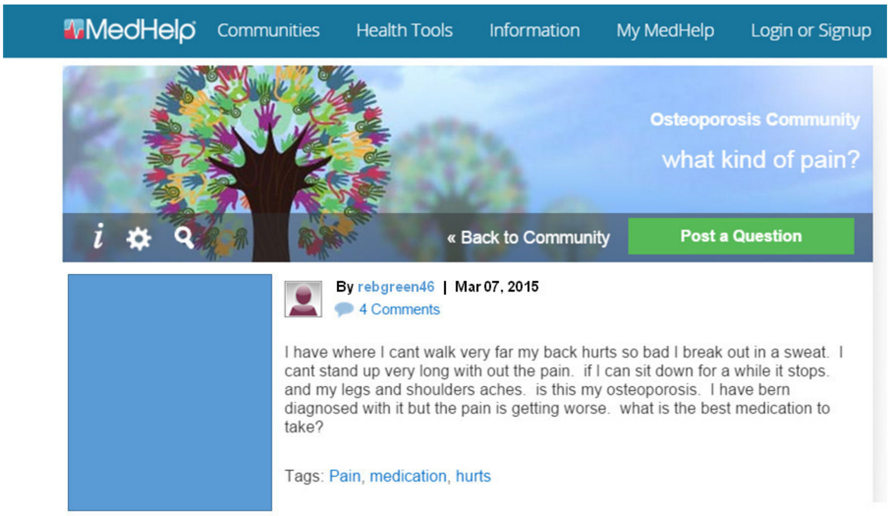
Facility: Great Clinic
Reason: Blood test

Fig. 2. Snapshot from HIS (structured data source)



| Patient name | Patient surname | Medical problem |
|---|---|---|
| *Rebecca* | *Greenfield* | *Diabetes* |
| *Rebecca* | *Greenfield* | *Osteoporosis* |

Fig. 3. Data extracted from HIS database by wrapper agent



By rebgreen46 | Mar 07, 2015
4 Comments

I have where I cant walk very far my back hurts so bad I break out in a sweat. I cant stand up very long with out the pain. if I can sit down for a while it stops. and my legs and shoulders aches. is this my osteoporosis. I have bern diagnosed with it but the pain is getting worse. what is the best medication to take?

Tags: Pain, medication, hurts

Fig. 4. A post extracted from a forum webiste by wrapper agent

integration mechanisms for either structured or semi-structured data sources (or both) should be investigated, as for instance agent-based approaches, also in order to fully exploit the cooperation of agents we outlined in the main architecture to provide effective automation in the integration process

- finally, a significant further work is required to implement both user and disease agents and to explore how to leverage corresponding networks to improve integration and the quality of health-related information.

## Acknowledgment

This work was supported by the "Programma Operativo Nazionale Ricerca & Competitivita' 2007–2013" within the project "PON04a2_C - Smart Health 2.0".

## References

[1] K. J. Cios and W. Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine*, vol. 26, pp. 1–24, 2002.

[2] R. Haux, "Health information systems - past, present, future." *I. J. Medical Informatics*, vol. 75, no. 3-4, pp. 268–281, 2006. [Online]. Available: http://dblp.uni-trier.de/db/journals/ijmi/ijmi75.html#Haux06

[3] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1186/2047-2501-2-3

[4] J. Fisher and M. Clayton, "Who gives a tweet: Assessing patients interest in the use of social media for health care," *Worldviews on Evidence-Based Nursing*, vol. 9, no. 2, pp. 100–108, 2012. [Online]. Available: http://dx.doi.org/10.1111/j.1741-6787.2012.00243.x

[5] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health." in *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011. [Online]. Available: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html#PaulD11

[6] C. Batini, M. Lenzerini, and S. B. Navathe, "A comparative analysis of methodologies for database schema integration," *ACM Comput. Surv.*, vol. 18, no. 4, pp. 323–364, December 1986. [Online]. Available: http://doi.acm.org/10.1145/27633.27634

[7] A. Doan and A. Y. Halevy, "Semantic-integration research in the database community," *AI Mag.*, vol. 26, no. 1, pp. 83–94, March 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1090488.1090497

[8] J. F. Dipnall, M. Berk, F. N. Jacka, L. J. Williams, S. Dodd, and J. A. Pasco, "Data integration protocol in ten-steps (dipit): A new standard for medical researchers," *Methods*, vol. 69, no. 3, pp. 237 – 246, 2014, recent development in bioinformatics for utilizing omics data. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1046202314002382

[9] C. Heeter, "{EHR} progress and future outlook," {*AORN*} *Journal*, vol. 97, no. 3, pp. C7 – C8, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001209213001506

[10] S. Sachdeva and S. Bhalla, "Semantic interoperability in standardized electronic health record databases," *J. Data and Information Quality*, vol. 3, no. 1, pp. 1:1–1:37, May 2012. [Online]. Available: http://doi.acm.org/10.1145/2166788.2166789

[11] A. Sheth, S. Agrawal, J. Lathem, N. Oldham, H. Wingate, P. Yadav, and K. Gallagher, "Active semantic electronic medical record," in *The Semantic Web - ISWC 2006*, ser. Lecture Notes in Computer Science, I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, Eds. Springer Berlin Heidelberg, 2006, vol. 4273, pp. 913–926. [Online]. Available: http://dx.doi.org/10.1007/11926078_66

[12] A. Baird, F. North, and T. S. Raghu, "Personal health records (phr) and the future of the physician-patient relationship," in *Proceedings of the 2011 iConference*, ser. iConference '11. New York, NY, USA: ACM, 2011, pp. 281–288. [Online]. Available: http://doi.acm.org/10.1145/1940761.1940800

[13] D. Lee, R. Cornet, F. Lau, and N. de Keizer, "A survey of snomed-ct implementations," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 87 – 96, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046412001530

[14] Health Level Seven international, *http://www.hl7.org/index.cfm.*

[15] J. Puustjärvi and L. Puustjärvi, "Ontology-based integration of clinical documents," in *Proceedings of the 14th International Conference on Information Integration and Web-based Applications &#38; Services*, ser. IIWAS '12. New York, NY, USA: ACM, 2012, pp. 342–347. [Online]. Available: http://doi.acm.org/10.1145/2428736.2428799

[16] D. Perez-Rey, V. Maojo, M. Garca-Remesal, R. Alonso-Calvo, H. Billhardt, F. Martin-Snchez, and A. Sousa, "Ontofusion: Ontology-based integration of genomic and clinical databases," *Computers in Biology and Medicine*, vol. 36, no. 78, pp. 712 – 730, 2006, special Issue on Medical Ontologies. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010482505000740

[17] J. H. D. Cho, P. Sondhi, C. Zhai, and B. R. Schatz, "Resolving healthcare forum posts via similar thread retrieval," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, ser. BCB '14. New York, NY, USA: ACM, 2014, pp. 33–42. [Online]. Available: http://doi.acm.org/10.1145/2649387.2649399

[18] G. Eysenbach, "Infodemiology and Infoveillance," *American Journal of Preventive Medicine*, vol. 40, no. 5, pp. S154–S158, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.amepre.2011.02.006

[19] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, April 2011, pp. 702–707.

[20] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic," *PLoS One*, vol. 6, no. 5, 2011.

[21] A. Longheu, V. Carchiolo, and M. Malgeri, "Personal health record feeding via medical forums," in *Computer Supported Cooperative Work in Design (CSCWD), Proceedings of the 2014 IEEE 19th International Conference on*. IEEE, 2015. [Online]. Available: toappear

[22] A. Longheu, V. Carchiolo, and MicheleMalgeri, "Medical data integration with snomed-ct and hl7," in *New Contributions in Information Systems and Technologies*, ser. Advances in Intelligent Systems and Computing, A. Rocha, A. M. Correia, S. Costanzo, and L. P. Reis, Eds., vol. 353. Springer International Publishing, 2015, pp. 1165–1171. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16486-1_115

[23] P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization, 2nd ed.* Amsterdam: John Benjamins, 2007.

[24] R. Baeza-yates and B. Ribeiro-Neto, *Modern Information Retrievial.* Seattle, Washington, United States: ACM Press, 1999.

[25] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093 – 1113, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2090447914000550

[26] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proceedings of the First ACM Conference on Online Social Networks*, ser. COSN '13. New York, NY, USA: ACM, 2013, pp. 27–38. [Online]. Available: http://doi.acm.org/10.1145/2512938.2512951

[27] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, January 2008. [Online]. Available: http://dx.doi.org/10.1561/1500000011

[28] Natural Language Toolkit chunk package, *http://www.nltk.org/api/nltk.chunk.html.*

[29] SNOMED CT, *http://www.ihtsdo.org/snomed-ct.*

[30] IHTSDO SNOMED CT Browser, *http://browser.ihtsdotools.org/.*

[31] A. Longheu, V. Carchiolo, and M. Malgeri, "Schema and data integration for relational and object-oriented data sources," in *Computer Science and Informatics (CSI2002), Proceedings of the 2002 Sixth International Conference on*, 2002.

[32] Fast Healthcare Interoperability Resources, http://hl7.org/implement/standards/fhir/.

[33] Microsoft HealthVault, https://www.healthvault.com/it/en-US.

[34] Patientslikeme, https://www.patientslikeme.com/.

[35] PHR - Medline, http://www.nlm.nih.gov/medlineplus/personalhealthrecords.html.

[36] OpenEMR, http://open-emr.org/.

[37] MedHelp, http://www.medhelp.org/.

# 9th International Workshop on Multi-Agent Systems and Simulation

MULTI-AGENT systems (MASs) provide powerful models for representing both real-world systems and applications with an appropriate degree of complexity and dynamics. Several research and industrial experiences have already shown that the use of MASs offers advantages in a wide range of application domains (e.g. financial, economic, social, logistic, chemical, engineering). When MASs represent software applications to be effectively delivered, they need to be validated and evaluated before their deployment and execution, thus methodologies that support validation and evaluation through simulation of the MAS under development are highly required. In other emerging areas (e.g. ACE, ACF), MASs are designed for representing systems at different levels of complexity through the use of autonomous, goal-driven and interacting entities organized into societies which exhibit emergent properties The agent-based model of a system can then be executed to simulate the behavior of the complete system so that knowledge of the behaviors of the entities (micro-level) produce an understanding of the overall outcome at the system-level (macro-level). In both cases (MASs as software applications and MASs as models for the analysis of complex systems), simulation plays a crucial role that needs to be further investigated.

## Topics

MAS&S'15 aims at providing a forum for discussing recent advances in Engineering Complex Systems by exploiting Agent-Based Modeling and Simulation. In particular, the areas of interest are the following (although this list should not be considered as exclusive):

- Agent-based simulation techniques and methodologies
- Discrete-event simulation of Multi-Agent Systems
- Simulation as validation tool for the development process of MAS
- Agent-oriented methodologies incorporating simulation tools
- MAS simulation driven by formal models
- MAS simulation toolkits and frameworks
- Testing vs. simulation of MAS
- Industrial case studies based on MAS and simulation/testing
- Agent-based Modeling and Simulation (ABMS)
- Agent Computational Economics (ACE)
- Agent Computational Finance (ACF)
- Agent-based simulation of networked systems
- Scalability in agent-based simulation

## Steering Committee

**Cossentino, Massimo,** ICAR-CNR, Italy
**Fortino, Giancarlo,** Universita della Calabria, Italy
**Gleizes, Marie-Pierre,** Universite Paul Sabatier, France
**Pavon, Juan,** Universidad Complutense de Madrid, Spain
**Russo, Wilma,** Universita della Calabria, Italy

## Event Chairs

**Fortino, Giancarlo,** Universita della Calabria, Italy
**Fuentes-Fernández, Rubén,** Universidad Complutense de Madrid, Spain
**Migeon, Frederic,** IRIT - University of Toulouse
**Seidita, Valeria,** Università degli Studi di Palermo, Italy

## Program Committee

**Antunes, Luis**
**Arcangeli, Jean-Paul,** Université Paul Sabatier, France
**Bernon, Carole,** Universite Paul Sabatier, France
**Botía, Juan,** Universidad de Murcia, Spain
**Botti, Vicente**
**Cossentino, Massimo,** ICAR-CNR, Italy
**Davidsson, Paul,** Malmö University, Sweden
**Garro, Alfredo,** University of Calabria, Italy
**Gomez-Sanz, Jorge J.,** Universidad Complutense de Madrid, Spain
**Gravina, Raffaele,** University of Calabria, Italy
**Guerrieri, Antonio,** University of Calabria, Italy
**Hassan, Samer,** Universidad Complutense de Madrid, Spain
**Jedrzejowicz, Piotr,** Gdynia Maritime University, Poland
**Klügl, Franziska,** Örebro Universitet, Sweden
**López-Paredes,** Adolfo, INSISOC - University of Valladolid, Spain
**Lorscheid, Iris**
**Molesini, Ambra**, Università di Bologna, Italy
**Niazi, Muaz,** COMSATS Institute of IT, Pakistan
**Nunes, Ingrid,** UFRGS
**Petta, Paolo,** OFAI, Austria
**Picard, Gauthier,** EMSE, Saint Etienne, France
**Ribino, Patrizia,** Istituto di Reti e Calcolo ad Alte Prestazioni - Consiglio Nazionale delle Ricerche
**Terna, Pietro,** Università di Torino, Italy
**Vasconcelos, Wamberto,** University of Aberdeen, United Kingdom
**Vizzari, Giuseppe,** Università di Milano Bicocca, Italy

DOI: 10.15439/2015F287
ACSIS, Vol. 5

# A Heuristic for Problem Formalization in Agent Based Simulation Studies

Massimo Cossentino[‡], Carmelo Lodato[‡], Patrizia Ribino[‡] and Valeria Seidita[§][‡]
[‡]Istituto di Reti e Calcolo ad Alte Prestazioni
Consiglio Nazionale delle Ricerche
Palermo, Italy
Email: {cossentino,c.lodato,ribino}@pa.icar.cnr.it
[§]Dip. di Ingegneria Chimica Gestionale Informatica Meccanica, University of Palermo
Palermo, Italy
Email: {valeria.seidita}@unipa.it

***Abstract*—Agent Based Modeling and Simulation (ABMS) is considered an effective approach for conducting simulation studies in many fields. In order to develop high quality simulation models, methodological approaches are demanded. In such direction we are moving by proposing a heuristic for the formalization of agent based simulation problems. The proposed heuristic is based on some guidelines developed for identifying the main elements of the problem domain description by analysing verbs and their common taxonomy in grammar.**

## I. Introduction

SIMULATIONS are used in several contexts for evaluating the behaviour of complex systems and for understanding how the numerous variables constraint the performance of such systems. Especially, a simulation model gives the opportunity to make experiments and to identify errors in such a way that would often be infeasible in the real world. A simulation study is a process that allows to define a simulation model of a real system and to make experiments with that model.

Among simulation models, Agent Based Modelling and Simulation (ABMS) is becoming a widely used approach for conducting simulation studies in many disciplines (such as social sciences, economy, traffic and transportation) that deal with complex systems characterized by the presence of autonomous and active entities [1, 2, 3, 4, 5, 6, 7].

Such increasing interest is mainly due to the possibility to include in such models some aspects that are commonly not considered by traditional approaches, such as local interactions, entity organizations, impact of environmental changes on entity behaviours, entity heterogeneity and so on.

Whether on the one hand these features make the ABMS paradigm very useful for the representation of highly quality models, on the other hand it requires appropriate methodological approaches for addressing such a kind of simulation study. At present, there are some proposals in such direction but several issues are still open and a comprehensive methodological approach is still lacking.

The work we present in this paper is part of a more ambitious objective: the development of a complete methodological approach (we named MAMAS, *Methodological Approach for Multi-Agent Simulations*) that covers the entire life cycle of a simulation study taking into account each facet of the agent based modelling paradigm. In so doing, the experience we collected in the latest years about methodological approaches in the field of agent oriented software engineering (hereafter AOSE) consolidated our opinion that a solid foundation for design processes lies on the use of system metamodels. This is also recognized in the field of agent based simulations where some proposals about system metamodels [8, 9, 10, 11] for creating a common ground for several agent based simulation domains are proposed. For these reasons, we are defining MAMAS approach starting from the metamodel we proposed in [11]. Specifically we are specializing the set of activities that are commonly defined for generic simulation studies in the ABMS context by using the metamodel as activity's benchmark. Thus we are defining specific guidelines and work products for these activities wherever are not defined yet.

The importance of the problem formalization in order to produce complete and well defined simulation models is well recognized in literature [12, 13]. In our work, problem formalization assumes a very important role for the identification of simulation goals and all the multi-agent system goals influencing the simulation goal itself. The concept of goal, in our approach, spreads over the whole MAMAS process. For this reason we spent a great effort in studying a formalized way for identifying and modelling goals. We also studied how they influence the whole simulation study, starting from the very early stages of simulation analysis i.e. the *problem formalization*.

In such a landscape, the aim of this paper is to provide guidelines for performing a core activity of MAMAS process: *Problem Formalization*; the proposed contribution lies on the heuristic developed for identifying the main elements of the problem domain by using verbs and their common taxonomy in grammar.

The rest of the paper is organized as follows: in section II we detail the motivations for our work against the related works in literature, in section III we present a brief overview on the MAMAS process; in section IV we illustrate the proposed heuristic and finally some discussions and conclusions are drawn in section V.

978-83-60810-66-8/$25.00©2015, IEEE

## II. Motivation and Related Work

The advantages of simulation models for studying the behaviour of complex and dynamic systems are widely known. In particular multi agent models are among the most interesting ones due to some useful properties showed by this type of modelling [6]. Moreover, multi agent simulations is considered one of the killer applications of agent oriented technology [14]. Nevertheless, as far as we know, to date there is no comprehensive methodological approach that can provide guidelines for the development of such models. Indeed, traditional approaches are conceived for generic simulation studies and they do not adapt well to agent based simulations because they overlook some essential aspects that are fundamental in an ABMS approach, such as for example the role of the environment or agent interactions.

Only to cite a few, Balci in [15] suggests a process based on 10 phases and 13 credibility assessment stages. The credibility assessment starts from the first phase of problem formulation until the last phase of the presentation of simulation results. In [16] instead Balci states the importance of the VV&T (Validation Verification and Testing) techniques throughout the life cycle of a simulation study showing how they are categorized. Although these two works established a baseline for the development of credible simulation studies they are conceived for generic simulation approaches. We think that, from the perspective of agent-based simulations, the life cycle of a simulation study has to take into account some typical features of agent oriented technology. For example, it is important to focus on the environment model that plays a fundamental role in such kind of simulations where it commonly represents the real context in which the real actors work.

In [12] John S. Carson defines the human actors that have to compose the simulation team relating them to the common steps of a simulation study and providing some useful hints in order to conduct the activities of the process. Just to give an example, Carson states that during *Problem Formulation* the team should develop a list of specific questions the model should address and he also states that it has to focus on model boundary and scope, level of detail and project scope. While during the *Model Development* phase he suggests two main activities have to be performed: *(i)* development of data structures to represent the data needed by the model and *(ii)* translation of the modelling assumptions in a specific document written in the language or representation required by the adopted simulation package. Hence, although the guidelines of *Problem Formulation* are surely helpful also in the context of agent based simulations, the second one is not applicable due to the different technological simulation paradigm.

In [13] Averill M. Law proposes a seven-step approach for conducting a simulation study. In his work, Law highlights the importance of the validity of a simulation model. Thus, he presents practical techniques and guidelines for developing valid and credible models. In particular, he declares that the keystone for building valid and credible models lies on (i) formulating the problem precisely; (ii) interacting with the decision-maker on a regular basis throughout the simulation project to ensure that the correct problem is being solved and to promote model credibility; (iii) using quantitative techniques to validate components of the model; (iv) performing sensitivity analysis to determine important model factors; (v) comparing model and system results for an existing system (if any); (vi) using a Turing test to compare model and system output data; (vii) reviewing of model results and animations to see if they appear to be reasonable. As well as the process proposed by Balci in [15], the process proposed by Law covers the whole life cycle of a simulation study providing right-minded guidelines for each activity. But as the previous ones, such guidelines are not effective enough for agent based simulation studies because too general.

At the same way, as far as we know, the methodological approaches currently proposed for agent oriented modelling give a rundown of the overall life cycle of a simulation study and only in some cases they focus on single steps, such as for example model definition and validation.

In [17] the authors present a model-based methodological framework for designing multi-agent simulations. The aim of the authors in this paper is the introduction of a consistent use of the agents during the entire life cycle of a simulation study. In particular they individuate three roles involved in a MABS (Multi-Agent Based Simulation) the thematicians that are practically the domain experts, the modellers that have the responsibility to design the model and finally the computer scientist that have to implement the model. Each role deals with abstractions of the agent at different levels (i.e: real agents, conceptual agents and computational agents) thus defining the specific kind of model related to the level he belongs to (i.e: domain model, design model and operational model). At that time, in their operational model the computational agents do not own the peculiar features of agents such as proactivity or autonomy due to the lack of development language supporting them. At the time was common the use either object-oriented or procedural or functional languages to implement the specifications described in the conceptual agents. These deficiencies are nowadays overcome by the several agent-oriented development paradigm. Hence according to Drogoul *et.al*, we believe that a methodological approach for simulation study has to be based on model transformations from different levels of abstractions but we also want to fill the gap between the conceptual model and the operational one.

In [18] the authors proposed a standard protocol for conceiving agent-based models, named ODD (Overview, Design concepts, Details). Such protocol provides guidelines in order to structure the information needful for agent-based models in an established sequence following a top down approach. It is composed of seven steps grouped in three main blocks: Overview, Design concepts and Details. The core of the protocol is the description of the design concepts following a detailed check-list of questions that allow to examine particular aspects of the system to be simulated such as emergence, adaptation, prediction etc...

The easyABMS methodology [19] is based on an iterative process that covers the common activities of a simulation study and produces several models according to the specific phase of the simulation study. Such models are produced according to a reference meta-model that is characterized by elements that are quite common in the AOSE context. It lacks of explicit representation of space, entity features and organizational structure.

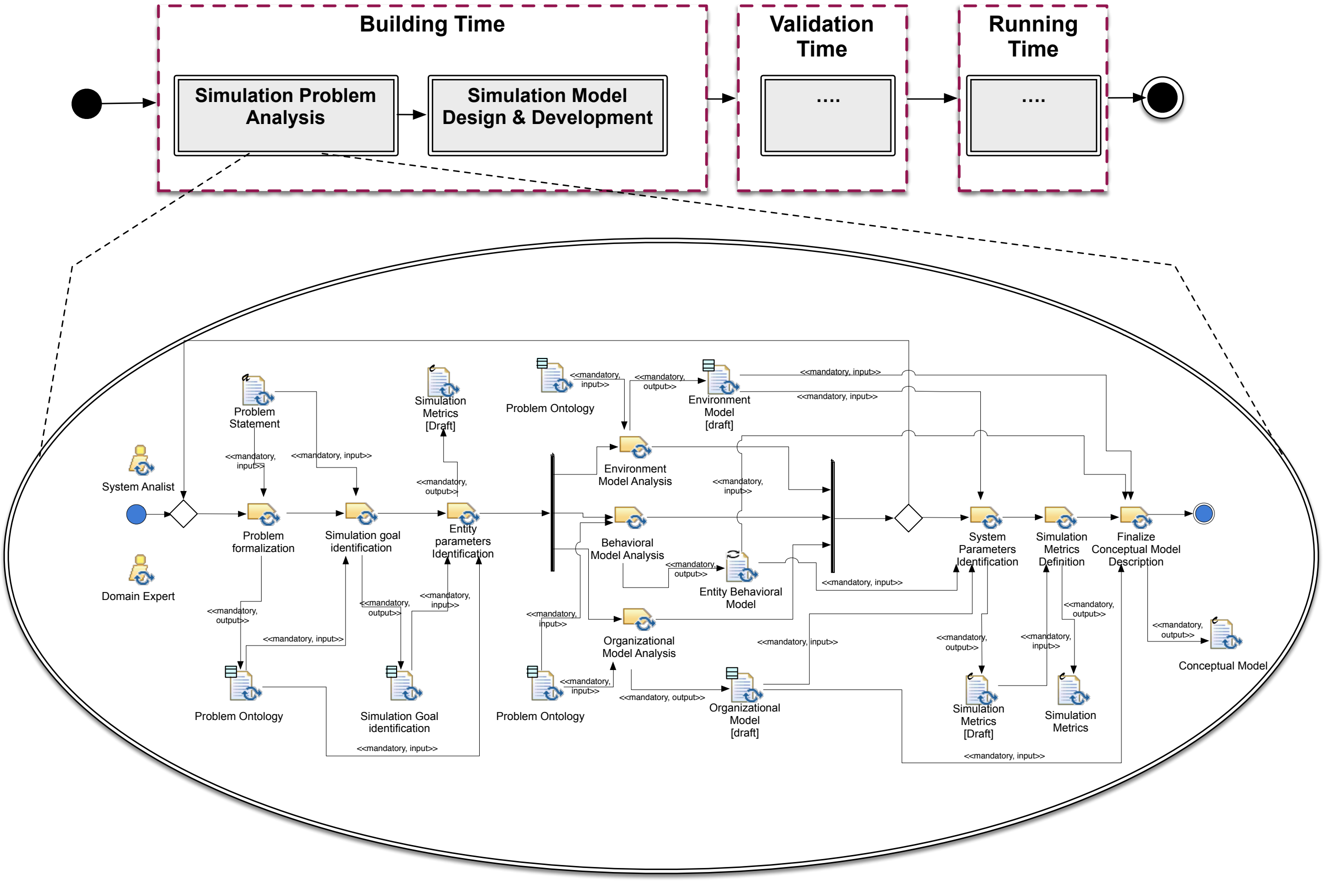


Fig. 1. A sketch of MAMAS process with zooming on Simulation Problem Analysis Phase

In [20] Cioffi introduces a general methodology for social simulations. It is an iterative process that begins with a referent system in the real world (he names *explanandum*). Then techniques of abstraction, formalization, programming are used to develop a simulation model (he names *explanans*). As well as techniques for Verification and Validation are reported such as Code Walk-Through, Parameter Sweeps, Histograms and so on. Social simulations find a natural representation by means of agent oriented techniques. As Cioffi states, agent based models have a more sophisticated landscape and actors that come closer to emulating humans through various aspects of reasoning, decision- making, and behaviours. But he only focuses on three main components of the agent based model (i.e: agents, rules, and environments) thus adapting the methodology only for this aspects.

Agent-Based Social Simulations are also the core of MAIA [21]. MAIA is a framework based on a meta-model that supports conceptualization of a agent-based simulations. It provides some guidelines in order to adequately capture, analyse, and understand the domain of application. It helps the modellers to explicitly report the motivations behind modelling choices. Such framework is focused on social and institutional structures and its meta-model aims to describe those systems where the key components are individuals and institutions. We think that some other model components have to be considered when we address agent based simulation studies. In [11] we conducted a study that highlighted the presence of an organizational aspect in many practical agent based simulations. According to the authors of [20] and [21] we claim that a methodological approach should certainly consider agents, rules, norms and environments but also other aspects such as the organizational one.

In [22] the focus is on the use of simulation in biological systems. In such a context the authors propose a revision of the Agent and Artifact (A&A) metamodel for adaptation to the Systems Biology, they made experiments using the case study of glycolysis and shown how to model and simulate the metabolic pathway.

In [23] the authors propose a method to integrate simulation-based approach within AOSE methodologies as a new fragment of the methodology. The main idea behind the use of simulation in the software engineering process is that the simulation can be used to predict the dynamics of the system to be developed and also the run-time properties that can be induced by design, before the system is completely developed.

From the literature it arises that several issues have to be still addressed for the definition of an appropriate methodological approach for agent based simulation studies. This work goes in such direction by addressing a crucial aspect: the *Problem Formalization*.

## III. AN OVERVIEW OF MAMAS PROCESS

In order to understand the aim of this paper we need to briefly introduce the context in which this contribution is plugged in. As we previously said, we are working on the development of a complete methodological approach for addressing agent based simulation studies (hereafter MAMAS) taking into account specific facets of agent based simulation paradigm.

From a literature review (see Section II), it arises that the main stages of a generic simulation study are: *(i)* the simulation model development, *(ii)* its validation and *(iii)* the execution of experiments on such model with the related analysis. We have defined in the MAMAS process three macro phases according to the stage they refer, we named them respectively *Building Time*, *Verification&Validation Time* and *Running Time* (see Fig. 1).

For the scope of this paper, let us focus on the *Building Time* phase. The final aim of this phase is to produce the agent based simulation model. This usually requires the transition from real-world system to a simulated one throughout four main artefacts:

- the *Problem Domain Description*, also called *Problem Statement*, containing an informal description both of the real-world problem to be addressed by the simulation study and of the simulation objectives;
- the *Domain Model*, also called *Conceptual Model*, that is the model of the system to be simulated emerging from the study of the *Problem Domain*. This model contains a more formalized form of knowledge about the *Problem Domain*. In particular, it may contain knowledge about real-world agents, behaviours, rules and so on;
- the *Design Model* is the model derived from the Domain Model by describing the details of the architecture of the agent based system that have to be realized for simulating the real system. A Design Model contains knowledge about design agents, interaction, behaviours that is a formal refinement of the previous one.
- the *Computational Model* is the implementation of the Design Model on a specific agent platform. It handles the computational agents that are implementation on a specific agent platform.

For supporting this transition in a systematic way, several activities are necessary. In MAMAS process, we grouped them in two phases: *Simulation Problem Analysis* and *Simulation Model Design & Development* respectively. The former supports the transition between the *Problem Domain Description* to the *Domain Model*, while the second produces the *Design Model* and the *Computational Model*. These last phase could be performed by using classical AOSE approach.

Now, let us focus on the *Simulation Problem Analysis* phase (see the zooming of Fig.1). According to the ABMS literature, we recognize the importance of the problem formalization as a mean for producing well defined simulation models. We have already experienced (see [24]) in the AOSE context how to perform the problem formalization through the use of ontological approaches in a way useful for identifying goals of multi agent systems that are the leading elements for the development of design models. Such goal identification approach is based on the definition of a *Problem Ontology* starting from the *Problem Statement* of the multi agent system to be developed.

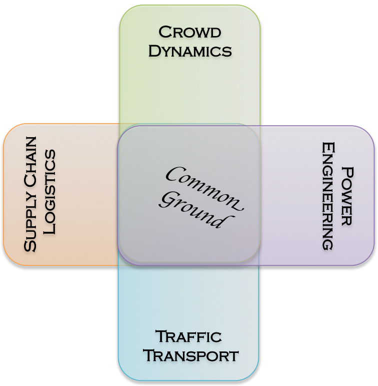


Fig. 2. Several simulation problem domains share a common ground that aims to provide a starting point for the development of an agent based simulation model.

In the same way, we think that in the ABMS context simulation goals along with agent goals are leading elements for the development of the conceptual model. Thus we are adapting the approach proposed in [24] for the identification of simulation goals and all the agent goals that influence the simulation goals itself. It is worth noting that in this context the concept of goal assumes different meanings. Commonly in generic simulation studies the term *goal* is used for referring to the simulation objectives, namely what are the real issues the simulation has to address. In such a sense, simulation goals are linked to the interest that the simulation team sees in the results of the simulation study. Conversely, agent goals refer to the states of the world an agent wants to achieve. Thus, they are linked to the interest the single entities have in the system they live in. This system is the means the simulation approach uses to achieve the intended simulation goals.

Thus, in order to adapt the approach we proposed in [24] in the context of ABMS, we need to define appropriate guidelines for building the *Problem Ontology* starting from the *Problem Domain Description* of a simulation study. Such guidelines are founded on the metamodel for agent based simulation problems proposed in [11] that will be presented in the following sections.

A detailed description of all the activities shown in Fig.1 is out of the scope of this paper, however it is important to note the role of the problem Ontology as input of several activities thus making necessary the problem formalization we propose.

## IV. PROBLEM FORMALIZATION

In [11] we already addressed the definition of a metamodel for agent based simulation problems. In that work we adopted a systematic approach to review the existing literature about agent-based simulation studies in order to identify what elements are commonly used for describing simulation problems and what elements are used in specific application domains. In

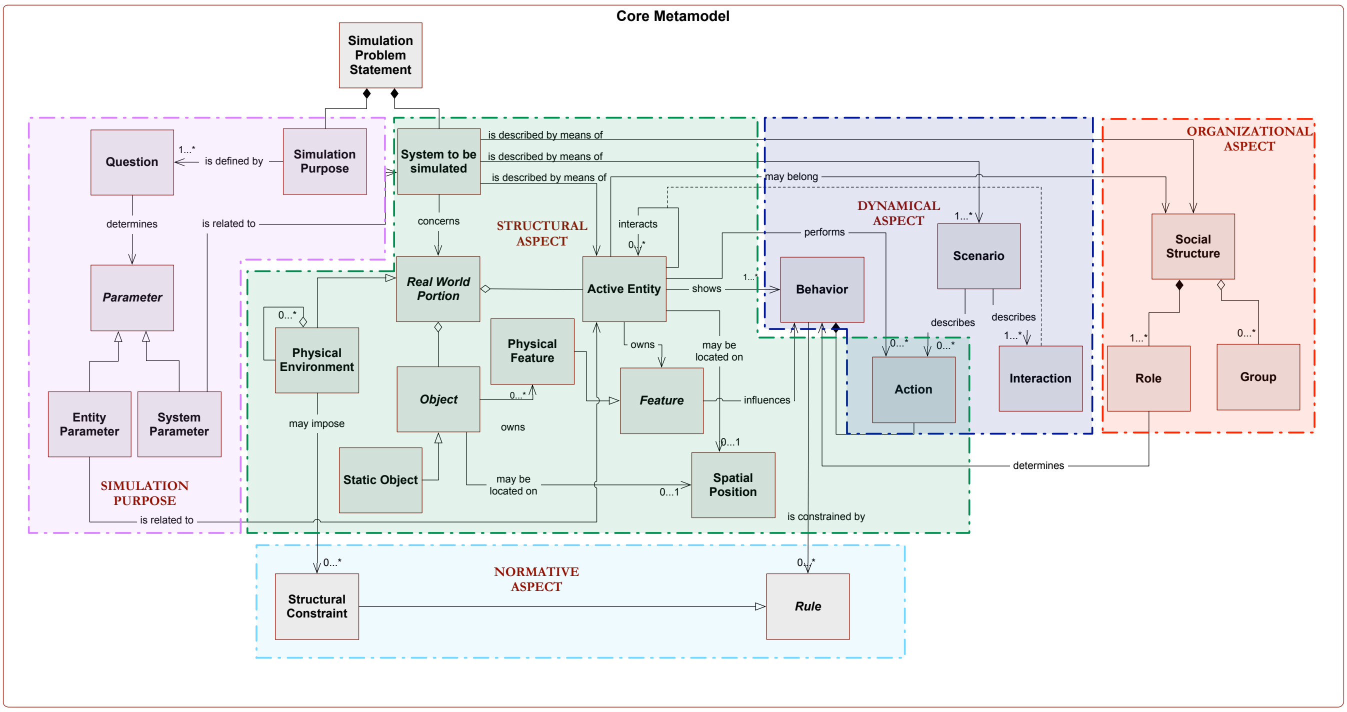


Fig. 3. The Core metamodel for agent based simulation problem.

so doing, we identified and analysed some problem domains where agent based simulations are frequently employed, thus determining a common ground formalized in a *Core Metamodel* that contains all the elements shared among the domains under analysis. Fig.2 illustrates the domains we covered in that review and the result is the metamodel shown in Fig.3. Such a metamodel shows that real-world systems commonly simulated by means of multi agent systems share some features we grouped in five key categories:

- the *Simulation Purpose* grouping elements that are related to the questions defining the goal of the simulation;
- the *Structural Aspect* namely the configuration of the real world influencing the simulation problem;
- the *Dynamical Aspect* represented by actions, interactions performed by someone/something for producing something or for achieving a particular end;
- the *Organizational Aspect* represented by the element that defines the social structure by means of group and roles;
- the *Normative Aspect* given by regulations that commonly constrains the dynamic or physical aspects of the system such as structural constraints or social norms.

Such metamodel underpins the *Simulation Problem Analysis* phase and each activity is devoted to manage some portion of it (see zooming of Fig.1). In particular, in this paper we detail the *Problem Formalization* activity that is devoted to instantiate the elements belonging to the structural aspect of the *Problem Ontology*. In the following we introduce the guidelines for performing this activity by means of a case study.

### A. Case Study

During the latest years we carried out several experiments in the field of goods management within logistic districts, studying how to store them in metropolitan distribution centres and the way to efficiently deliver them throughout the city. By working on this problem we had the possibility to conduct several experiments on agent based methodological approaches for simulation studies The focus of our studies was on one single node of the supply chain; the node was responsible for managing containers automatically unloaded by means of some AGVs (Automatic Guided Vehicle) provided with forklift for pallet handling; that is representative for a large number of real logistic warehouse-related problems.

Throughout the following subsections we use this case study in order to illustrate the proposed approach to problem ontology definition; an excerpt of the Problem Statement document is reported below.

*a) Problem Statement:* A logistic district is a large area composed of several warehouses where some freight forwarders may deliver their container. Inside a logistic district, several articulated lorries, coming from extra urban areas, arrive all the time and are arranged in warehouse bays. Each articulated lorry transports one 40-foot standard ISO (*International Organization for Standardization*) type container. Whenever a lorry reaches the warehouse, it docks into a bay for unloading its cargo. Commonly, a container holds several kinds of goods contained in boxes and grouped in pallets. These latter are *EUR pallets* (a standard ISO pallet measures 800x1200 mm). Each pallet must be unloaded from the container and carried into a well defined warehouse area dedicated to the sorting of goods, this operation is carried by the AGVs. In this area, each pallet is opened and its contents (packages of goods) placed on different sorters according to their destinations. These operations averagely last 5 minutes per pallet.

Commonly, a real logistic warehouse is basically composed of the following elements: gates, recharging areas (for AGVs), sorting areas, sorter places, buffer areas, paths and waypoints. The *Gate* corresponds to the warehouse bay and it is holds an unloading platform; lorries park here waiting for unloading their cargo (pallets). The *recharging area* is a specific area provided with sockets for recharging AGV's batteries when necessary, in particular the AGV moves towards the nearest free recharging area when its battery has depleted to a specific level. The *Sorting Area* is the place in which goods, contained in the pallets, are processed by the Sorter, an automatic device able to collect incoming goods and to forward toward a new destination. AGVs deliver goods to the Sorter using several input points (called *Sorter Places*). Goods are placed on hold when the Sorter is busy (all Sorter Places occupied by AGVs delivering some good) in the *Buffer Area*, this area is also used for temporarily parking AGVs that are waiting for a new mission. AGVs move following some *Paths* realized by means of optical guides connecting particular floor markers (*Waypoints*). AGVs have a value for their speed, their turning radius, loading capacity and type of guidance and lie in a particular area called parking area when unused.

A Waypoint has a unique ID and may be recognized by an AGV during its movement; a Waypoint is positioned outside each area of interest (Sorting Areas, entrance of unloading platforms and so on) or two meters before each cross between optical guides thus pointing out a stopping position for each AGV willing to pass a cross.

Paths are divided in path sections that are delimited by waypoints; there may be two kind of waypoints: waypoints positioned at the end of a path and the one positioned in the mid of a path, in the first case the waypoint is adjacent to one other waypoint only in the second case to two ones.

An automatic warehouse may be configured with several physical layouts (different disposition of optical lines, different number of entrances or exits, etc.) and equipped with some machinery. Each layout may impose limits to the use of different resources, the number of AGVs working simultaneously may be bounded by the set of available optical paths and their lengths; at the same time, specific equipment (i.e. the number and the performance of the AGVs, sorter capacity, etc...) and the choice of specific business strategies, such as path reservation strategy or AGVs scheduling, may constraint the performance of the warehouse thus greatly limiting its ability to quickly unloading arrived trucks.

Smaller vehicles (eco-friendly trucks) move packages out of the warehouse to their new destination (usually in town). The transport of pallets toward the sorting area is committed to AGVs with optical guidance; AGVs move, in order to load or unload pallets, by following the guidelines painted on the floor and engage a path between two adjacent waypoints, only one AGV at a time may engage such a path.

Given a warehouse made as described above the main aim of this simulation study is: how may we improve the throughput (the number of pallets unloaded per hour) of the warehouse?

### B. Guidelines for Problem Formalization

The aim of the Problem Formalization activity is to identify and instantiate all the metamodel elements composing the *Structural Aspect* of the domain under study hence all the real world concepts, or part of it, influencing the simulation problem. The result is a structural diagram, an UML class diagram, that we call Problem Ontology. This activity is deeply grounded on the analysis of the problem statement, hence on the analysis of portions of text from which the analyst has to identify the following elements: *active entity*, *object*, *feature*, *physical feature*, *spatial position* and *action*.

Before going on in the activity description, let us illustrate a verbs classification (shown in Table I) that proved useful in analysing the text for creating the Problem Ontology diagram.

| | | | | |
|---|---|---|---|---|
| **Dynamic Verbs (or Action Verbs)** | Dynamic verbs normally describe action that can be done or something that happens. They also describe action or events that may have a beginning and an end. | Transformational | They indicate events *without* a temporal extension and entail a *state change*. | wake up, leave, die… |
| | | Continuous | They indicate events *with* a temporal extension and they *have not* an implicit *goal* to reach. | sleep, work, help, improve… |
| | | Resulting | They indicate events *with* a temporal extension and they *finish* when a specific *goal* is reached. | lern, fall, build, complete… |
| **Stative Verbs** | Stative verbs refer to not changing or static condition. | Perception | They indicate perceptive processes | see, like, feel, listen, sound,.. |
| | | Cognition | They indicate cognitive process such as thinking, imaging, remembering and so on | think, suppose, remember,… |
| | | Relation | They indicate a relationships among objects or person | have, consist of… |

TABLE I. VERBS TAXONOMY

In grammar, verbs may be grouped in two main categories: dynamic verbs and stative verbs. A dynamic verb is a verb having a duration and referring to a continuous and progressive action of the subject, this is the opposite of stative verbs. Dynamic verbs are moreover distinguished in verbs showing a changing in the state of the element they refer to and/or verbs that make clear reference to a goal, an objective, to reach (for instance, work, help, sleep ...). In contrast, stative verbs refer to situations in which there is no obvious action (for instance, think, suppose, listen ...) or that relate different elements in a sentence (for instance, have, be, consist of, composed of ...).

We use this classification in order to identify in a text the elements of interest for our problem domain. As already said, the main focus when analysing a problem statement during a simulation study is on all those elements having an influence on the simulation goals. This verb classification allows us to retrieve from the text elements who performs an action when that action is intentionally performed by some active entity in order to cause a state change in the domain. It is also useful for discriminating among all the entities that actively produce a state change and all the ones that may be considered objects, resources or everything else is also used to reach a specific goal. Moreover, the focus on verbs allows us to identify the right relations among entities in the domain and above all the qualities and parameters belonging to entities that have to be taken into account when identifying all the goals of the simulation study and the metrics to be used for measuring their fulfillment. Although the latest part of the previous statement is not in the scope of this paper, it is worth underlining that the construction of a complete and well built problem ontology is the foundation, in our approach, for representing the domain as well as for identifying the system goals that purposefully influence the simulation goal; we already reported our work on the retrieval of goals from ontology in [24].

The heuristic we propose for identifying the metamodel elements from the problem statement prescribes to analyse the problem statement text in order to retrieve nouns, adjectives, verbs and relationships among them. This approach may resemble the well known Abbot's technique for identifying objects and their attributes/operations/relationships from use cases description (see [25]). Moreover, particular kinds relationship, such as aggregation or composition, may be identified also through the genitive or possessive case.

The following subsections report the guidelines for implementing the heuristic, they start from the identification of *active entities*, *objects* and *actions* that are related and identifiable through subjects, verbs and direct objects in sentences. *Active entities*, *objects* and *actions* are drawn by means of classes. [1]

It is worth noting the proposed guidelines help in identifying nouns, verbs, etc. that are candidate to become active entities, objects, actions and so on. The final decision whether they are really to be inserted in the Problem Ontology with that specific meaning is always left to the designer. The guidelines are not intended to define an algorithm for the automatic (i.e. non-supervised) production of a Problem Ontology diagram.

Fig. 8 reports a part of the Problem Ontology for the analysed problem statement. In the following subsections such ontology will be built step-by-step by applying the proposed guidelines.

*1) Preliminary Step-Nouns Identification:* A preliminary step is required: it prescribes to analyse the problem statement and to prepare a list of all nouns present in the text. Obviously, not all the nouns found during this work are relevant for the problem, we are constructing a model hence a view on a specific area of interest. The problem domain we are analysing as a case study falls in the area of the simulation of processes. It takes place in a warehouse, and it aims to improve its throughput. Hence, we may say, for instance, that *logistic district* is not a noun of interest because we are focusing on the warehouse that is a part of logistic district. The same is for *urban area* because, in this case, it clearly does not influence the real process in the warehouse and consequently our simulation study. It is also frequent that synonyms are used in the text. This is to be avoided when possible and existing synonyms are to be reported only once with the most immediately recognizable name. For instance, in the case study problem description text, it is possible to note that bay is used as a synonym of gate. The result of this activity in the reported case study produces the following list of nouns: lorry, pallet, AGV, path, gate, sorting area.

*2) Active Entity Identification:* The identification of active entities is performed in the following way:

```
For each noun in the previous list

-   analyze sentences in which it is the
    subject of a verb

-   if there exists at least one sentence
    in which it is the subject of a
    dynamic verb (or it is in some kind of
    relationship with a dynamic verb)
      -   analyze the verb
      -   if it is a resulting or a
          transformational verb (Table I)
          the considered noun is an active
          entity
```

Some active entities may be in composition/aggregation, specialization or association relationship with other active entities, this kind of relationships are identified by means of stative verbs, relation kind.

In the presented case study there are not evidences of stative verbs such as cognition and perception verbs, however, generally speaking, they both may be verbs whose subject is an active entity.

The result of this activity in the reported case study brings to the identification of sentences like:
"*AGVs move*, in order to *load* or *unload* pallets, by following guidelines (hence paths) painted on the floor"
"*Lorries park* here (ref. gates)"
"*AGV moves* towards the nearest recharging area"

In these sentences two active entities may be identified: *AGV* and *Lorry* (the use of the singular form of the nouns is advisable). A first portion of Problem Ontology Diagram may now be built by reporting in it the two identified active entities (see Fig. 4).

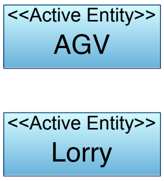


Fig. 4. Problem Ontology-Step 2 (active entities)

[1] It is worth noting that the guidelines for performing Problem Formalization may be associated to every kind of notation, we choose UML class diagram because its extended usage makes it easier to understand the composition of the Problem Ontology diagram.

*3) Actions Identification:* The identification of actions is performed in the following way:

> For each previously identified active entity, associate the corresponding action in the sentence and draw them in the diagram.

This step may be performed in a iterative fashion with the previous one (active entities identification).

The result of this activity in the reported case study brings to the identification of the following actions: move, load, unload, follow, park.
Such elements are reported in the Problem Ontology diagram thus obtaining its new release as reported in Fig. 5.

The notation we adopt for this diagram prescribes the direction of the relationships went from the active entity (the actor performing an action) towards the performed action.

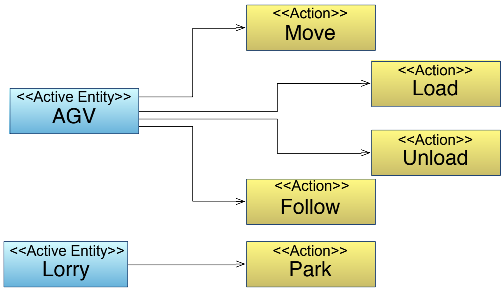


Fig. 5. Problem Ontology-Step 3 (actions)

*4) Objects Identification:* The identification of objects is performed in the following way:

> All nouns that are direct objects in sentences where an active entity is a subject are *objects*. All the other nouns that are not active entities are candidate objects. Candidate objects that are subject of stative-relation or dynamic-continuos verbs are *objects*.

In order to definitively identify all the *objects*, the designer has to examine the structure of sentences and to analyse if a candidate object is related to another one or if it is related to adjectives, anyway she has to look through which kind of verbs they are related to. This is complemented by the following step. Moreover, as for active entities, objects may be related by association, composition/aggregation and generalization relationships.

The result of this activity in the reported case study brings to the identification of sentences like:
"AGVs move, in order to load or unload *pallets*, by following guidelines (hence *paths*) painted on the floor"
"Lorries park here (ref. *gates*)"
"AGV moves towards the nearest *recharging area*"
"Paths *are divided* in path sections that *are delimited* by waypoints; there *may be* two kind of waypoints..."

In these sentences the following objects may be identified: *Recharging Area*, *Pallet*, *Path*, *PathSection*, *Waypoint*, *MidWayPoint*, *FinalWayPoint*, *Gate*. Some of them are related by composition/generalization relationships as described in the new release of the Problem Ontology diagram reported in Fig. 6.

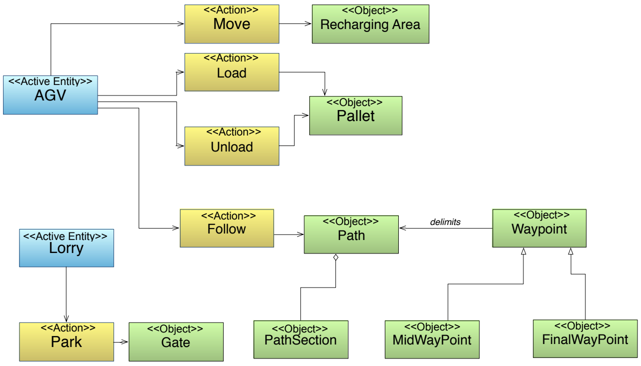


Fig. 6. Problem Ontology-Step 4 (objects)

*5) Feature Identification:* The identification of features is performed in the following way:

> For each candidate object analyse if it is related to other candidate objects through relation or continuous verbs. We may find two cases:
>
> - noun + relation or continuous verb + adjectives, the noun is an *object* and the relationship with the adjective is described with a predicate
> - noun + relation or continuous verb + noun, the first noun is an *object*, the second one may be represented through another *object*, related with UML relationships, or through attributes.

The choice between representing relations as attributes rather than other objects depends on what the analyst wants to underline and on the specific problem domain. In the presented case study, for instance, speed is represented as an attribute of AGV and not as an object because we found not necessary to represent speed as an object while simulating processes in warehouses. The contrary happens for Sockets that in the real word own a specified position and physical attributes that, in the designer's understanding, may influence the simulation goal (warehouse throughput optimization).

The result of this activity in the reported case study brings to the analysis of sentences like:
"in the first case the waypoint is adjacent to one other waypoint only in the second case to two ones."
"AGVs have a value for their speed, their turning radius, loading capacity and type of guidance."
"The recharging area is provided with sockets for recharging AGV's batteries when necessary"

The first sentence brings to the identification of the IsAdjacent predicate between two waypoints. The second one lists some attributes of AGVs (speed, turning radius,...), the last one identifies two new objects (Socket and Battery) and the relationship between Socket and Battery.

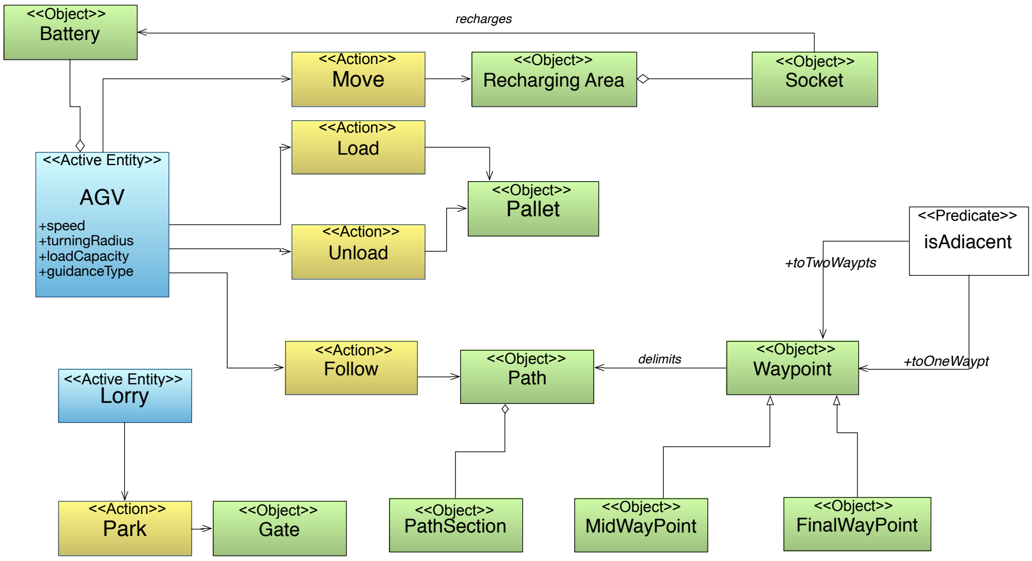


Fig. 7. Problem Ontology-Step 5 (Features)

Such elements are reported in the Problem Ontology diagram thus obtaining its new release as reported in Fig. 7.

*6) Spatial Position Identification:* The identification of features is performed in the following way:

Guidelines in this case are the same of *Feature*, the designer has to look at continuous or relation verbs denoting positioning of objects in the space and choose if they may be treated as objects, attribute or predicates.

The result of this activity in the reported case study brings to the analysis of sentences like:
“a Waypoint is positioned outside each area of interest”
“AGVs lie in a particular area called parking area when unused”

The first sentence allows to identify two relations, the first between the concepts Waypoint and InterestPoint whereas the second between the concepts AGV and ParkingArea; see Fig. 8 where, as already anticipated, the complete resulting Problem Ontology is reported.

## V. Conclusions

Simulation studies conducted by using agent based modelling and simulation approaches had reported in the latest period good results, however they still lack a rigorous methodological approach for going, step by step, from simulation problem analysis to the implementation of the agent based simulation system. We may say that in the case of simulation studies the development of agent based systems is a part that resides within the overall simulation study and its requirements are greatly affected from the simulation requirements themselves. In this context an important role is covered by the simulation problem analysis phase that aims at identifying and describing the domain under study along with the simulation goals.

We claim that simulation goals descending from the problem domain, are greatly related to the goals of the agent based systems, although they cover a different scope with respect to agent goals. Indeed, simulation goals are related to the interest that the simulation team sees in the results of the simulation study. Whilst agent goals refer to the states of the world an agent wants to achieve in the system they live. This system is the means the simulation approach uses to achieve the intended simulation goals. Thus, simulation goals constraint the scope of the problem to be addressed by the simulation study and they guide the definition of the problem elements that are useful to model.

In this context it is of high importance to have some guidelines for extracting the goal of the simulation and the system goals from the description of the problem domain. In some previous work [24] we already experienced how to extract goals from a formalized description of the problem domain and we want to somehow apply the same approach for the future development of our work on agent based simulation study.

In this paper we explore the step before the identification of goals and provide a set of guidelines for modelling the structural aspect, hence the configuration of the real word influencing the simulation goal, of the problem domain. We propose an heuristic based on the analysis of verbs and nouns in the problem statement.

This work is the logic progression of what reported in [11] where we identified which elements have to be present in the problem statement of a simulation study; the result of that work was a metamodel whose elements (the part related to the structural aspect) are instantiated in the simulation problem analysis phase by means of the guidelines we propose in the proposed contribution.

For the future, we are going to complete all the activities of the simulation problem analysis phase as well as the whole methodological approach for multi agent simulation study.

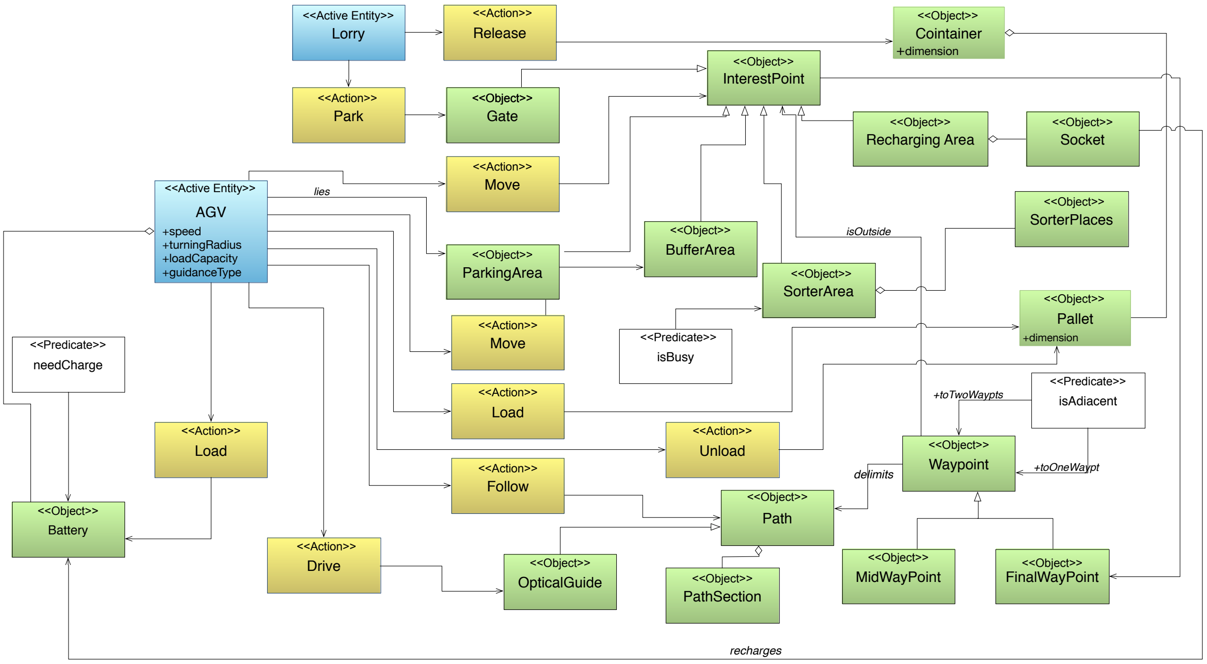


Fig. 8. The Problem Ontology diagram resulting from the application of the proposed heuristic for Logistic Warehouse case study.

## References

[1] M. Wooldridge and N. Jennings, "Intelligent agents: Theory and practice," *The Knowledge Engineering Review*, vol. 10, no. 2, pp. 115–152, 1995.

[2] M. J. Wooldridge, *Introduction to Multiagent Systems*. John Wiley & Sons, Inc. New York, NY, USA, 2001.

[3] J. Georgé, M. Gleizes, P. Glize, and C. Régis, "Real-time simulation for flood forecast: an adaptive multi-agent system staff," in *Proceedings of the AISB*, vol. 3, 2003, pp. 109–114.

[4] M. J. North and C. M. Macal, *Managing business complexity: discovering strategic solutions with agent-based modeling and simulation*. Oxford University Press, 2007.

[5] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 3, pp. 7280–7287, 2002.

[6] N. Gilbert, *Agent-based models*. Sage, 2008, no. 153.

[7] D. Helbing and S. Balietti, "How to do agent-based simulations in the future: From modeling social mechanisms to emergent phenomena and interactive systems design," Chapter "Agent-Based Modeling" of the book "Social Self-Organization" by Dirk Helbing (Springer, Berlin, 2012), 2013.

[8] F. Klügl and P. Davidsson, "Amason: Abstract meta-model for agent-based simulation," in *Multiagent System Technologies*. Springer, 2013, pp. 101–114.

[9] A. Helleboogh, G. Vizzari, A. Uhrmacher, and F. Michel, "Modeling dynamic environments in multi-agent simulation," *Autonomous Agents and Multi-Agent Systems*, vol. 14, no. 1, pp. 87–116, 2007.

[10] R. Siegfried, A. Lehmann, R. El Abdouni Khayari, and T. Kiesling, "A reference model for agent-based modeling and simulation," in *Proceedings of the 2009 Spring Simulation Multiconference*. Society for Computer Simulation International, 2009, p. 23.

[11] P. Ribino, V. Seidita, C. Lodato, S. Lopes, and M. Cossentino, "Common and domain-specific meta-model elements for problem description in simulation problems," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*. IEEE, 2014, pp. 1467–1476.

[12] J. S. Carson *et al.*, "Introduction to modeling and simulation," in *Simulation Conference, 2005 Proceedings of the Winter*. IEEE, 2005, pp. 8–pp.

[13] A. M. Law, "How to build valid and credible simulation models," in *Simulation Conference (WSC), Proceedings of the 2009 Winter*. IEEE, 2009, pp. 24–33.

[14] F. Klügl, "Engineering agent-based simulation models?" in *Agent-Oriented Software Engineering XIII*. Springer, 2013, pp. 179–196.

[15] O. Balci, "Guidelines for successful simluation studies (tutorial session)," in *Proceedings of the 22nd conference on Winter simulation*. IEEE Press, 1990, pp. 25–32.

[16] ——, "Validation, verification, and testing techniques throughout the life cycle of a simulation study," *Annals of operations research*, vol. 53, no. 1, pp. 121–173, 1994.

[17] A. Drogoul, D. Vanbergue, and T. Meurisse, "Multi-agent based simulation: Where are the agents?" in *Multi-agent-based simulation II*. Springer, 2003, pp. 1–15.

[18] V. Grimm, U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, G. Huse *et al.*, "A standard protocol for describing individual-based and agent-based models," *Ecological modelling*, vol. 198, no. 1, pp. 115–126, 2006.

[19] A. Garro and W. Russo, "easyabms: A domain-expert oriented methodology for agent-based modeling and simula-

tion," *Simulation Modelling Practice and Theory*, vol. 18, no. 10, pp. 1453–1467, 2010.

[20] C. Cioffi-Revilla, "Computational social science," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 3, pp. 259–271, 2010.

[21] A. Ghorbani, P. Bots, V. Dignum, and G. Dijkema, "Maia: a framework for developing agent-based social simulations," *Journal of Artificial Societies and Social Simulation*, vol. 16, no. 2, p. 9, 2013.

[22] S. Montagna, A. Ricci, and A. Omicini, "A&a for modelling and engineering simulations in systems biology," *International Journal of Agent-Oriented Software Engineering*, vol. 2, no. 2, pp. 222–245, 2008.

[23] A. Molesini, M. Casadei, A. Omicini, and M. Viroli, "Simulation in agent-oriented software engineering: The soda case study," *Science of Computer Programming*, vol. 78, no. 6, pp. 705–714, 2013.

[24] P. Ribino, M. Cossentino, C. Lodato, S. Lopes, L. Sabatucci, and V. Seidita, "Ontology and goal model in designing bdi multi-agent systems." *WOA@ AI* IA*, vol. 1099, pp. 66–72, 2013.

[25] B. Bruegge and A. H. Dutoit, *Object-Oriented Software Engineering Using UML, Patterns, and Java*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. ISBN 0136061257, 9780136061250

DOI: 10.15439/2015F248
ACSIS, Vol. 5

# Developing an Integrative Modelling Language for Enhancing Road Traffic Simulations

Alberto Fernández-Isabel
Rubén Fuentes-Fernández
{afernandezisabel@estumail, ruben@fdi}.ucm.es
Universidad Complutense de Madrid
Madrid, Spain

***Abstract*—Road traffic is a pervasive aspect in modern societies that affects millions people. The study of its multiple aspects is a very demanding task. Due to its complexity, traffic simulations become a key tool. Their development demands multidisciplinary teams, where communication problems are frequent. Model-driven engineering alleviates this situation providing graphical instruments for designing Modelling Languages (MLs) and semi-automatic transformations. This work presents a model-driven infrastructure composed by an integrative ML, a model editor, and a code generator. The ML is based on related literature and facilitates modelling different theories and simulations based on them. It considers the roles of individuals involved in road traffic, and partially adopts agent-based methodologies to model their decision-making. A case study shows how to produce a simulation specification adapting an existing traffic theory to the ML, and adjust this specification to a simulation platform for testing. It provides the basis for comparison with related work.**

## I. Introduction

ROAD traffic is a complex phenomenon. Its study requires considering a large amount of variables, and it affects a variety of aspects, such as pollution, economical factors, leisure organisation, and health issues. The individuals involved play multiple roles in a broad range of scenarios, being able to establish complex relationships among them. These features make difficult studies based on experiments in real settings, which leads researchers to limit the variables considered and focus only on very specific aspects. In order to alleviate these restrictions, traffic simulations appear as a possible solution. Nevertheless, simulations present their own weak points [1]. Some of the most relevant are related to the difficulties to align (and check this alignment) the theories, goals, code, and results of the simulation, particularly because of the different backgrounds of people involved and the use of implicit information.

Approaches based on Model-Driven Engineering (MDE) [2] have been proposed to deal with these issues. They are organised around *models*, which are compliant with MLs, and generate other artefacts via semi-automatic *transformations*. MDE processes are usually incremental and iterative, allowing introducing improvements and modifications at any part of their workflows. MDE requires an initial effort for developing the elements of its infrastructure. This effort is higher than just implementing a simulation, but it compensates it with reusability (i.e. the resources can be used as a basis for other projects) and the explicit description of all the information (with MLs, models, and transformations).

Our approach provides a complete and integrative MDE infrastructure for road traffic simulations, focused on modelling individual behaviours. It introduces a Traffic Modelling Language (TML) and two development tools. A graphical editor supports describing the specifications, and a code generator helps to produce the semi-automatic transformations to generate the source code for a target simulation platform from those specifications.

The TML pursues being able to integrate (i.e. support the modelling and combination of) different theories related to road traffic. It considers the multiple viewpoints of the involved individuals (i.e. drivers, pedestrians, and passengers) and the roles in development teams (e.g. traffic expert and programmer). In this sense, it is a Domain-Specific ML (DSML) for this kind of problem. Faced to the traditional dichotomy in MLs between general and specific ones [3], our work chooses limiting the applicability of the approach to traffic simulations in order to provide better support in terms of guidance and tools to experts. The language is also intended to be platform-independent, so the details of the target platform can be considered just in late design tasks.

Following common practices in MDE, a metamodel describes the TML. It is organised using inheritance and composition hierarchies. The inheritances provide specialisation of concepts, while compositions are based on relationships of purpose, functional groups, physical links, or similarity. The metamodel is divided into three clusters: a *Mental cluster* where the different features of individuals are considered, an *Environmental cluster* to specify environment information, and an *Interactive cluster* to represent the interactions among individuals and the environment and the decision-making.

The *Mental cluster* considers the psychological features of individuals [4]. It plays a role similar to the mental state of the agent paradigm [5]. It adopts the BDI model [6], incorporating some of its knowledge concepts to model the current information that an individual or group possesses.

The *Environmental cluster* is based on the Driver-Vehicle-Environment (DVE) [7] approach. It considers that individuals can interact among them or with the environment, either directly or using their means of transport. These dynamic interactions influence the individual behaviours. This assumption

978-83-60810-66-8/$25.00©2015, IEEE

fits with Agent-Based Modelling (ABM) [8], where agents are intentional entities that can establish communications among them for different purposes (e.g. collaborate or interact).

The *Interactive cluster* models the decision-making of the individuals. It adapts this aspect from methodologies based on Agent-Oriented Software Engineering (AOSE) (e.g. INGENIAS [9] or Tropos [10]). *Goals* represent the people's objectives, and *tasks* are the instructions to execute in order those satisfy goals. These elements are considered in a *perceive-reason-act* cycle [11].

Regarding the development tools, the graphical editor allows designing the model instances compliant with the TML. The code generator takes as input these instances, and provides a set of functionalities to generate source code and adapt it to the target traffic simulation platform.

A case study shows the suitability of the MDE infrastructure to develop traffic simulations. It specifies the work in [12] using the ML. The graphical editor tool supports and guides this process, which produces a model specification and default source code templates for the primitives of the ML. These artefacts are the input of the code generator tool. It supports the completion of these templates with graphical wizards that assist users. This code is finally adapted to a specific traffic simulation platform, MATSim [13]. For this purpose, the code generator allows adding specific classes and code snippets to modify available code (e.g. using libraries or algorithms).

The rest of the paper is organised as follows. Section II presents the basic concepts of MDE and the related tools. The TML is introduced in Section III through its metamodel and clusters. Section IV presents the two development tools based on the TML, the model editor and the code generator. The case study in Section V illustrates the application of the approach. Then, Section VI compares this with related work. Finally, Section VII discusses some conclusions and future work.

## II. Model-Driven Engineering

MDE [2] is a development methodology that is composed around *models*, in contrast to traditional approaches that are based on source code. The development process is focused on the production of iterative and incremental specifications of models going to abstract to accurate, where developers refine and add new elements to them at each step. During this process, *transformations* are introduced in order to automate repetitive modifications in models. For instance, generating patterns or concrete specialisations to target platform in order to produce models. Other related elements (e.g. source code or documentation) are compliant to these considerations, as they can be obtained from models using manual settings and transformations.

This development approach is based on modelling languages. In the case of graph-oriented languages, which are the most popular ones [14], the main instrument to achieve these definitions is the metamodel. Metamodels are commonly selected to describe their abstract syntax, but also they can be used to define their specific syntax or semantics [15]. Also, these metamodels are defined using meta-modelling languages. The Meta-Object Facility (MOF) [16] provided by the Object Management Group (OMG) is the standard in the domain. Nevertheless it presents some limitations. The absence of extensive tool support promotes that users frequently choose alternative languages or develop their own related tools. The Ecore meta-modelling language [17] is considered as an alternative as it is supported by multiple Eclipse modelling tools. These tools are organised around the Eclipse Modelling Framework (EMF) [17] and the Graphical Editing Framework (GEF) [18]. Also, Ecore adopts the Object Constraint Language (OCL) [19] to define model constraints and it is almost compliant to Essential MOF (EMOF). EMOF is a part of MOF focused on object oriented concepts and able to specify reflective operations. These features encourage this approach to select Ecore as its meta-modelling language in order to develop the TML.

Fig 1 shows the principal primitives of Ecore. An instance of *EClass* plays the role of its similar entities at the model level (i.e. classes). It clusters *EAttribute* and *EReference* elements. *EAttribute* instances provide features coming from *EDataTypes* (i.e. primitive types) to *EClass* instances. These primitive types include the most common (e.g. integer, char or string). An *EReference* instance symbolises a binary relationship in only one direction among two *EClass* instances. It allows creating containment and non-containment relationships. The *EReferenceType* of a specific *EReference* instance is indicated by its target *EClass*. Multiple *EClass* instances can be considered by *ESuperType* relationship in order to express inheritance among them. *EPackage* instances contemplate the possibility of grouping the structures of the metamodel.

Regarding the transformations, they are the other core instrument of MDE. They present different types of inputs and outputs, being able to be classified in [20]: Model-to-Text (M2T), Text-to-Model (T2M) and Model-to-Model (M2M) [21]. These transformations can be developed using general-purpose programming languages or transformation languages. In the first case, a module uses programming structures to manage its inputs and outputs. In the second case, the module is developed using a specific language for transformations and presents an engine that executes it in order to accomplish the process.

In this work, a module defined by a general-purpose programming language (i.e. Java) is adopted (see Section IV), as it can use techniques from reflection-based programming [20] and integrates several wizards that assist developers.

## III. Traffic metamodel

Road traffic is a pervasive phenomenon that involves elements and situations. In order to study it, there are different theories that consider its aspects from multiple backgrounds and purposes. The same situation occurs with traffic simulations, where infrastructures differ on their modelling approach and goals.

In order to facilitate the integration and modification of elements in the TML and its study, this approach uses a metamodel [17] defined with Ecore.

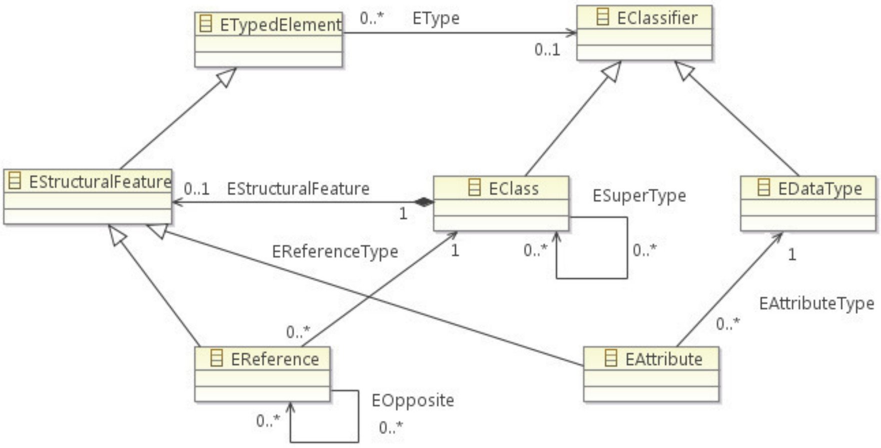


Fig. 1. Extract of the Ecore model selected from [17]

Regarding the concepts of the TML,they are mainly based on ABM [8] and structured into three clusters. The *Mental* and *Environmental* clusters gather the different concepts obtained from traffic literature. The *Mental cluster* represents the inner state of the participants in traffic [4]. The *Environmental cluster* includes the DVE approach [7]. These clusters have similar structures (see later in this section). The *Interactive cluster* is focused on representing the goals and actions of people involved in traffic. It is based on the guidelines of methodologies coming from Agent-Oriented Software Engineering (AOSE), integrating a *perceive-reason-act* cycle [11].

The core concept of the metamodel is the *Person* meta-class. It represents the types of people involved in traffic. According to their means of transport, they can be drivers, passengers, or pedestrians. These *Person* instances can interact with an *Environment* instance. This interaction is direct (in the case of pedestrians), or indirect when a *Vehicle* instance is used for it (for drivers and passengers). People' features are modelled with *Profile* instances, and the information they possess with *Knowledge* instances. Their acts are motivated by *Goal* instances, and the potential ways to achieve them are represented by *Task* instances. *Evaluator* instances determine how people have actually to act according to the circumstances, and *Actuator* instances execute the planned tasks.

The previous elements are arranged in inheritance hierarchies, adding the needed specialisation and structure to the metamodel. All concepts inherit from the *GeneralElement* meta-class (see Figs. 2 and 3). This meta-class provides the *EInherits* reference in order to represent inheritance among elements of the same type in model instances. The *GeneralRelationship* meta-class (see also Figs. 2 and 3) supports the introduction of relationships (e.g. affects or influences) among other elements. The *RInherits* reference allows its specialisation. Both types of references are constrained by expressions written with OCL [19]. For instance, constraints only allow inheritance among instances of the same type of meta-classes (e.g. a *Knowledge* instance only extends another *Knowledge* instance using a *EInherits* reference).

The internal structure of the *Mental* and *Environmental* clusters allows composition hierarchies using the *XComponent* (e.g. *KComponent* or *VComponent*) meta-classes. These meta-classes can be decomposed into others of their same types. All these compositions are constrained by OCL expressions. For instance, a *Profile* instance can be decomposed only into *PComponent* instances, while these *PComponent* instances can be only decomposed into others of the same type.

The meta-classes of the metamodel include attributes and predefined methods. Attributes can be specific for certain meta-classes or common (with similar name and meaning) to several meta-classes. An example of the first case is the *AvailableArea* attribute in the *Environment* meta-class; the *XValues* attributes (e.g. *PValues* or *EValues*) for storing the impact that an element has in the rest of the elements of a model instance are examples of the second group. Methods are placeholders for specifications that describe behaviour or attributes derived from others. For instance, code snippets can be attached to these methods in the model specification for code generation.

Next sub-sections discuss these aspects in detail. Subsection III-A describes the mental state and features of participants in traffic. Subsection III-B focuses on concepts to describe the traffic setting according to the DVE model [7], i.e. vehicles and the environment. Subsection III-C introduces the concepts to represent interactions among the previous elements and decision-making.

### *A. Mental cluster*

The *Mental cluster* (see Fig. 2) represents the different concepts that can appear in the road traffic domain influencing the behaviour of individuals [4]. These concepts are classified as features of people or their current state.

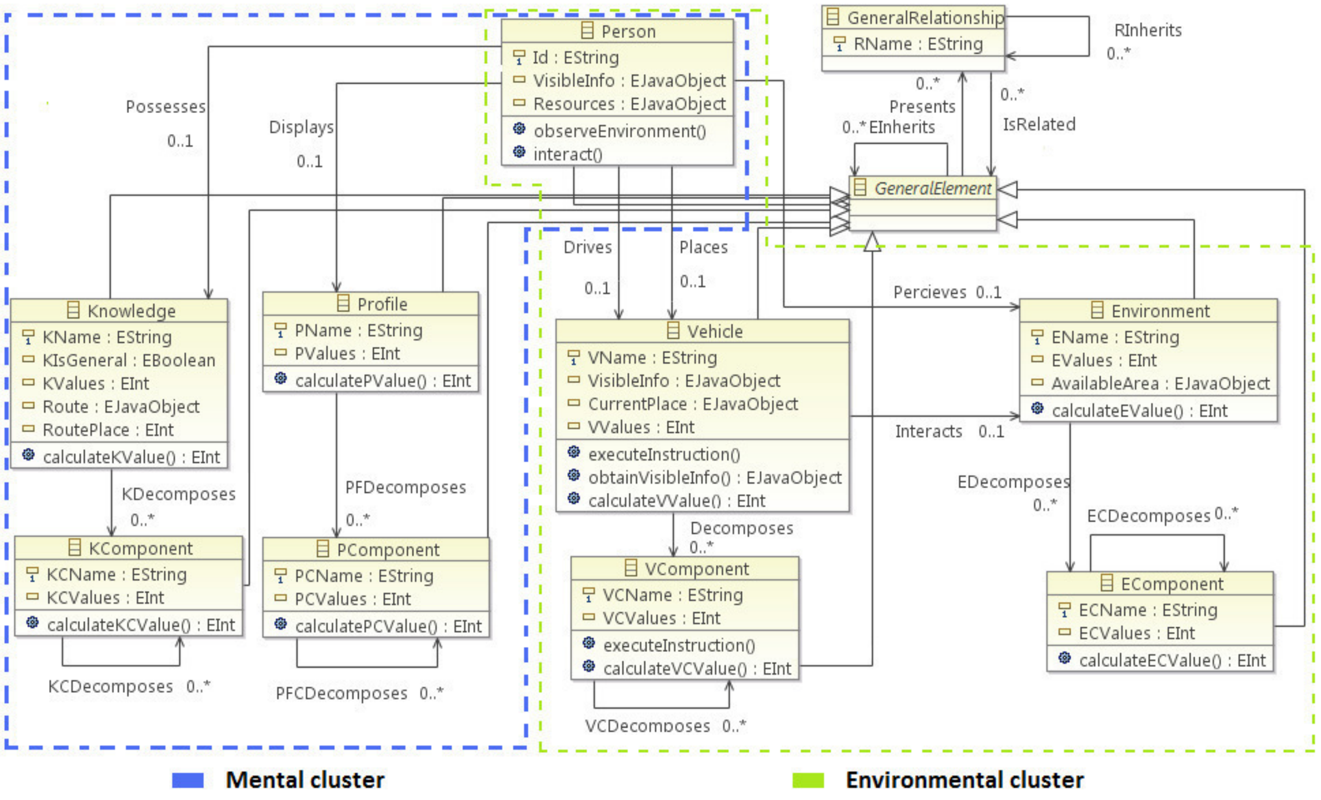


Fig. 2. Excerpt of the *Mental* and *Environmental* clusters of the metamodel.

The cluster includes three main meta-classes: *Person*, *Profile*, and *Knowledge*. *Profile* represents the different features of people in traffic. *Knowledge* considers the current mental state (but the goals) that a *Person* uses when dealing with traffic. It can be factual (e.g. traffic signs), procedural (e.g. how to overtake a vehicle), and normative (e.g. drivers should respect safe distances with other vehicles) knowledge. *Profile* instances describe people features (e.g. age or fatigue).

Both knowledge and features of people can specify information that does not change in simulation time (e.g. gender or meaning of signs), or does it (e.g. stress or mood). Proper *calculate* methods and their associated attributes must be specified to describe how to calculate them later.

The instances of the *Knowledge* meta-class and their composition meta-classes can represent information belonging to individuals (e.g. the current route), or global information available for every participant in the simulation (e.g. the speed limit in a specific type of road). The *KIsGeneral* attribute differentiates both uses.

This cluster is closely related to the agent paradigm [5]. For instance, the *Knowledge* meta-class can consider the *Beliefs* of people involved in road traffic, which are contemplated in the BDI model [6].

### B. *Environmental cluster*

The *Environmental cluster* (see Fig. 2) adapts the concepts of the DVE model [7], as this is focused only on the driver role and the TML considers others. Thus, here it is considered that an individual can get information from the environment (any participating person) and the vehicle (only drivers and passengers). These elements can be extended to facilitate the potential accommodation of other theories.

The cluster considers how individuals relate to their means of transport and environment. It comprehends three main meta-classes: *Person*, *Environment*, and *Vehicle*. *Environment* represents the place where people (i.e. *Person* instances) interact, including the physical conditions that can occur (e.g. weather and road conditions). A model specification has a unique *Environment* instance shared by all the individuals. The *Vehicle* represents the means of transport, considering the different roles of people in road traffic (i.e. driver, passenger, and pedestrian). Drivers and their passengers relate to the *Environment* through their vehicles, but only drivers can use them to act on it. In the case of pedestrians, they have a direct relationship with the environment.

The mutual influences among *Person*, *Environment*, and *Vehicle* instances because of their relationships are partially represented in the metamodel with some attributes. The *Environment* meta-class has an *AvailableArea* attribute. It indicates the part of the environment that can be perceived. The *Person* and *Vehicle* meta-classes include the *VisibleInfo* attribute to specify which information from the *Environment* instance can be perceived in or through their instances.

### C. Interactive cluster

The *Interactive cluster* (see Fig. 3) describes how *Person* instances act on the traffic situations considered by the *Mental* and *Environmental* clusters (see Sections III-A and III-B). Its components are organised into two groups. The first one describes the objectives of people and their capabilities to achieve them. The second one represents the elements that carry out the acting cycle.

The first group includes the *Goal* and *Task* elements. These two concepts come from Multi-Agent Systems (MAS) [22], and agent-based methodologies. These methodologies include a specific acting architecture where agents play multiple roles and try to meet the requirements of their different goals. These goals are enabled according to the agents' mental states and are directly related to task elements that can satisfy them. These goals can be decomposed into others, generating *OR* or *AND* compositions. Tasks can be decomposed in a similar way in order to describe complex jobs.

In our work, the *Mental cluster* represents the mental state of agents, and the *Environmental cluster* provides information from the environment and the vehicle (only in the case of drivers and passengers). The *Goal* meta-class represents a state of some traffic elements a person aspires to keep or reach, and the *Task* meta-class models person's capabilities. Both meta-classes hold specific attributes to characterise them. *Goals* have *Satisfaction* attributes that represent their satisfaction conditions. *Tasks* include *Instructions* attributes to specify the atomic actions that implement them.

These meta-classes can be decomposed into others of their type (constrained with OCL expressions), following the structure already seen for sub-components in the other two clusters. However, semantics are different. Here, they are related to satisfaction instead of determining the features of a component. The *Goal* and *Task* meta-classes present the *GType* and *TType* attributes in order to specify the type of compositions (e.g. *OR* or *AND*). The *GType* attribute represents the type of goal satisfaction compositions, while the *TTYpe* attribute indicates if the current task is accomplished by completing one or all its sub-tasks. These semantics are flexible, as both attributes could be modified to support different structures and classifications.

The second group represents those elements of a person that are in charge of evaluating the actual known state and executing actions. It follows a classical *perceive-reason-act* cycle [11] with evaluators and actuators (based on [23]). The information perceived from the environment is stored in the elements of the *Environmental cluster* (including the *Person* meta-class), the reasoning is carried out by *Evaluator* instances, and the acting is achieved with *Actuator* instances.

*Evaluator* instances can be decomposed into others using the *EVDecomposes* reference, being able to distribute the liabilities among them. In the case of *Actuator* instances, they cannot be decomposed into others. They can use inheritance through *EInherits* references, but each *Person* instance (i.e. a person type modelled) can only present one *Actuator* instance related to it (see the *Utilizes* reference in Fig 3).

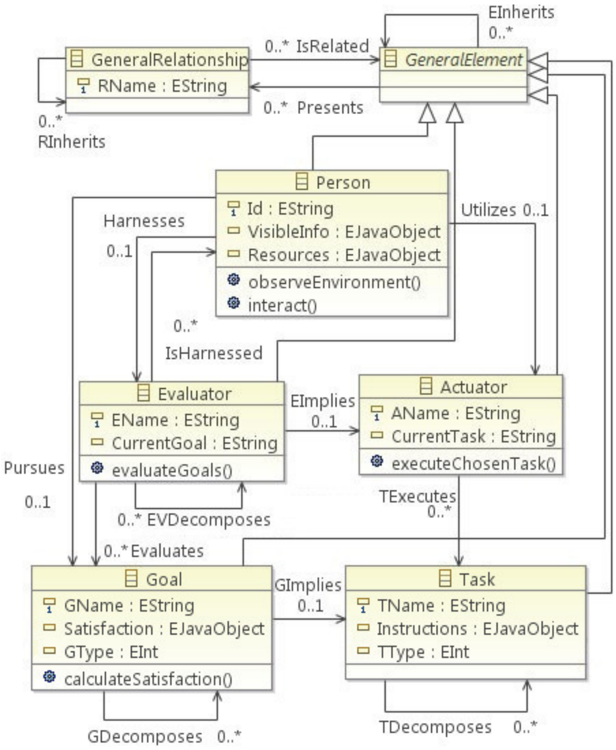


Fig. 3. Excerpt of the *Interactive cluster* of the metamodel.

*Evaluator* instances assess the information obtained from the *Environment* and *Vehicle* instances, the elements they are composed, and the available relationship instances linked to their *Person* instance. From that, they update the internal state of the *Person* instance. All this information determines the current state of goals. Once a candidate goal is selected, an *Actuator* instance picks its associated tasks. It executes these *Task* instances through its *Instructions* or subtasks.

## IV. Development tools

The MDE approach presented in this paper is supported by two main tools: a graphical editor where the model specifications are developed, and a code generator where multiple operations to produce source code from models are achieved through semi-automatic transformations. Their implementation is based on the Eclipse Modelling Framework (EMF) [17] and the Graphical Editing Framework (GEF) [18].

The graphical editor is an Eclipse plug-in that guides users in the development of model specifications. It generates models compliant with the TML through a visual interface. This interface provides a canvas and a palette for displaying the model and the concepts of the metamodel. The model generated can be validated to ensure its compliance.

The code generator tool takes as input the model specifications produced by the graphical editor. In a first step, it associates to the classes in the model the source code EMF generates automatically for their meta-classes. Also, it provides options to integrate other external files, e.g. specific libraries coming from the target platform or even the entire simulation with its associated dependencies. This can be used later to modify the preliminary code.

Over that input, the tool presents a graphical interface for displaying the information captured in the model specifications, allowing an intuitive navigation of them. The information about a selected element instance includes the methods associated (original and newly created), the decomposed elements it presents, the *GeneralRelationship* instances where it acts as the origin, and the elements from which it inherits (see Fig.4).

The code generator implements operations related to source code transformation and model adaptation, and operations related to specialisation to target simulation platform. Most of them are partially automated through wizards in order to provide guidance to users. A text editor, an internal graphical editor, and a compiler are integrated in the infrastructure in order to support these features.

Regarding the code transformation, main functionalities are: source code injection for platform adaptation, design and storage of self-contained *Interactive clusters* (see Section III-C), and cluster integration.

The injection of source code offers two alternatives. They are based on techniques from reflection-based programming [20] in order to modify and compile dynamically the default EMF implementation. The first one redefines only the body of the methods of the classes adding different instructions, using suitable code snippets for the target simulation platform. The second one is more complex, being able to complete the entire class or extending it from another one of the same type previously redefined. This allows adding new attributes and methods. As these operations require some programming skills, the graphical interface and the integrated modules (i.e. text editor and compiler) assist to examine the metamodel and model elements, and their code. This facilitates these tasks and produces a more intuitive development environment. There is also on-line help and examples to guide users in this point.

The design and storage of self-contained *Interactive clusters* uses the integrated graphical editor. It provides (in a similar way to the graphical editor plug-in explained above) a canvas and a palette to create the multiple elements (i.e. *Goal*, *Task*, *Evaluator*, *Actuator*, and *GeneralRelationship* instances) and a validation tool. After that, the generated cluster can be loaded into the tool in order to perform other tasks, such as code injection or storage of the development stage. In the last case, a wizard creates a compressed file including the current stage and the model designed, which allows continuing with the graphical design in the future.

The cluster integration functionality is linked to the previous one. It merges a model specification only with elements from the *Mental* and *Environmental* clusters previously loaded in the tool, and a stored self-contained *Interactive cluster*. A graphical wizard facilitates the process showing to users the available elements in each cluster in order to link them through references from the TML. Also, *GeneralRelationship* instances can be added or completed (i.e. relationships with origin in elements of *Mental* and *Environmental* clusters and destination in elements of *Interactive cluster* or vice versa). This functionality is particularly useful since models of the *Interactive cluster* can frequently be reused with different models of the *Mental* and *Environmental* clusters. Moreover, their attached code is the most depending on the target platform. Once the integration is completed, the rest of the tool functionalities can be used taking the new integrated model specification as the current one.

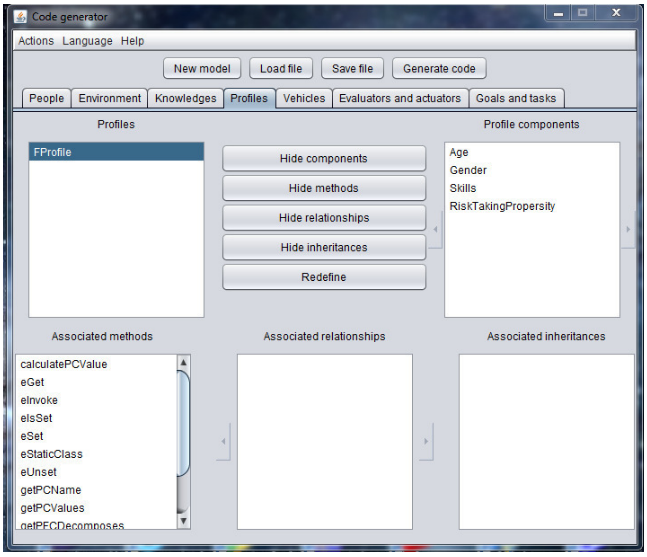


Fig. 4. Graphical interface of the code generator tool.

Regarding the platform adaptation, the code generator tool presents two main functionalities that promote the specialisation of the design and a better integration between the model specification and the target traffic platform: a dynamic compiler insertion and new classes generation.

The dynamic insertion assists in the attachment of libraries provided by the target platform (i.e. as external files) to the code generator, making them available to the compiler. Once the library or libraries are selected, the process is internally managed by the tool making it transparent to users. It allows producing platform-related elements in the source code of a model specification class.

The new class generation functionality supports the creation of new classes extending the original ones coming from these external libraries. It uses a wizard that eases the selection among the available classes. These classes are inserted dynamically in the path of the compiler and can be used as the others. This allows the creation of new objects of these classes in the model specification classes or vice versa.

Finally, when both the code transformation and platform adaptation are completed, the final file is produced. It can be generated using two different approaches: a specific plug-in and a new adjusted platform, being both supported by wizards to make the process more intuitive. The specific plug-in approach builds a compressed file packaging the model specification, the dependencies, and the classes generated that could be used or inserted into the target platform. The new adjusted platform approach integrates the entire target simulation platform and their dependencies (if they are needed), using external libraries, with the model specification and the classes modified and generated. The result is a runnable compressed file that contains the target platform. This platform is able to develop simulations considering the model specification inserted.

In both cases, configuration files can be created by another wizard. These files can be added to the compressed files in order to indicate parameters related to the simulation that must be considered.

The functionalities of these tools support our MDE approach for traffic simulation. They facilitate the process allowing the graphical examination of elements and the integration of multiple artefacts (e.g. model specifications or code snippets). Also, they encourage reusability and incremental development, reducing manual coding.

## V. Case study

The case study shows the use of the MDE infrastructure (i.e. the TML and the development tools) in order to produce a road traffic simulation. It integrates a model specification (compliant with the TML) that adapts a theory of the domain with the specialisation to a target traffic simulation platform.

In this case, the selected traffic theory [12] describes a classification of potential risk factors for drivers, and how these factors can influence their behaviour. It is modelled with the proposed TML, and uses the resulting model for generating source code through a semi-automatic process. Specific code snippets and classes are inserted for generating an adaptation that can run a simulation using the MATSim platform [13].

The original classification of risk factors presents multiple aspects structured around two main concepts: *Individual differences* and *Situational factors*. This classification does not follow the DVE [7] approach required by the TML, but its adaptation seems feasible, as both cover common aspects. For instance, the *Vehicle size* factor in [12] can be represented through the *VComponent* meta-class, the *Age* factor with the *PComponent* meta-class, and the *Trip purpose* factor using the *KComponent* meta-class.

The first step to model the traffic theory is focused on elaborating a *modelling plan*. This describes an initial evaluation of the elements coming from the traffic theory to model that can match with the types of the metamodel. Once this task is completed, these elements are mapped to the selected types of the *Mental* and *Environmental* clusters (see Sections III-A and III-B) that fit properly with them. This planning produces a *starting schema* as a result. These guidelines are represented as a model using the graphical editor tool as follows.

Users start producing a simple structure that contains only the needed instances from the root meta-classes of the metamodel to represent the concepts of the theory (i.e. the starting schema). In this case, the *FPerson* class (an instance of the *Person* meta-class) is the root of the model design. The other classes are related to it using their appropriate references, e.g. connecting the *FEnvironment* class (an instance of the *Environment* meta-class) and the *FKnowledge* class (an instance of the *Knowledge* meta-class) to the *FPerson* class.

The previous structure is the basis to integrate the rest of the theory. After creating it, users add the elements of the TML that represent the other factors considered in [12], and link them with the relevant main elements following the *modelling plan*. The *FProfile* class acts as a root of its own tree substructure, being decomposed into two *PComponent* children. They represent *Individual Differences* and *Individual* factors from the theory. Each one of them is in turn decomposed into several children (e.g. *Individual Differences* into *Age* and *Gender*; and *Individual* into *Impairment* and *Hurry/Distraction*). The *FKnowledge* class is decomposed into two *KComponent* children (i.e. *Trip purpose* and *Length of drive*). The *FVehicle* class is decomposed into two *VComponent* children (i.e. *Size* and *Performance Characteristics*). Finally, the *FEnvironment* class is decomposed into four *EComponent* children (e.g. *Weather* and *Road condition*). This completes the adaptation of the original model to the TML (see Fig. 5).

Once the model specification based on the *Mental* and *Environmental* clusters is completed, the next step is designing the elements of the *Interactive cluster*. This can be done with the graphical editor tool adding the elements previously defined, or with the graphical editor integrated in the code generator tool. In order to show how models can be reused, the second option is chosen, developing a self-contained model specification.

A self-contained specification is an independent model that comprises only elements of the *Interactive cluster* and *GeneralRelationship* instances. These components can be developed according to the requirements of a particular simulation platform, promoting their specialisation. The resulting model can be merged with other models based on the *Mental* and the *Environmental* clusters, which facilitates the integration of multiple traffic theories with the target platform.

In this case, the self-contained model takes as basis the approach presented in [23]. It provides a *Goal* and *Task* tree structure with *AND* and *OR* compositions. That comes from agent-based methodologies and the BDI model [6]. This tree structure presents tasks associated with most of the goals. These tasks achieve the actions of the individuals following a set of instructions.

The root goal called *ArrivedFastDestination* represents the basic goal of individuals involved in traffic. It is decomposed into two sub-goals that must be fulfilled (i.e. *AND* composition): *Actuated* and *EndedRoute*. In turn, the *Actuated* goal is decomposed into a set of alternative goals (i.e. *OR* composition) that represents the different actions individuals can choose while they are interacting in road traffic (see Fig. 6). These goals are decomposed into the alternatives to achieve them (e.g. the *SearchedObstacle* goal is decomposed into the *SearchedOnLeft* or *SearchedForward* sub-goals). When any of these goals that represent actions is satisfied, the *Actuated* goal is satisfied too. Meanwhile, the end of the route is

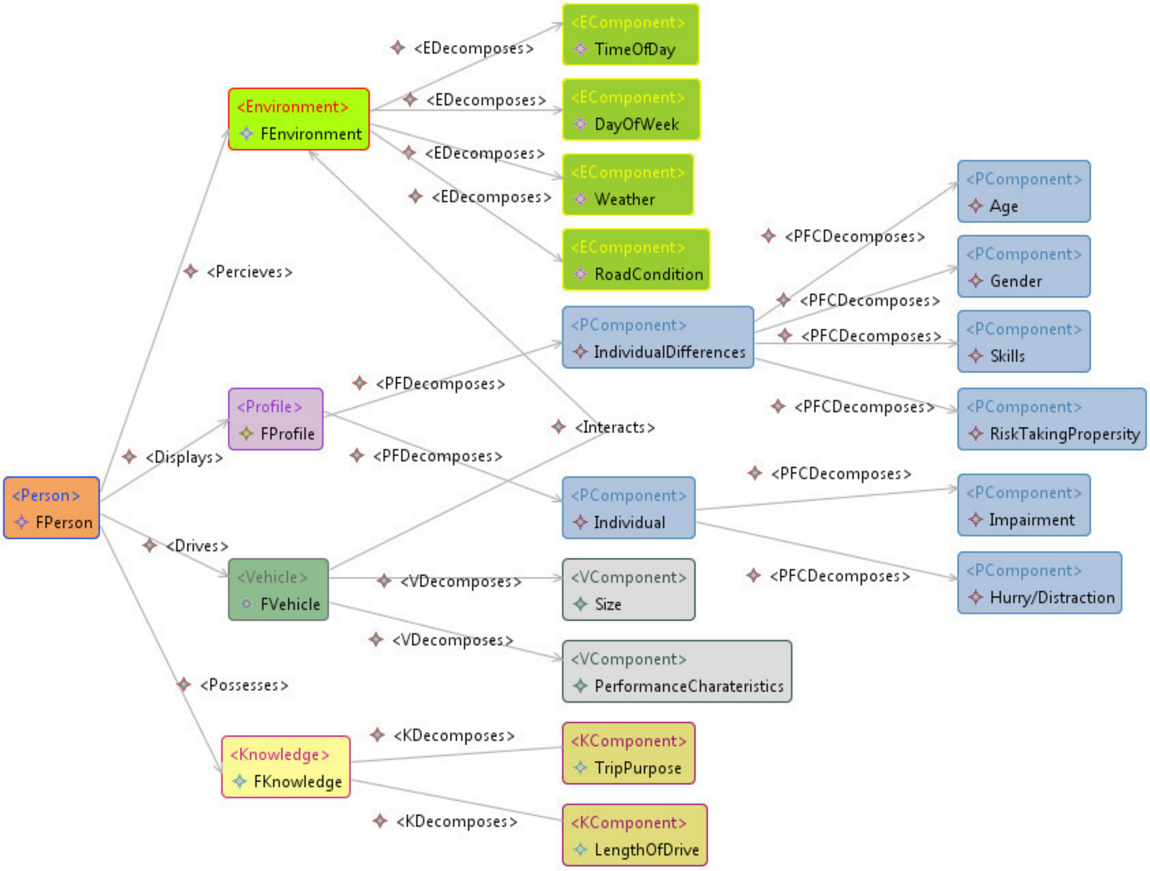


Fig. 5. Excerpt of the TML specification of the *Factors* model.

checked (*EndedRoute* goal satisfaction). If the *EndedRoute* goal is satisfied, then individuals have achieved their purpose, satisfying the *ArrivedFastDestination* root goal; if not, the process starts again. This sequence of actions models the processes in real-life, and assumes that at least one action must be done to reach the root goal.

In this self-contained model specification, *Goal* and *Task* instances present their own attributes and methods to manage these type of compositions. In *Goal* instances, the *GType* attribute indicates the type of goal composition (i.e. *AND* or *OR*). Code snippets are inserted into the body of the *calculateSatisfaction* method. These code snippets check if the sub-goal elements are satisfied. In *Task* instances, the *TType* attribute indicates the type of task composition (i.e. *AND* or *OR*), while code snippets complete the body of the *setInstructions* method. These code snippets validate if the associated sub-tasks and the atomic instructions are achieved successfully.

The *Interactive cluster* of the metamodel provides the means to model a *perceive-reason-act* cycle [11] of people. It uses the *Evaluator* and *Actuator* meta-classes based on [23].

In this case, *Evaluators* have a hierarchical decomposition that follows the goal tree. The root *Goal ArrivedFastDestination* is considered only by the *EvaluateDestination Evaluator* instance, the *EndedRoute Goal* is checked only by the *EvaluateRoute Evaluator* instance, and the *Actuated Goal* and the rest of *Goals* related to actions are controlled by the *EvaluateActions Evaluator* instance. This last *Evaluator* is in charge of selecting the best *Goal*. It considers the input parameters provided by both the *Mental* and the *Environmental* cluster, or the other two *Evaluator* instances. It also evaluates if the satisfaction of the selected *Goal* is produced in order to check if its parent *Goal* instances can be satisfied.

Once the structure in charge of evaluating the goals is completed, an *Actuator* instance is added to the self-contained model specification. It considers every task, executing its instructions when the appropriate evaluator selects its associated goal.

After that, this part of the model specification must be stored using the corresponding wizard of the code generator tool to generate a compressed file. In turn, the model based on the *Mental* and *Environmental* clusters generated in the graphical editor is loaded as input by the code generator tool. Then, navigating through the elements of the model specification, code snippets can be inserted into the *setXValues* body method of each one of the elements with the purpose of redefining the

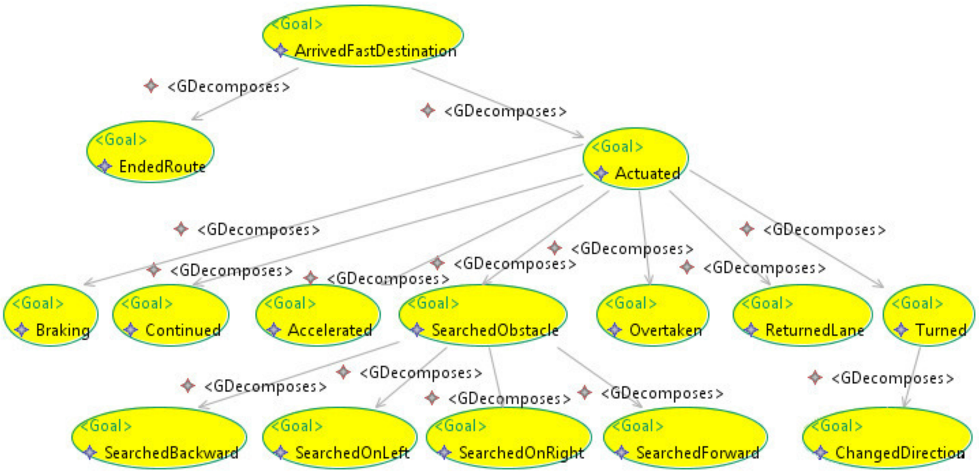


Fig. 6. Excerpt of the goal tree structure of the self-contained model specification.

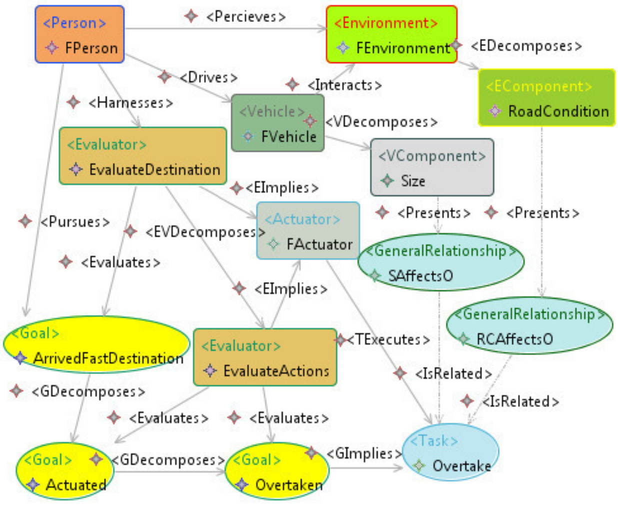


Fig. 7. Excerpt of the elements implicated in the overtaken interaction.

calculation procedure of the *XValue* attributes (see Section III). In this case, the code snippet applies a formula based on Fuzzy logic [24]. The value of an element is obtained adding every value of the children and its own value and dividing this result by its number of children plus one, establishing a relationship among the children components and the parent. This step can be changed to consider other formulas and theories.

This model based on the *Mental* and *Environmental* clusters can be stored using the appropriate code generator functionality. This allows reusing the fuzzy formula and the structure generated in other projects.

As soon as the fuzzy logic is inserted into the model, the wizard in charge of the cluster integration functionality can be selected to apply it. The wizard links the *FPerson* instance with the root *Goal* instance of the self-contained model specification (i.e. *ArrivedFastDestination*) through a *Pursues* reference. Then, the *FPerson* instance is linked to the *Evaluator* root instance (i.e *EvaluateDestination*) and vice versa using the references *Harnesses* and *IsHarnessed* respectively. Finally, the *Actuator* instance is connected to the *FPerson* instance by means of *Utilizes* reference. When the process is completed and every reference is established, both clusters become a single model specification.

The same wizard supports the integration of *GeneralRelationship* instances among both clusters of the single model specification. These *GeneralRelationship* instances indicate the influence of target elements over the *Task* instances associated with each *Goal* instance. These *GeneralRelationships* are considered by the appropriate *Evaluator* instance in order to select the best candidate *Goal*. It evaluates the *Xvalues* attributes (previously configured using Fuzzy logic) of each related element of the *Mental* and *Environmental* clusters in order to generate real-time decisions.

Here, the addition of *GeneralRelationship* instances follows the factors structure. *Environment* factors (i.e. *RoadCondition* or *TimeOfDay*) directly affect *Overtake* or *Brake* instances, establishing multiple *GeneralRelationship* instances among them (see example in Fig.7). *IndividualDifferences* factors (i.e. *Age* or *RiskTakingPropensity*) affect *Accelerate* or *ReturnLane Task* instances. *Individual* factors (i.e. *Impairment* or *Hurry/Distraction*) affect *SearchObstacle* or *Accelerate Task* instances. *Vehicle* factors (i.e. *Size* or *PerformanceCharacteristics*) affect *Accelerate* or *Turn Task* instances. *Knowledge* factors (i.e. *TripPurpose* or *LengthOfDrive*) affect *Overtake* or *SearchObstacle Task* instances.

When the integration of these *GeneralRelationship* instances is completed, all these concepts and other possible evaluation criteria must be considered by the *Evaluator* instances. For this, the appropriate code snippets must be inserted into the body of the *evaluateGoals* method.

In order to illustrate how the code generator tool generates specific source code for a target platform, this case study considers MATSim. This agent platform presents different functionalities, but it only supports route configuration and optimisation using a path to follow (i.e. the interactions of individuals are not considered). It is made available to the code generator tool as an external compressed file, adding all its libraries and dependencies to the compiler.

The integration requires developing and adding a new class using the related wizard. This class is in charge of merging the model specification and the platform by establishing communications between them through programming procedures (i.e. the model specification is encapsulated being able to be integrated as a class in the MATSim source code).

Some classes of the MATSim platform need to be extended in order to consider the model specification structure. The original classes only plan the route, and now they must also consider, for instance, overtaking or lane changes. The new classes are integrated into the project, being considered by the compiler of the code generator.

The *Instructions* attribute of *Task* instances must be adapted and redefined to the way of functioning and the source code provided by MATSim. This allows generating the proper platform-specific actions to get the intended behaviour.

Once the specialisation is achieved, a configuration file is generated through the corresponding wizard of the code generator tool. The initial parameters of the *XValues* attributes are defined. These parameters can be modified with the purpose of obtaining different influences of elements. Also, another class is developed in the code generator and integrated into the path of its compiler in order to load this configuration file when the road traffic simulation starts.

Finally, the code generator tool produces a compressed file. It directly runs the MATSim platform with the embedded model specification generated and integrating the configuration file.

## VI. Related work

Road traffic simulation is related to multiple areas of research. The presented approach considers the modelling of people's behaviour and environmental features affecting traffic, and the development process of simulations.

Existing road traffic simulation platforms are mainly based on multiple drivers that follow paths, though some of them allow random behaviours. The differentiation of their features, the decision-making and the interaction among them or with the environment are considered only in limited ways [13], [25].

Pedestrians and the influence of people around (e.g. passengers) over drivers are also important elements to evaluate, but frequently disregarded. For instance, [26] can model pedestrian interacting with the environment and drivers, but passengers and their possible impact are not contemplated.

The proposed metamodel considers the different roles of the individuals involved in traffic (i.e. drivers, pedestrians and passengers). It corresponds to microscopic models, as it models the multiple individual artefacts involved in road traffic (e.g. instances of *Person* and *Vehicle*). Although not fully considered now, mesoscopic models (i.e. those combining the individual and group levels), could also be integrated in the ML. The ABM approach adopted in the ML facilitates this extension, as it is frequently adopted for such kind of models [27]. The metamodel structure improves existing approaches in order to embody multiple social features. Most of approaches consider a fixed set of these features and their relationships, e.g. [9]. *Knowledge* instances are designed to be specialised and combined, providing instruments to add facts that affect groups of entities or the overall simulation.

Regarding the internal modelling of individual participants involved in traffic, there is not a widely accepted approach. Models range from simple, mainly reactive ones, to quite complex, usually deliberative. For instance, in [28] agents use simple logical rules to interact with the environment. This environment is mainly composed of crossroads where agents react to the behaviour of others. A more complex approach appears in [29], where driver's actions are decomposed into workflows considering the multiple situations that can occur during their execution.

The decisions achieved by agents in the previous approaches can be combined in hierarchical architectures, where there are several abstraction layers that organize acting. An example of this is the Michon's hierarchical control model for drivers [30]. The metamodel supports the hierarchical composition of most of its elements, but not the definition of abstraction layers as required by hierarchical architectures.

Another point of discussion in literature is related to which of the features of participants and the environment have influence on road traffic. Approaches such as [31], [32] review some of these features. The metamodel is intentionally open in this aspect. Meta-classes such as *Vehicle* and *Environment* in the *Environmental cluster*, and *Knowledge* and *Profile* in the *Mental cluster*, present sub-components to classify other related elements or characteristics. The metamodel also allows introducing additional concepts (i.e. through the *GeneralElement* meta-class) and relationships (i.e. through the *GeneralRelationship* meta-class), and extending them using the different inheritance hierarchies. These aspects entail that the TML is highly customisable for the multiple requirements of the road traffic domain.

Regarding the development process, most of reviewed works do not cite the approach they follow. Those that do it, in general assume common development processes focused on source code, where models play only a documentation and communication role. The advantages of MDE in this scenario have been already discussed in the related literature [33]: explicit representation of the information, higher involvement of experts, enhanced model validation, and reusability.

## VII. Conclusions

This paper has presented a metamodel that defines a TML to support a MDE approach for the development of road traffic simulations. It defines this extensible ML focused on the behaviour of individuals. Development tools compliant with the metamodel are provided to support the process. The adoption of MDE facilitates the exchange of information among groups of experts with different backgrounds. It also promotes the reutilisation of artefacts between projects, as there is a clear separation of concerns and all the information is explicit. For instance, this facilitates deployments in multiple platforms. It also encourages the incremental development, as models and transformations can be more easily modified than code.

The metamodel is designed with the purpose of being able to integrate multiple theoretical works from the domain. It follows an ABM [8] approach in order to consider the social interactions of the individuals involved in road traffic. It has three main clusters: a *Mental cluster*, an *Environmental cluster*, and an *Interactive cluster*. The first one includes the different psychological attributes that can influence the behaviour of individuals [4]. These concepts are classified into two groups: the features of people and their current state. It considers aspects of the mental state in the agent paradigm [5], particularly the BDI model [6]. The second cluster is based on the DVE approach [7] for modelling the different interactions of the individuals involved in road traffic, considering the relationships among vehicles, environment, and people. Many of the existing studies can fit their concepts into this structure, as it includes widely accepted notions to describe traffic settings. The last cluster uses concepts like goal and task. A *perceive-reason-act* cycle [11] is integrated through the evaluator and actuator concepts.

The meta-classes of the metamodel are designed to support internal hierarchical substructures, e.g. the container *Profile* meta-class and its sub-components with the *PComponent* meta-class. Inheritance between elements of the same type is introduced to make possible specialisations. Other types of relationships among elements are also considered.

Development tools are based on Eclipse facilities [17], [18]. The graphical editor provides a visual interface and a palette for describing the model specifications. These specifications can be validated to guarantee its compliance with the ML. The code generator provides a set of functionalities that guide users in the production of source code for a given target traffic simulation platform.

The case study exemplifies the use of the complete MDE infrastructure. The development tools support the design of a model specification according to a theory, its specialisation to the MATSim platform, and the generation of the source code associated. The MATSim platform presents a route optimisation feature to simulate individuals involved in traffic. This feature does not consider interactions among individuals and only generates a path to follow. Here, it is improved adding decision-making actions based on [23] through a goal-task hierarchical structure with *OR* and *AND* compositions. A tailored *perceive-reason-act* cycle [11] is integrated using the *Evaluator* and *Actuator* meta-classes. Also, the resulting model integrates a taxonomy related to the traffic domain based on the risk factors for drivers [12]. Individual actions are influenced by the related factors, producing different behaviours in individuals when those factors change.

The presented approach has several open issues. The TML must be tested with other types of road traffic theories (e.g. interactions among drivers and pedestrians) in order to check its primitives and structure. The development tools must also be used to generate source code specialisations for other traffic agent platforms (e.g. SUMO [25] or VISSIM [26]). The introduction of social norms, the influence of traffic signals (e.g. crossings or traffic lights) and the types of vehicles (e.g. ambulances or motorbikes) could be considered.

## ACKNOWLEDGMENT

This work has been done in the context of the project "Social Ambient Assisting Living - Methods (SociAAL)" (grant TIN2011-28335-C02-01) supported by the Spanish Ministry for Economy and Competitiveness, and the research programme MOSI-AGIL-CM (grant S2013/ICE-3019) supported by the Autonomous Region of Madrid and co-funded by EU Structural Funds FSE and FEDER.

## REFERENCES

[1] A. Crooks, C. Castle, and M. Batty, "Key challenges in agent-based modelling for geo-spatial simulation," *Computers, Environment and Urban Systems*, vol. 32, no. 6, pp. 417–430, 2008.

[2] R. France and B. Rumpe, "Model-driven development of complex software: A research roadmap," in *2007 Future of Software Engineering*. IEEE Computer Society, 2007, pp. 37–54.

[3] A. Van Deursen, P. Klint, and J. Visser, "Domain-specific languages: An annotated bibliography." *Sigplan Notices*, vol. 35, no. 6, pp. 26–36, 2000.

[4] D. Shinar, *Psychology on the Road. The Human Factor in Traffic Safety*. John Wiley & Sons, 1978.

[5] Y. Shoham, "Agent-oriented programming," *Artificial Intelligence*, vol. 60, no. 1, pp. 51–92, 1993.

[6] A. S. Rao and M. P. Georgeff, "An abstract architecture for rational agents," in *Proceedings of Knowledge Representation and Reasoning (KR&R-92)*, vol. 92, 1992, pp. 439–449.

[7] A. Amditis, K. Pagle, S. Joshi, and E. Bekiaris, "Driver–vehicle–environment monitoring for on-board driver support systems: Lessons learned from design and implementation," *Applied Ergonomics*, vol. 41, no. 2, pp. 225–235, 2010.

[8] M. A. Janssen, "Agent-based modelling," *Modelling in Ecological Economics*, pp. 155–172, 2005.

[9] J. Pavón, J. J. Gómez-Sanz, and R. Fuentes, "The INGENIAS methodology and tools," *Agent-Oriented Methodologies*, vol. 9, pp. 236–276, 2005.

[10] P. Bresciani, A. Perini, P. Giorgini, F. Giunchiglia, and J. Mylopoulos, "Tropos: An agent-oriented software development methodology," *Autonomous Agents and Multi-Agent Systems*, vol. 8, no. 3, pp. 203–236, 2004.

[11] J. Lind, "Issues in agent-oriented software engineering," in *Proceedings of the First International Workshop on Agent-Oriented Software Engineering (AOSE)*. Springer, 2001, pp. 45–58.

[12] T. A. Ranney, "Psychological factors that influence car-following and car-following model development," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 2, no. 4, pp. 213–219, 1999.

[13] Transport Systems Planning and Transport Telematics Group, Transport Planning Group and Senozon Company, "MATSim, Multi-agent transport simulation," http://www.matsim.org/, 2015, [Online: accessed 08-May-2015].

[14] J. Bézivin, "Model driven engineering: An emerging technical space," in *Generative and Transformational Techniques in Software Engineering*. Springer, 2006, pp. 36–64.

[15] S. Kent, "Model driven engineering," in *Integrated Formal Methods*. Springer, 2002, pp. 286–298.

[16] Object Management Group, "Meta-Object Facility (MOF) Core Specification, Version 2.4.2," 2014.

[17] D. Steinberg, F. Budinsky, E. Merks, and M. Paternostro, *EMF: Eclipse Modeling Framework*. Pearson Education, 2008.

[18] D. Rubel, J. Wren, and E. Clayberg, *The Eclipse Graphical Editing Framework (GEF)*. Addison-Wesley Professional, 2011.

[19] Object Management Group, "Object Constraint Language (OCL), Version 2.4," http://www.omg.org/, 2014, [Online: accessed 07-May-2015].

[20] K. Czarnecki and S. Helsen, "Feature-based survey of model transformation approaches," *IBM Systems Journal*, vol. 45, no. 3, pp. 621–645, 2006.

[21] M. Wimmer and L. Burgueño, "Testing m2t/t2m transformations," in *Model-Driven Engineering Languages and Systems*. Springer, 2013, pp. 203–219.

[22] W. Van Der Hoek and M. Wooldridge, "Multi-agent systems," *Foundations of Artificial Intelligence*, vol. 3, pp. 887–928, 2008.

[23] A. Fernández-Isabel and R. Fuentes-Fernández, "An agent-based platform for traffic simulation," in *Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011*. Springer, 2011, pp. 505–514.

[24] C. P. Pappis and E. H. Mamdani, "A fuzzy logic controller for a traffic junction," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 7, no. 10, pp. 707–717, 1977.

[25] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumo-simulation of urban mobility-an overview," in *SIMUL 2011, The Third International Conference on Advances in System Simulation*, 2011, pp. 55–60.

[26] Visual Solutions, Incorporated, "VisSim, A graphical language for simulation and model-based embedded development," http://www.vissim.com, 2015, [Online: accessed 08-May-2015].

[27] M. Vasirani and S. Ossowski, "A market-inspired approach to reservation-based urban road traffic management," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2009, pp. 617–624.

[28] A. Doniec, R. Mandiau, S. Piechowiak, and S. Espié, "A behavioral multi-agent model for road traffic simulation," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 8, pp. 1443–1454, 2008.

[29] B. Burmeister, A. Haddadi, and G. Matylis, "Application of multi-agent systems in traffic and transportation," in *IEEE Transactions on Software Engineering*, vol. 144, no. 1. IET, 1997, pp. 51–60.

[30] J. A. Michon, "A critical view of driver behavior models: what do we know, what should we do?" in *Human Behavior and Traffic Safety*. Springer, 1985, pp. 485–524.

[31] H. Greenberg, "An analysis of traffic flow," *Operations Research*, vol. 7, no. 1, pp. 79–85, 1959.

[32] P. Paruchuri, A. R. Pullalarevu, and K. Karlapalem, "Multi agent simulation of unorganized traffic," in *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*. ACM, 2002, pp. 176–183.

[33] R. Fuentes-Fernández, S. Hassan, J. Pavón, J. M. Galán, and A. López-Paredes, "Metamodels for role-driven agent-based modelling," *Computational and Mathematical Organization Theory*, vol. 18, no. 1, pp. 91–112, 2012.

DOI: 10.15439/2015F213
ACSIS, Vol. 5

# Multi-Agent System Simulation of Indoor Scenarios

Rafael Pax
Universidad Complutense de Madrid
Facultad de Informática,
28040 Madrid, Spain
Email: rpax@ucm.es

Juan Pavón
Universidad Complutense de Madrid
Facultad de Informática,
28040 Madrid, Spain
Email: jpavon@fdi.ucm.es

***Abstract*—This paper presents a flexible agent decision model for the simulation of indoor scenarios. There are different kind of applications with varying requirements, from the typical emergency evacuation where the physical interaction of crowds of agents are more relevant, to those that demand more sophisticated agent decision models as when testing smart environment applications. Existing tools usually focus on one of these issues, looking for efficiency in the solutions.**

**The agent decision model in this paper tries to get a balance between efficiency and flexibility in the specification of the agent behavior in simple and complex situations. This is applied in a simulation framework for indoor scenarios, although it could be extended to other settings.**

## I. Introduction

TESTING APPLICATIONS for smart environments is a difficult task. It requires the installation of sensors and actuators, the communications and the software for the control system, and the participation of persons who have to play the different scenarios. This is costly, both in economic sense as well as in time. Also, there are some situations that cannot be tested for practical reasons (e.g., a fire, people accidents). Furthermore, from the point of view of the developers, who are used to iterative processes, it is difficult to repeat the tests if they have to perform these with persons. At least for these reasons it is interesting to use simulation tools that provide some support for the development of smart environment applications.

A relevant aspect to be considered in this kind of tests is the modelling of the behavior of humans under different situations. The behavior for these scenarios requires at least the following: interactions among agents, with the environment, and the process for decision making.

There are several tools for simulation and design of how people behave in indoor scenarios (some of them commercial) [1]–[6]. They focus on the design of spaces and 3D appearance, where the agents are seen more like a crowd that can be characterized by simple behaviors with a fixed number of parameters, instead of considering them as individuals. Although they are appropriate to simulate specific scenarios, it is important to consider the human and social behavior of individuals when simulating how people interact with their environment, including other individuals.

Other works have better addressed the specification of the agent behavior, such as [7]–[13]. However, they have not sufficiently taken into account the methodological aspects for a design process when developing the agents' behavior. This is relevant when the simulation framework has to be used for different purposes and by other developers. In those cases, there is a need for a clearer agent model, with some support for the design at a higher level of abstraction that can be easily translated to an implementation. This is the purpose of MASSIS (**M**ulti-**A**gent **S**ystem **S**imulation of **I**ndoor **S**cenarios), a framework for modelling and simulation of the decision-making process of agents in multiple situations in indoor scenarios domain.

The rest of the paper is structured as follows. Section II describes the MASSIS framework. Section III explains how is modelled the agent's behavior. Section IV shows a case study to illustrate the approach. Finally, Section V presents our conclusions.

## II. The MASSIS framework

MASSIS is an agent-based simulation framework for indoor scenarios. It has a component-based architecture, built on open-source software components. An overview of the framework is shown in Figure 1.

Graphical modelling of the indoor environment is supported by SweetHome3D [14]. This is a well known package that is used to model all components involved in the simulation environment, such as walls, doors, stairs, people, etc. (see Fig. 2). Sweet Home 3D allows to design buildings with enough realism, in a relatively short time. It also allows the integration of extensions, providing significant flexibility when adding new features. MASSIS adds a set of plugins for this application, which lets the user to specify the characteristics of the elements of the building that will act as agents. In the case of people, can be weight, speed, inherent characteristics of the person (fear, courage, etc.) and link to their behavior. It is also possible to specify elements of the environment, such as sensors and actuators, which will have a reactive behavior.

When the building is created, the SweetHome3D representation of the building is transformed inside the simulation engine, which adapts it to the internal representation of MASSIS. One of the important issues when modeling

This work has been developed in the context of the project MOSI-AGIL-CM (with grant S2013/ICE-3019, by the Directorate General for Universities and Research of CAM) and by the Programa de Financiación de Grupos de Investigación UCM-Banco Santander with reference GR3/14.

978-83-60810-66-8/$25.00©2015, IEEE

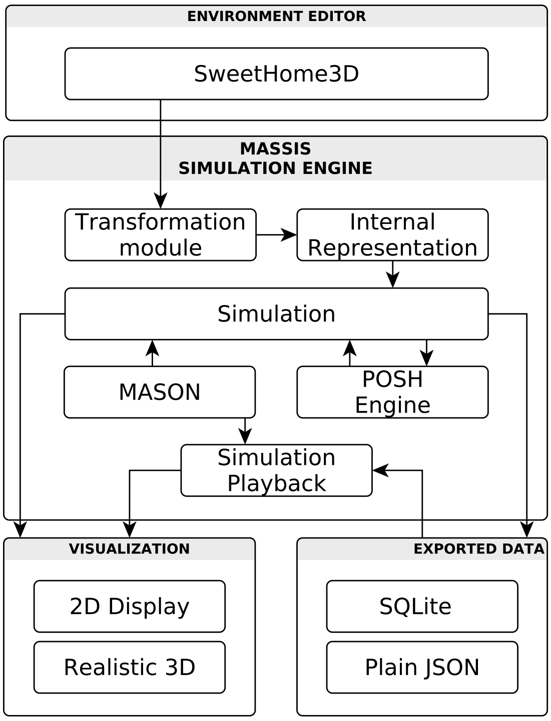


Fig. 1. MASSIS framework overview

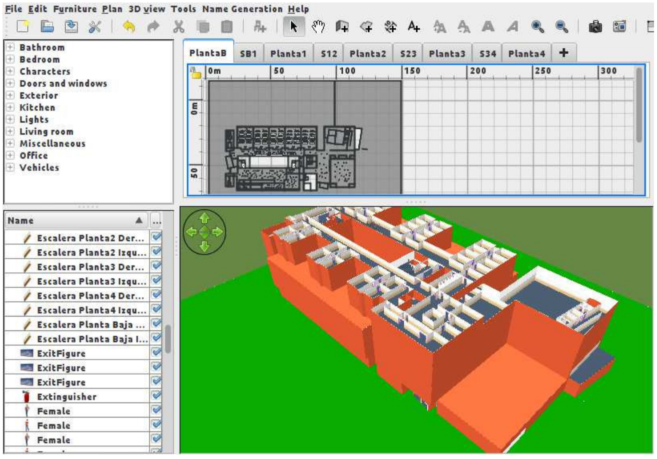


Fig. 2. Screenshot of the MASSIS's 3D editor, based on SweetHome3D.

indoor scenarios is the accuracy of the representation. Inside a building it is very common to have various elements at very small distances. If the elements of space are assigned to cells, the accuracy is reduced considerably. To solve this problem, MASSIS represents the elements in the building without discretizing the space; each element is represented as a polygon. For efficient computation of the locations of agents, data structures appropriate to this representation are used, such as different models of quadtree and polygon meshes. These data structures are used both for locating agents, perception and pathfinding.

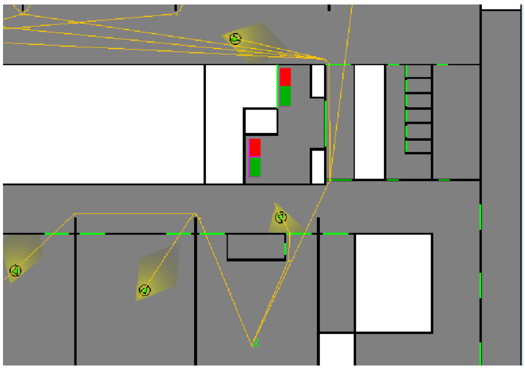


Fig. 3. Standard schematic display of MASSIS, showing bodies radio, vision areas and paths (*black circles green arrows*, *yellow polygons* and *yellow lines*, respectively). The green lines are doors, and the red and green squares are stairs.

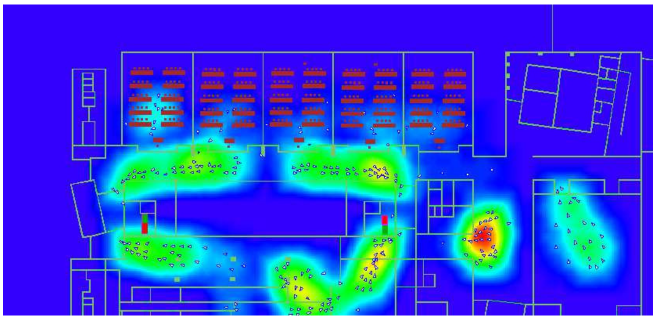


Fig. 4. Example of an user-defined schematic display: crowd density.

As simulation engine, MASSIS uses MASON [15], a lightweight multi-purpose agent-based simulation library. MASON has been chosen because it provides a good support for agent-based simulation platform, with well proven efficiency. Also, the clear separation of the simulation core and the GUI, allows MASSIS to use the MASON simulation core, while using its own display system.

The agents' behavior is controlled with the Pogamut's [16] POSH engine. (See Section III for more details)

All the changes made in the environment are reflected in real time by MASSIS' 3D (Fig. 9) and 2D (Fig. 3) displays, and can be logged in JSON format (Listing. 1), as a single zipped file or in a SQLite database for further analysis. Although 3D display is more realistic, the 2D view is useful for analysis and debugging. Also, the 2D visualization API allows the creation of user-defined layers (Fig. 4).

Once a simulation is performed, the exported data can be used to playback all events that have occurred during the execution of the simulation, i.e., the agents will behave in the same way they did during the simulation.

## III. Agent Decision Model in MASSIS

The aspects of human behavior that are of interest for modelling indoor scenarios in the MASSIS framework are the mechanisms that humans use to deal with problems reasoning

```
{
    "velocity": { "x": 32,"y": 58},
    "visionRadio": 300,
    "maxforce": 10,"maxspeed": 15,
    "properties": {"steering.separation":
        70,...},
    "locationState":{
    "angle": 0.7853982,"floorId": 8,
    "centerX": 4975.2285,"centerY":
        4108.2695,...},
    "id": 3673
}
```

Listing 1. An agent's saved state in *JSON* format.

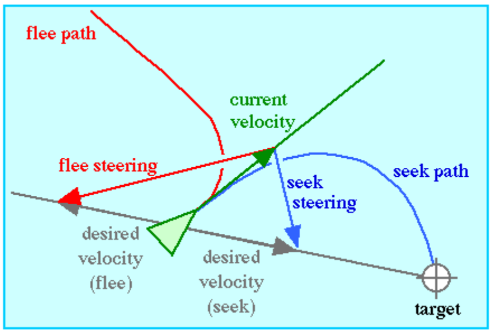


Fig. 5. *Seek and Flee* Steering behavior [17].

from context, making use of collective intelligence, and how this intelligence is used in problem solving.

Each agent in MASSIS has its own behavior, which is computed at a high-level (reactive plans) and at a low level (speed, position, angles, density, etc.)

When the high level decides *what to do next*, the action is executed by the low-level module, which carries all the necessary operations (movement, animation, etc.) Both high-level and low-level behaviors are affected by the state of the agent, altering the decision making process (e.g., a *scared* agent may choose a different route to reach a target, probably a longer one) and the action execution (it will be moving faster).

### A. Low-level behavior

The low-level module deals with the perception of the environment and a set of basic behaviors for interacting with it. These behaviors are mostly a set of *steering behaviors* [17], which control the most basic movement component of the agent.

Using a simple force model, steering behaviors produce smooth, life-like movements, providing agents the ability to navigate around the environment in a realistic manner. The forces applied to the agent can be combined in order to create more complex behaviors (e.g. collision avoidance, path-following, leader following, queuing, etc.). Fig. 5 shows the forces present in the basic *Seek and Flee* behavior.

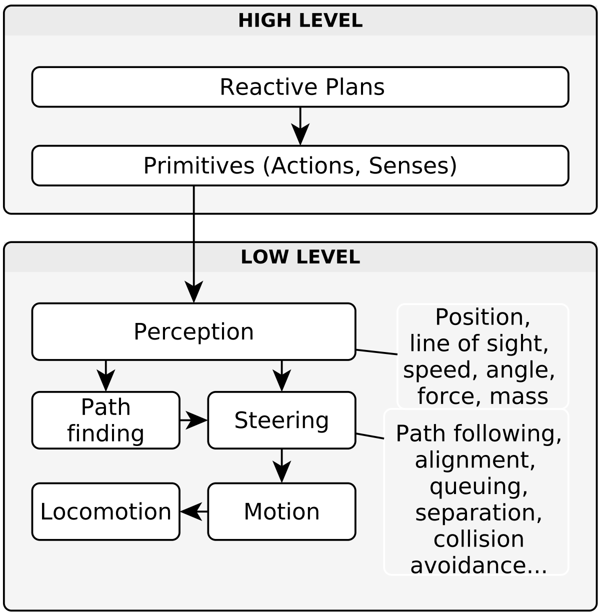

HIGH LEVEL
Reactive Plans
Primitives (Actions, Senses)
LOW LEVEL
Perception
Position, line of sight, speed, angle, force, mass
Path finding
Steering
Path following, alignment, queuing, separation, collision avoidance...
Locomotion
Motion

Fig. 6. MASSIS 's human behavior agent model.

### B. High-level behavior

The high-level behavior deals with decision making, learning and communications with other agents and makes decisions based on the knowledge of the environment, which is provided by the low-level module.

The architecture of this behavior follows the BOD (Behavior Oriented Design) method [18]. This method for building agents combines the advantages of Behavior Based AI [19], [20] and object-oriented design approaches.

In MASSIS this is applied to facilitate the design of agents that are capable of running in parallel and of generating a behavior that can satisfy multiple objectives that may conflict with each other.

The difficulty of making an autonomous agent is that many of the goals that the agent wants to be accomplished must be carried out at the same time. An agent may have the desire to be loved, be promoted at work and having breakfast in the morning.

Additionally, these goals must be achieved in an unpredictable environment, which can complicate or even make easier the way in which the agent tries to accomplish its goals. Developing a system of agents under BOD involves dividing the implementation into two different parts:

1) A library of Behavior modules. They consist of a set of classes representing a set of modules for perception, action and learning. These are *primitives*, actions and senses that can be called from the mechanism of action selection. They also provide a place where certain states and knowledge can be stored in order to perform those actions, and they contain code that describes any sense that needs to be carried out to acquire that state and

```java
public ActionResult run() {
    boolean isInLoc;
    SimulationObject target=getTarget();
    Location tLoc=target.getLocation();
    isInLoc=agent.approachTo(tLoc);
    if (isInLoc) {
        return ActionResult.RUNNING_ONCE;
    }
    else {
        return ActionResult.FINISHED;
    }
}
```

Listing 2. Example of the primitive action *Go to target*

```
(C search-for-object
vars($type, $storeFlag="IS_NEAR_TARGET")
(elements
 ((has-target (trigger ((HasTarget))) approach
    ))
 ((is-object-visible
  (trigger ((SeesElement($attr=$type,$value=1)
      true ==)))
  setTarget($target="?LAST_SEEN_OBJ")))
((search ExploreAction)))
)
```

Listing 3. POSH code of parametrized competence *search-for-object*. *ExploreAction* is a primitive action, *approach* is a competence and words preceded by *$* are variables.

knowledge. In brief, they determine *how* to do something. These senses and actions are created in the native language for the problem space (in the case of MASSIS, Java; see for instance the code in the listing 2)

2) POSH Dynamic action selection scripts. These allow to determine priorities between modules. The BOD architecture uses a POSH dynamic plan when an action should be carried out.
A POSH(Parallel-rooted, Ordered Slip-stack Hierarchical) plan is a prioritized set of conditions and the related actions to be performed when the conditions have been met.
It consists of drive collections, competences, and action patterns.
    - Drive collections are the root of every POSH plan. On the action selection step, the drive collections select which goal the agent must try to accomplish. They can be seen as a set of conditional rules, that are evaluated from highest to lowest priority.
    Every time the condition of the drive collection with highest priority is satisfied (a higher rule interrupts a lower one), the POSH engine executes the corresponding action pattern or competence.
    - Competences are a set of nested if-then conditional trees, which can be reused several times inside the reactive plan. They differ from the drive collections in the way they are executed; rules they do not interrupt each other.

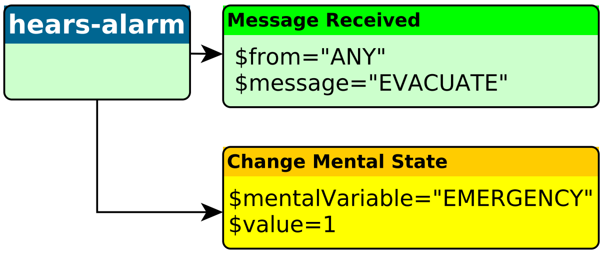


Fig. 7. *Mental state* modification in a POSH plan (*yellow*), triggered by the sense "Message Received" (*green*) on the Drive Element "hears-alarm"(*blue*)

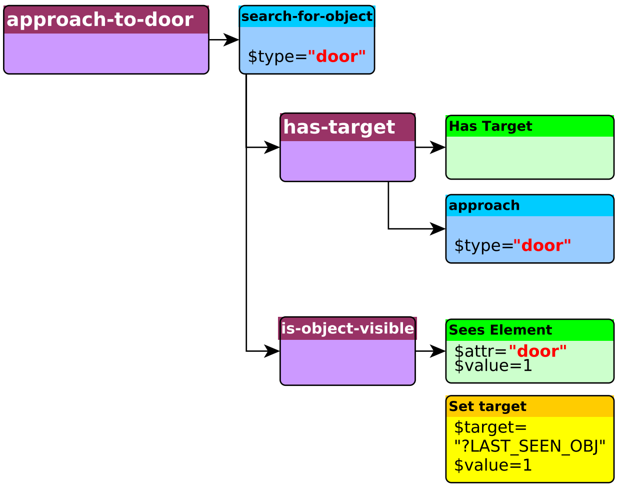


Fig. 8. Propagation of the value "door" in the parametrized competence of searching an object. **Note**: Some elements were omitted for clarity.

    - Action Patterns are simple sequences of actions. Although they are not very flexible, they provide a layer of abstraction very useful when grouping actions.

MASSIS encourages the use of variables and the agent's *Mental State* in POSH plans. *Mental States* are intended for representing the knowledge of the agent about its environment as a set of key-value pairs, but can be used for any other purpose, such as storing control variables in order manage the plan execution (see Fig. 7).

Also, as MASSIS uses the Pogamut's extension of the POSH language, *actions*, *senses*, *action patterns* and *competences* can be parametrized. This provides considerable flexibility, allowing the reuse of elements in the reactive plans (see Fig. 8 and Listing 3).

## IV. Case study: Emergency simulation

Public buildings have some protocols for dealing with emergency situations, which may involve, for instance, evacuation of the building. Testing these protocols requires some planning and cost. Simulation can help to this task. This case study addresses this kind of situation for the building of the Facultad de Informática at UCM. In this scenario, a teacher is responsible for guiding students safely to the building's exit in case of emergency. For illustrating purposes, this is the protocol for a teacher in an emergency situation:

> *When the alarm sounds the teacher of the group should go to the classroom door and order the students to close the windows if there is a fire. If instead of a fire there is a bomb threat, windows and doors should be left open. Students will leave the classroom through the door and they will be waiting for the teacher outside. The teacher will be the last person to leave the classroom. Once there is nobody in the classroom, the teacher will place a chair at the entrance of it,as an indication that the room has been evacuated entirely. Then the teacher will guide students toward the nearest exit.*

Modelling the teacher agent involves the following basic skills:

- Hearing and vision capabilities.
- Ability to communicate with other agents by voice.
- Movement.
- Interaction with objects in the environment: Taking an object, carrying it , dropping it.

These skills are candidates to be *primitive* actions and senses. These primitives, such as the movement one (Listing 2), are used by the reactive plan as *Triggers* of *Drive Elements*, components of *Action Patterns*, or they form part of one or more *Competences*. Figure 10 shows part of the teacher's reactive plan. Figures 9 and 11 show the initial state of the simulation. When the alarm sounds, and the teacher goes to the door (using *go-to-nearest-door* competence). Figure  12 illustrates the moment when the teacher tells the students that they must close the windows (*act-windows* competence). When the windows are closed, and the students outside (Fig. 13), the teacher takes the first chair he sees and he moves it to the

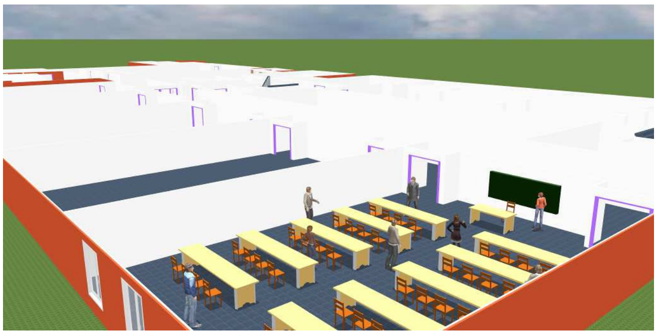


Fig. 9. 3D view of the simulation

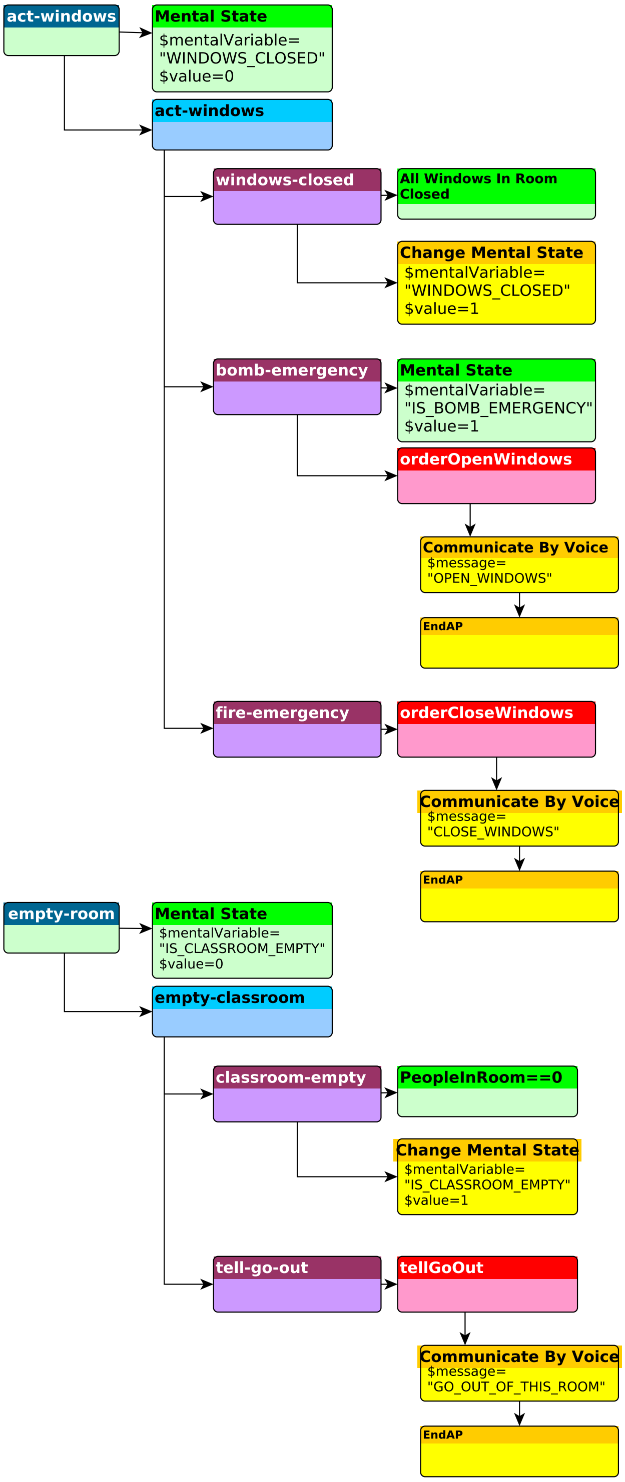


Fig. 10. Partial overview of the Case Study POSH plan

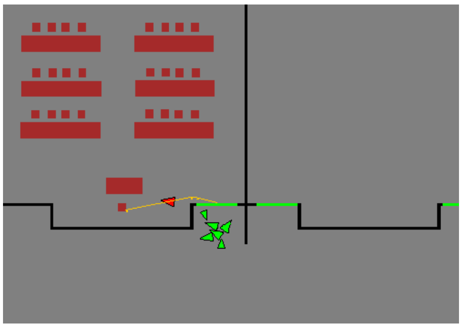


Fig. 13. Teacher going to take the nearest chair.

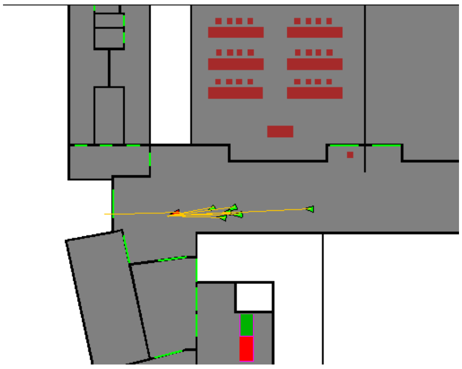


Fig. 14. The students follow the teacher for escaping from the building.

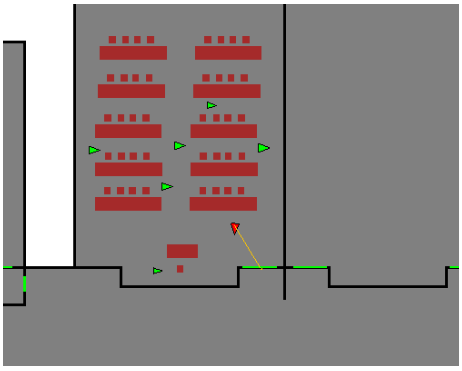


Fig. 11. Teacher going to the door.

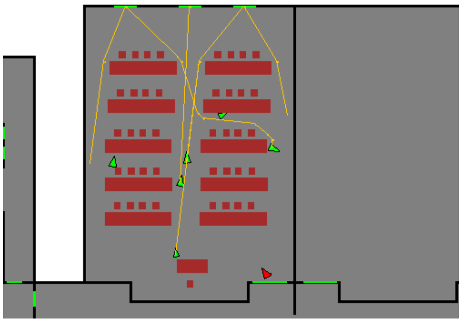


Fig. 12. The students receive the ”close windows” message, and they proceed to close the windows.

door. After that, the teacher guides his students to the nearest exit (cf. Fig. 14).

## V. Conclusion and Future Work

This paper has presented MASSIS, a multiagent-based simulation framework that supports the decision making process of humans when solving problems. Agent behavior is structured in low-level and high-level behavior components, extending Pogamut's POSH implementation model, with the addition of features that facilitate the separation of decision-making process and low level actions. MASSIS provides a rich set of low-level behavior components for the simulation of indoor scenarios. This has required to MASSIS the extension of the SweetHome3D environment with plugins for linking agent's behavior in the simulation. In order to apply MASSIS to other kind of scenarios (e.g., a city), new low-level behavior components should be implemented and integrated with another graphical design package that supports the definition of the new environment. In this sense, MASSIS can be easily extended. MASSIS provides as well a rich log capability, which can be the basis for further analysis of the scenarios. The integration of existing analysis tools is one of the most relevant issues for the next version ofthis framework.

## Acknowledgment

A special thanks to P.R. Hijas for her comments that greatly improved the development of the MASSIS' framework.

## References

[1] Legion, “Science in Motion,” http://www.legion.com, [Online; accessed Mar. 2015].

[2] M. Owen, E. R. Galea, and P. J. Lawrence, “The exodus evacuation model applied to building evacuation scenarios,” *Journal of Fire Protection Engineering*, vol. 8, no. 2, pp. 65–84, 1996.

[3] PedGo, “TraffGo HT.” http://www.traffgo-ht.com/de/pedestrians/products/pedgo/index.html, 2006, [Online; accessed Mar. 2015].

[4] M. MacDonald, “STEPS,” http://www.steps.mottmac.com/, 2009, [Online; accessed Mar. 2015].

[5] T. Engineering, “Pathfinder,” http://www.thunderheadeng.com/pathfinder/, 2006, [Online; accessed Mar. 2015].

[6] Golaem, “Golaem Crowd: Artist-Driven Crowd Simulation,” http://golaem.com/content/products/golaem-crowd/overview, 2011, [Online; accessed Mar. 2015].

[7] X. Pan, C. S. Han, K. Dauber, and K. H. Law, “A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations,” *Ai & Society*, vol. 22, no. 2, pp. 113–132, 2007.

[8] L. Saîfi, A. Boubetra, and F. Nouioua, “Approaches to modeling the emotional aspects of a crowd,” in *Modelling and Simulation (EUROSIM), 2013 8th EUROSIM Congress on*. IEEE, 2013, pp. 151–154.

[9] M. Software, “Simulating Life,” http://www.massivesoftware.com/, 2002, [Online; accessed Mar. 2015].

[10] S. Wu and Q. Sun, “Computer simulation of leadership, consensus decision making and collective behaviour in humans,” *PloS one*, vol. 9, no. 1, 2014.

[11] A. C. Bicharra, N. Sánchez-Pi, L. Correia, and J. M. Molina, “Multi-agent simulations for emergency situations in an airport scenario,” *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 1, no. 3, pp. 69–73, 2013.

[12] T. Bosse, M. Hoogendoorn, M. C. Klein, J. Treur, C. N. Van Der Wal, and A. Van Wissen, “Modelling collective decision making in groups and crowds: Integrating social contagion and interacting emotions, beliefs and intentions,” *Autonomous Agents and Multi-Agent Systems*, vol. 27, no. 1, pp. 52–84, 2013.

[13] R. Hocevar, F. Marson, V. Cassol, H. Braun, R. Bidarra, and S. R. Musse, “From their environment to their behavior: a procedural approach to model groups of virtual agents,” in *Intelligent Virtual Agents*. Springer, 2012, pp. 370–376.

[14] E. PUYBARET, “Sweet Home 3D,” http://www.sweethome3d.com/, 2005, [Online; accessed Mar. 2015].

[15] S. Luke, C. Cioffi-Revilla, L. Panait, K. Sullivan, and G. Balan, “Mason: A multiagent simulation environment,” *Simulation*, vol. 81, no. 7, pp. 517–527, 2005.

[16] J. Gemrot, R. Kadlec, M. Bída, O. Burkert, R. Píbil, J. Havlíček, L. Zemčák, J. Šimlovič, R. Vansa, M. Štolba *et al.*, “Pogamut 3 can assist developers in building ai (not only) for their videogame agents,” *Agents for Games and Simulations*, pp. 1–15, 2009.

[17] C. W. Reynolds, “Steering behaviors for autonomous characters,” pp. 763–782, 1999.

[18] J. J. Bryson, “Intelligence by design: principles of modularity and coordination for engineering complex adaptive agents,” 2001.

[19] R. A. Brooks, “A robust layered control system for a mobile robot,” *Robotics and Automation, IEEE Journal of*, vol. 2, no. 1, pp. 14–23, 1986.

[20] R. A. Brooks., “Intelligence without reason,” *The artificial life route to artificial intelligence: Building embodied, situated agents*, pp. 25–81, 1995.

DOI: 10.15439/2015F116
ACSIS, Vol. 5

# Simulation as a Service: A Design Approach for large-scale Energy Network Simulations

Thomas Preisler, Tim Dethlefs and Wolfgang Renz
Faculty of Engineering and Computer Science,
Hamburg University of Applied Sciences,
Berliner Tor 7, 20099 Hamburg, Germany
Email: {thomas.preisler,tim.dethlefs,wolfgang.renz}@haw-hamburg.de

*Abstract*—**In the ongoing GEWISS project it is planned to implement a geographical heat information and simulation system. It shall provide a planning and simulation tool for the interlinking of urban development and district heating network development to support the political decision making process in the City of Hamburg. The system shall combine macroscopic and microscopic simulations to a co-simulation system. The simulation as a service approach is presented as a loosely-coupled scalable solution to realize large-scale energy network simulations. It is based on cloud computing technologies for the optimal utilization of computing resources in heterogeneous simulation-infrastructures. This approach can be used to realize simulation systems integrating Multi-Agent System (MAS) based simulations and other simulation technologies. For practical evaluation, two implementation approaches based on a MAS platform as a service-oriented solution will be presented and compared to an approach involving standard web-service technologies.**

## I. Introduction

SIMULATIONS take a significant stake in finding intelligent solutions and concepts for future Smart Energy Networks. They are vital to test the suitability of such concepts and solutions and therefore, needed before an actual realization can take place. This regards today's energy grid that undergoes a structural change towards the so-called *Smart Grid* as well as other energy networks as e.g., *heat supply systems*. Besides the realization of control and coordination strategies for the Smart Grid, the uniformed planning of both urban development and energy network development is essential. The holistic contemplation of such a *Smart City* requires the combination of different simulation concepts. Regardless of the domain, it is essential to test operational concepts with respect to their cost effectiveness and to optimize them if necessary [1], [2]. Contrasting to the test of operational concepts, which often requires fine-granular simulations with consideration of microscopic aspects, in terms of planning questions for urban and energy network development macroscopic simulations are of concerns.

The *simulation as a service* approach presented in this paper is related to the *GEWISS* project [3]. The goal of this project is to develop a geographical information and simulation system in order to support the political decision making in terms of an interlinking between urban development and district heating network development. It shall combine macroscopic modeling approaches from urban development regarding the evolution of building structures in a quarter, with microscopic simulations of the heat energy demand of buildings respectively building blocks. A multi-agent based co-simulation system is envisioned, where the microscopic behavior of single buildings or building blocks will be modeled as agents and the selection of appropriate sub-simulations will be mapped to agent-based goal deliberation. The latter will not be part of this paper. This paper focuses on presenting a possible approach to support large-scale co-simulations of macroscopic and microscopic simulation components, that can be rolled out to a heterogeneous IT-infrastructure. The goal of the proposed approach is to utilize cloud computing technologies and service-oriented design concepts in order to build loosely-coupled simulation systems, that integrate different simulation types and models in a service-oriented fashion. Cloud computing technologies and approaches should be used to form a private cloud on heterogeneous university IT-infrastructure, consisting out of different types of servers, in order to utilize the available computation resource optimally and to integrate slow but cheap computers like, e.g., *Raspberry Pi* single-board computer meaningfully.

The remainder of this paper is structured as follows. Section II describes current cloud computing infrastructures and related simulation frameworks and approaches as well as a related approach in the domain of modeling and simulating the heat energy demand of buildings. In Section III the GEWISS project will be described in more detail before the concept of simulation as a service will be described in Section IV. Section V describes and discusses possible implementation alternatives and challenges, followed by Section VI where the implementation alternatives will be evaluated in terms of their scalability. Finally, Section VII concludes the paper.

## II. Related Work

This Section introduces examples for state of the art cloud computing technologies, and describes related simulation approaches and frameworks, as well as related work in terms of the *GEWISS* Project.

### A. Cloud Computing Technologies

Cloud computing [4] is seen as a new approach to IT infrastructure management, facilitating a pay-as-to-go usage model. Computational resources are made available on a demand-driven basis instead of statically dedicated physical systems.

978-83-60810-66-8/$25.00©2015, IEEE

Therefore, the approach minimizes idle times and optimizes the utilization of resources, which leads to a minimization of resource dissipation. The authors of [5] bring forward the argument, that by adapting cloud computing ideas for optimal resource utilization, it is reasonable that existing computers in a company or university network could contribute their spare resources in a private cloud. This approach is picked up to propose the *simulation as a service* concept to realize large-scale simulations as required in the *GEWISS* project, that scale well and support optimal utilization of a heterogeneous IT-infrastructure. Cloud applications can be built on the *Infrastructure as a Service* (IaaS) or the *Platform as a Service* (PaaS) layer. On the IaaS layer, access to the cloud is granted by virtual machines that allow fine-grained control of the software stack and provide low-level aspects like operating systems. On the PaaS layer, a cloud operator establishes a new software layer with a dedicated middleware programming interface, and thus, lower level details are abstracted. PaaS facilitates the development of applications on top of the given platform, but restricts the types of applications to those which are supported by the platform. The *Software as a Service* (SaaS) layer are user-ready applications running in the cloud, which are typically built upon the IaaS or PaaS layer. There are many different approaches and technologies for the implementation and operation of cloud applications. This Section will focus on two different PaaS solutions and describe them exemplary. PaaS solutions are more suitable for the realization of simulations as a service than IaaS approaches, as they abstract from the operating system level and offer an Application Programming Interface (API) for the realization of scalable applications on top of a cloud infrastructure. The *Mesos* platform [6] offers a thin resource sharing layer, that enables fine-grained sharing across diverse cluster computing frameworks by providing a common interface for accessing cluster resources. *Mesos* focuses on providing shared commodity clusters between diverse cluster programming frameworks like *Hadoop* [7] or *MPI* [8] to improve cluster utilization. Resources are shared in *Mesos* to allow frameworks to achieve data locality by taking turns in reading data stored on each machine. It introduces a distributed two-level scheduling mechanism called *resource offers*. A resource offer encapsulates a bundle of resources that a framework can allocate on a cluster node to run tasks. *Mesos* decides how many resources it offers each framework, while the frameworks decide which resources they accept and which computations are executed. *Mesos* delegates the control over the scheduling to the frameworks. This decentralized scheduling model may not always result in globally optimal scheduling, but according to [6] it performs well in practice and allows the frameworks to meet goals such as data locality near perfectly. While *Mesos* focuses on resource-sharing in datacenters, *JadexCloud* [5] is a PaaS infrastructure to develop, deploy and manage distributed applications with a strong focus on distribution transparency. It is based on a Multi-Agent Systems (MAS) platform and allows to build cloud-based agent applications with service-oriented communication. The key concept is a three layered model, that helps separating responsibilities and managing complexity. A daemon layer provides a node infrastructure for managing cloud resources. It automatically detects and announces nodes joining and leaving the network. The platform layer on top supports application related management tasks including the automatic deployment of application artifacts on different nodes as well as starting and stopping of components. Finally, the application layer facilitates the application development by providing the API and tools for debugging. The current version of the provided PaaS infrastructure [9] supports the management of non-functional requirements. It allows to define non-functional properties for services that are monitored automatically, e.g., a wait-queue property that counts the request currently waiting to be answered. Based on this non-functional properties the service selection can be automated by defining non-functional requirements like, e.g., finding the service with the smallest wait-queue. Due to its focus on cloud-components offering services (agents), and the automatic deployment of application artifacts to dynamically joining and leaving network nodes, *JadexCloud* seems to be well suited to realize the simulation as a service approach.

### *B. Simulation and Modeling as a Service Approaches*

In [10] it is described how simulation software can be accessed as service (SimSaaS) in a service-oriented architecture (SOA) approach. It is stated that carrying out an experiment can be achieved by connecting multiple simulation services to form a work-flow which represents how the experiment proceeds. The authors of [10] state that simulation services differ from regular web services as concepts of time and state are essential, thus, viewing them as stateful services, that treat each service request as a series of dependent transactions that are related to the previous requests as well as the model's current state. Therefore, different communication and synchronization paradigms are described and assessed in terms of their suitability to interconnect stateful simulation services. The approach presented in this paper adopts the ideas from [10] by combining them to a design approach for large-scale energy network simulation with the extension that stateless simulation services may also be suitable for the realization and integration of small sub-simulation models. A survey about current trends and technologies in the field of modeling and simulation as a cloud service is given in [11]. It defines the term Modeling and Simulation as a Service (MSaaS) as a model for provisioning modeling and simulation (M&S) services on demand from a cloud service provider (CSP), that keeps the underlying infrastructure, platform, and software requirements and details hidden from the user. In this case the CSP is responsible for licenses, software upgrades, scaling the infrastructure, and providing grade of service and quality of service as specified in service level agreements. The SaaS model provides access to hosted applications in a cloud environment, allowing users to access services at low cost and scale as needed. This model was extended in [12] to include high-performance hosted simulation and modeling services. The described *Polymer Portal* is a first-generation simulation and modeling as a service (SMaaS) platform. Contrasting to the approach presented in this paper, where the simulation as a service approach is presented as a modeling approach to build distributed, scalable simulation systems, the approach presented in [12] focuses on ready to use simulation- and modeling-services that users can access for a fee.

### C. *Simulation Frameworks and Approaches*

There exist many different tools and frameworks that developers can use to built simulations upon. Some of these simulation frameworks focus on specific application domains and thus, support the developers by providing domain specific models and tools, while other frameworks focus on a broader application-domain independent scope and support simulations in general. While application-domain specific frameworks take some of the implementation work from the developer and often allow a faster implementation, they restrict developers in terms of implementational freedom and limit them, if their problem does not fit into the domain specific scope of the framework. Examples for domain specific simulation frameworks are, e.g., *Mosaik*, a flexible Smart Grid co-simulation framework [13] and *RinSim*, a simulator for logistics problems [14]. An example for a general simulation framework is the *Jadex* project [15]. When it comes to providing simulation functionality as a service there a two practicable ways. The first one is to implement the simulation using an application-domain dependent or independent simulation framework (or of course implementing it completely from scratch). In this case, the implemented simulation system will be considered as a black box with defined input and output parameters when it is encapsulated by a service component, that provides the simulation functionality as a service method. An example for this implementation scheme is given in [16] where 4GL Matlab simulation models are encapsulated by service components, to provide their functionality via REST (Representational State Transfer) web-services. The other implementational alternative is to implement the whole simulation with service-based technology. In this case it is possible to provide both, the macroscopic functionality of the simulation itself, as well as the microscopic functionality of single simulation components out of the box, without the need to encapsulate it first. An example for such a simulation system is given in [17], where a Multi-Agent based self-healing resource-flow system was built that used services to provide the functionality of the simulated robots. The standardization of simulation interoperability resulted in the *High Level Architecture* (HLA), an IEEE standard for modeling and simulation [18]. The HLA is a technical architecture developed to facilitate the reuse and interoperability of different simulation systems and assets. It provides a general framework, which can be used by developers to structure and describe their simulation systems and to interoperate them with other simulation systems. But due to its complexity it is hardly used outside the military domain [19].

### D. *Modeling Urban Energy and Heat Demand*

In terms of modeling the urban energy and heat demand the approach presented in [20] and [21] is related to the undertaking in the *GEWISS* project. Both projects aim at modeling the energy respectively heat demand of urban buildings and building structures. But while the *GEWISS* project will follow a Multi-Agent based simulation approach, modeling the building and building structures as microscopic agents the approach presented in [20] and [21] uses 3D city models to calculate the heat demand of whole district areas. It uses the OGC Standard *CityGML* [22], an open, multi-functional model that can be used for geospatial transactions, data storage, and database modeling, for the modeling of 3D buildings. Based on the surface of the 3D models of whole building blocks their energy and heat demand is calculated as a macroscopic approach.

## III. Introducing the GEWISS Project

The requirements for the simulation as a service approach are derived from the *GEWISS* project. In this project a geographical heat information and simulation system will be developed to support the analysis of heat demand and urban development for the City of Hamburg (Germany). The goal is the development of a simulation tool (framework) that supports the interlinking of urban development and district heat network development by providing a planning tool for different scenarios.

### A. *Project Description*

In order to utilize the potential of climate protection and to allocate available resource with high cost-effectiveness, the strategic heat planing must be interlinked with urban development projections (and vice versa). In this case, an abstraction of the spatial location of existing buildings is for obvious reasons not possible. Aspects such as the conversion of urban areas, infill, redevelopment or demolition as well as renovation of buildings or building groups should be tailored to the local available heat sources. Conversely, this means that the grid-based heat supply should be planned with respect to existing and future building stock. For this indentation it is necessary that data will be collected and analyzed with respect to spatial location. The combination of data and analysis to the heat demand and urban development should take place in a geographic information system (GIS). The goal of the *GEWISS* project is to extend such a GIS with geographic heat information and simulation capabilities in order to provide planning assistance for future developments.

Therefore, a framework should be develop that allows users to simulate the medium and long-term development of a heat supply system. Here, the specification of both external conditions (e.g., the development of energy prices) as well as municipal political measures (e.g., regulations or financial incentives for certain structural measures) should be supported and also the analysis of their effectiveness. The goal hereby is to explore business models as well as different urban development plans in a systematic fashion. This should help the City of Hamburg to find measures to achieve an optimal solution to match environmental protection objectives in the heating sector. From the overall project goal a set of scientific and technical working tasks are derived, which are relevant for the functional and non-functional requirements of the simulation and information system to be built:

- The strategic conception of heat planning and urban development should be integrated, so that energy and emission reductions can be implemented efficiently.
- All areas of a heating system (generation, transmission, storage and demand) should be considered through a participatory modeling and simulation approach that involves all relevant stakeholders.
- As a core, an agent-based simulation with visualization capabilities should be realized.

### B. Simulation Requirements

A series of functional and non-functional requirements unfolds from the mentioned project goals. For example, the design of the simulation entities must consider all areas of the heating system, taking into account production, transport, storage and demand. This means, that the simulation tool must be able to support large-scale simulations with more than 100,000 simulation entities (agents). The agents that are designed to model buildings or building blocks may be required to support a rule-based temporal evolution. The considerations of important stakeholders implies on the one hand influences on the design of simulation scenarios, but allows on the other hand the scenario-dependent modeling of rule-based descriptions for the automatic simulation through autonomous agents. The validation of these descriptions is to be ensured against separate micro-simulations. In order to fulfill this requirements a simulation tool that supports the realization of co-simulations based on a service-oriented middleware is envisioned. Co-simulation is a prominent method to solve multi-physic problems. Such simulations combine well-established and specialized simulation tools for different fields [23]. In the context of the *GEWISS* project a co-simulation is distinguished by a distributed simulation as well as a distributed modeling (defined by the usage of different modeling tools). A distributed simulation in this context is understood as the integration of multiple, self-contained simulations. Additional information about the strong impact of non-functional requirements on large-scale systems can be found in [24].

## IV. The Concept of Simulation as a Service

The concept of the simulation as a service approach derives from the simulation requirements of the *GEWISS* project. The goal is to use cloud computing ideas and technologies to built loosely-coupled large-scale simulation systems. By adapting the SaaS concept to the simulation domain, simulations will be provided as services in a cloud infrastructure. The conceptual architecture and approach is depicted in Fig. 1. The idea is that simulations will be contained within a Simulation Service Component. This component encapsulates the functionality of the simulation and provides a service interface to execute it in a service-oriented fashion. This approach unifies the invocation of different simulation types while also allowing the combination of different implementation models and programming languages. Fig. 1 exemplifies two different simulation types that can be combined and invoked in an unified way following this approach.

A common modeling technique in the engineering domain is the usage of fourth generation languages (4GL). These are programming languages focusing on rapid application development. The expression was coined by [25]. Recently, they have gained new attention through the introduction of model based software development [26]. Examples for 4GL are MATLAB, Simulink, Modelica, GAMS (General Algebraic Modeling System). These languages are used in many research projects for rapid prototyping as well as the realization of models or algorithms, e.g., for demand side management and the integration of regenerative energy resources into smart grids [27]. Although they reduce the overall development effort through the usage of comprehensible application-oriented paradigms, the integration of such 4GL models into large-scale simulations is a considerable challenge [16]. Encapsulating their functionality by a Service Simulation Component allows to execute them in an unified way, while also enabling their execution in a scalable cloud infrastructure.

The other depict simulation type is a MAS based simulation. They are well-known and widely spread [28]. Here, different types of agents are used to model the reality and to examine different problems. Such simulations exhibit different grades of complexity and inter-simulation dependencies. If a simple MAS based simulation that is executed on a single network node is considered, it can easily be encapsulated by a Simulation Service Component as a whole. This will provide an unified service-oriented way of invocation and control the execution and the life-cycle of the simulation as a whole. If a more complex, distributed MAS based simulation is considered, a possible integration approach is to encapsulate a dedicated starting component with a Simulation Service Component, which starts the distributed application and collects the results to return them.

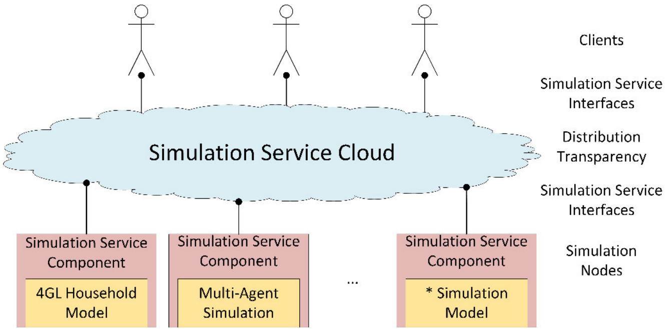

Clients
Simulation Service Interfaces
Simulation Service Cloud
Distribution Transparency
Simulation Service Interfaces
Simulation Service Component
4GL Household Model
Simulation Service Component
Multi-Agent Simulation
...
Simulation Service Component
* Simulation Model
Simulation Nodes

Fig. 1. Simulation as a Service Conceptual Architecture and Approach

Simulations that are encapsulated by a Simulation Service Component can either by called by clients directly if they require a specific simulation functionality, or more complex simulations can be composed out of multiple simulation services. In this case a simulation component can use the services offered by other simulation components, if they require the results provided by these simulation components as a sub-simulation. An example for this use-case is the realization of a mesoscopic co-simulation where a macroscopic simulation requires results from microscopic simulation components. This will be the case in the *GEWISS* project, were questions regarding the development of district heat networks shall be examined with regards to macroscopic simulations regarding urban development and microscopic simulations regarding the heat demand of households. If the composition of different simulation services is of concern, approaches like the Web Service Business Process Execution Languages (WS-BPEL) respectively the Business Process Model and Notation (BPMN) for the orchestration of web-services might be adoptable to the simulation domain.

Basically there are two main paradigms for the composition of simulation services. Fig. 2 depicts the hierarchical composition and execution of simulations as an UML sequence diagram. It illustrates the execution of a simulation $A$. During its execution, the simulation requests services of the sub-simulations $B$ and $C$, while sub-simulation $C$ requires the results from sub-simulation $D$ in order to answer the request from Simulation $A$. The hierarchical structure depict in Fig. 2 resembles the envisioned structure of the simulation system in the *GEWISS* project, where a main simulation (macroscopic) requires results from microscopic sub-simulations. Thereby, the sub-simulations are only active when their simulation services are called by the main simulation. The simulated time in this case can either be managed by the encapsulating root simulation and passed on to the sub-simulations or each sub-simulation can handle it's own time model, which allows to model interacting simulation with different time-scales (microscopic and macroscopic).

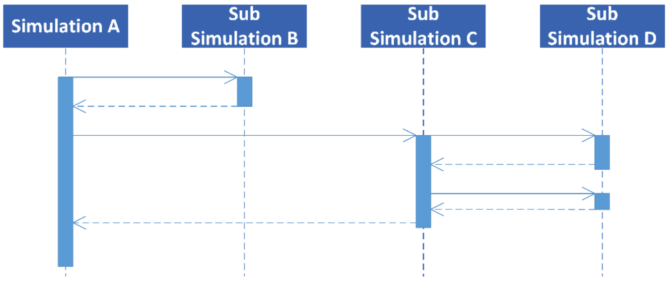


Fig. 2. UML sequence diagram: Hierarchical composition and execution of simulation services.

A different composition paradigm is depict in Fig. 3 where two simulations are executed parallel while exchanging data. This paradigm resemble the one described in [10] where it is stated that simulation services differ from web services as they have to be considered as stateful services, that treat each service request as a series of dependent transactions that are related to the previous requests as well as the model's current state. Regarding this composition paradigm of interlinked simulations it has to be ensured that their time model is synchronized, so that events are processed in a correct order.

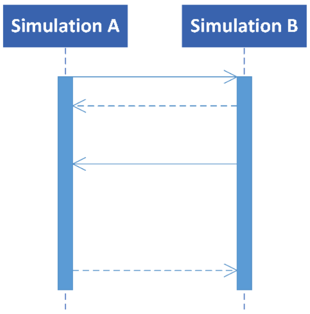


Fig. 3. UML sequence diagram: Parallel composition and execution of simulation services.

## V. Implementation Alternatives and Challenges

In general it is reasonable to facilitate established service-oriented technologies to provide simulations as a service. If distribution transparency, scalability and robustness properties should be adopted for the provided simulation services, it is reasonable to facilitate established cloud computing technologies as well. If the simulation is considered as a black box and not implemented with explicit cloud distribution in mind it might be more feasible to use the IaaS approach to abstract from the physical hardware in order to achieve the previous mentioned cloud computing properties. In this case the IaaS approach excels the PaaS approach, because the later one requires the application to be built on top of API of the according platform. Although it is also possible to encapsulate a black box simulation with a component regarding the API of the PaaS platform, but in this case the simulation does not profit from the distribution, scalability or robustness properties of the PaaS platform. So if an already existing simulation system should be provided as a cloud simulation service, a solution built upon a IaaS platform is advantageous. If the simulation is built from scratch or the simulation model is a simple one with a small footprint, it is beneficial to built it against the PaaS API respectively to encapsulate the simple model with a component that is built against the API.

The next Section will provide an evaluation of three different implementation alternatives and evaluate them in terms of their scalability. A case study was ventured, in order to show how a simulation as a service component could be built based on different technologies. A 4GL-model simulating the energy demand of a single household equipped with an electrical heater was used for this case study. In the following, the three implementation alternatives are described briefly before they are evaluated in the next Section.

**REST Webservices:** The first implementation follows the approach presented in [16] and encapsulated the 4GL-model with a REST web-service. In order to provide scalability, a load balancer was connected upstream to distribute the request based on a round-robin scheduling mechanism among multiple simulation nodes.

**Jadex NFP Service Selection:** The second implementation facilitated the approach for elastic component-based cloud applications presented in [9]. This approach introduces the concept of service selection based on non-functional properties. Again the 4GL-model is encapsulated by a service component. In this case the service is equipped with a non-functional property observer that observes the length of the wait queue of this service. If the simulation service is going to be requested by a client, a non-functional property evaluator evaluates the mentioned waiting queue size property and selects the best service (lowest number of waiting clients). Thus, an optimal distribution of client calls is achieved.

**JadexCloud Service Pool:** The third implementation uses the *JadexCloud* PaaS middleware [5]. Like other PaaS solutions it simplifies the construction of a cloud application by providing common base abstractions, tool sets and an API. The *JadexCloud* infrastructure has two main advantages for the development and deployment of distributed application. Any service offered by a component can be made available in a platform-controlled service pool. The utilization of the service pool deliberates the developer from starting and handling the life-cycle of the service-components himself. The service pool will start new service-components automatically if new resources are needed to handle incoming requests. Therefore, the service pool acts as a proxy handling all incoming service-requests and forwards them to service-components with free resources. The usage of the service pool, does not only simplifies the development of scalable services but also supports the dynamical deployment of service-components to remote platforms. If the service pool detects new *JadexCloud* platforms in the network range, it automatically deploys the governed service-components on this platforms and starts new instances on the remote platform, in case such are required to handle incoming service requests. So following this implementation scheme, the service component encapsulating the 4GL-model was equipped with a service pool to handle the automatic deployment on a cluster of *JadexCloud* nodes and to handle the life-cycle and load-balancing in order to achieve scalability.

## VI. Simulation and Evaluation of Implementation Alternatives

In order to evaluate the three proposed implementation alternatives in terms of their scalability three implementation prototypes were realized. Each of the three simulation setups will be described briefly before the results of the scalability analysis will be discussed. The simulation setup for the implementation based on REST web-services (1) is depicted in Fig. 4. Similar in all three setups is a client (a HTTP web-service client in this case) that places $n$ parallel requests. For this example, the 4GL-model simulating the energy demand of a single household is encapsulated by a REST web-service. The service is deployed to five homogeneous workstations running *Apache Tomcat 7*[1] as an application container. On one of these workstations the *Apache HTTP Server*[2] equipped with the *mod_jk*[3] Tomcat connector is used as a load-balancer. It uses a round-robin scheduling mechanism to distribute the calls to the simulation services in an equal way. For the purpose of testing the scalability and distribution properties only one service-component was deployed on each of the network nodes.

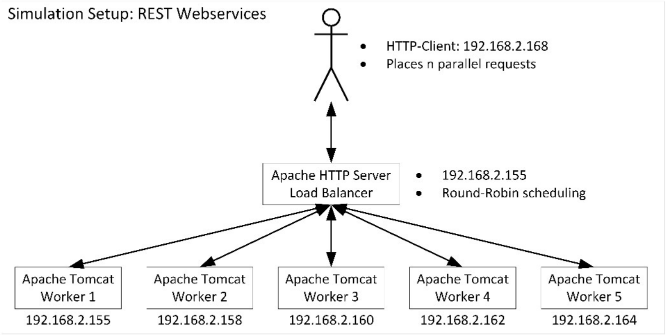


Fig. 4. Simulation Setup: REST Webservices.

The simulation setup for the service selection based on non-functional requirements (2) is depicted in Fig. 5. Again a client places $n$ parallel service calls. In this setup the simulation service encapsulating the 4GL-model is realized as a *Jadex* service and is equipped with a non-functional property monitor, that monitors the size of the service's wait-queue. For each of the service calls, the client queries the service monitors and uses a non-functional property evaluator to evaluate them with regards to quality criteria. The best service (lowest number of waiting requests) is selected and then called directly. Thereby, an optimal degree of capacity utilization is achieved as it is ensured that always the service with the lowest capacity utilization is ensured.

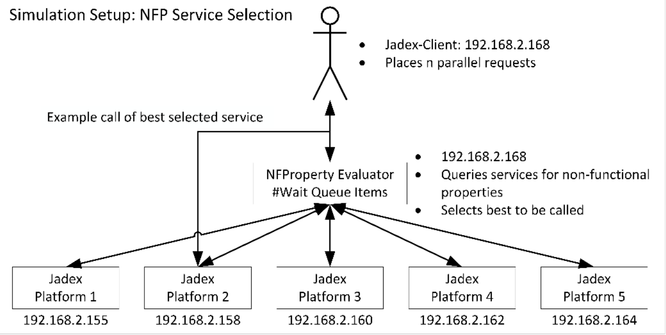

Simulation Setup: NFP Service Selection
Jadex-Client: 192.168.2.168
Places n parallel requests
Example call of best selected service
NFProperty Evaluator
#Wait Queue Items
192.168.2.168
Queries services for non-functional properties
Selects best to be called
Jadex Platform 1
192.168.2.155
Jadex Platform 2
192.168.2.158
Jadex Platform 3
192.168.2.160
Jadex Platform 4
192.168.2.162
Jadex Platform 5
192.168.2.164

Fig. 5. Simulation Setup: NFP Service Selection.

The simulation setup for the utilization of a service pool and the *JadexCloud* infrastructure (3) is depicted in Fig. 6. It uses the same client and simulation service component as described before, but contrasting to the previous described service selection based on non-functional properties, a global

[1] http://tomcat.apache.org/ (accessed September 25, 2015)

[2] http://httpd.apache.org/ (accessed September 25, 2015)

[3] http://tomcat.apache.org/connectors-doc/ (accessed September 25, 2015)

service pool is used to act as a proxy between the actual services and the client. The service pool deploys the service code automatically to all five workstations and handles the execution of one service-component on each workstation. Due to comparability reasons with the two other setups, only one service-component was started on each of the *JadexCloud* platforms, although it is possible to configure the service pool accordingly to make use of more components. Comparable to the first simulation setup the service pool use a round-robin based scheduling mechanism to distribute the service requests equally to all pooled services. The service distribution is completely transparent for the clients, as they only interact with the proxy.

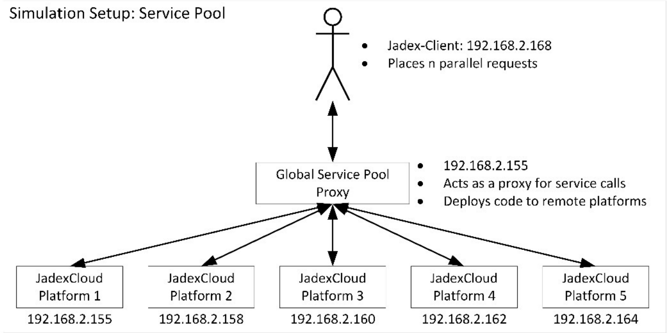


Fig. 6. Simulation Setup: Service Pool.

In order to analyze the scalability properties of the three presented implementation alternatives a number of experiments were conducted. The results of this analysis are shown in Fig. 7. The x-axis depicts the number of parallel calls on a logarithmic scale and the time in milliseconds it took to serve all the parallel requests is depicted on the y-axis (logarithmic scale as well). In order to dampen spikes for each number $n$ of parallel calls 10 experiments were ventured and the mean values were used for the evaluation. For $n = 10, 100, 1000$ parallel calls the three implementation alternatives perform quite similar and all three show a linear scalability. This behavior differs for $n = 10000$ parallel service requests. The implementation utilizing the non-functional property based service selection still scales linear. But both, the implementation using REST web-services and the *JadexCloud* service pool scale considerably worst. Therefore further experiments for these two implementations were conducted. As shown in the graph for $n = 2000, 3000, 4000$ parallel calls they still perform nearly linear. When the number of parallel calls reaches $n = 5000$ they lose this scalable behavior. We assume that this deviation between the three approaches is because the according load balancer (Apache HTTP Server or Jadex Service Pool) became saturated. If a large amount of parallel calls has to be processed by a load balancer it may become overloaded. Therefore, the load balancer itself could be the bottleneck of the system. This behavior differs from the service selection based on non-functional properties, as in this case the client itself evaluates the degree of capacity utilization of the offered services and selects the best. Therefore, the required distribution effort is shifted to the client. In order to further investigate this behavior, future work will analyze the scalability behavior of the three implementation alternatives on a System-on-a-Chip cluster. This cluster is characterized by a larger amount of machines with less computational power each in comparison to the workstation cluster used in this experiment. Thus, the scalability and load-balancing properties of the implementation alternatives will become more relevant.Also a combination of the non-functional property selection approach with the service pool approach is envisioned. Even if the scalability properties of the non-functional property selection approach excels the ones from the service pool approach, the dynamically deployment properties of the service pool approach are of great value, especially if an application is executed on dynamic infrastructure with joining and leaving network nodes. Therefore, a combination is envisioned where the default round-robing scheduling mechanism of the service pool is going to be replaced by one selecting the service based on non-functional properties like, e.g., the size of the wait-queue.

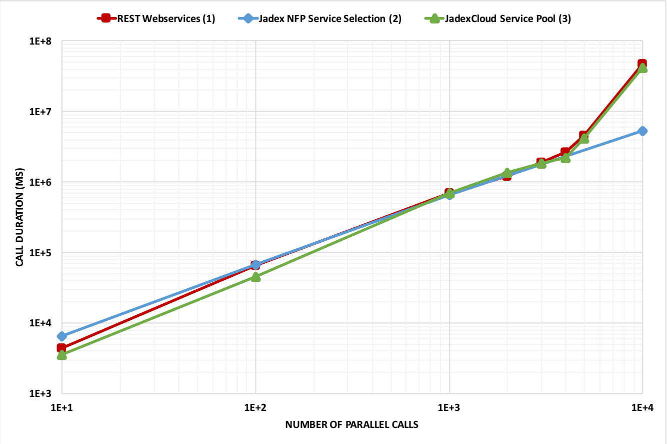


Fig. 7. Results of the Scalability Analysis.

## VII. Conclusion

In this paper we presented the concept of simulation as a service as a promising approach for the realization of large-scale smart-energy network simulations. It was described how the approach will be used to implement a geographical heat information and simulation system in the *GEWISS* project. The goal is to provide a planning and simulation tool for the interlinking of urban development and district heat network development in order to support the political decision making process in the City of Hamburg. The envisioned system will combine macroscopic and microscopic simulations to a co-simulation system. The simulation as a service approach was presented as loosely-coupled and scalable solution to realize a large-scale simulation system based on cloud computing technology in order to optimal utilize the computing resources of a heterogeneous university IT-infrastructure. Three different implementation alternatives were described and evaluated in terms of their scalability. The results encouraged the usage of the *Jadex* PaaS platform. A MAS platform that allows to built agent-based cloud application with service-oriented communication and supports the execution and deployment with service pools, service selection based on non-functional requirements and the automatic deployment of components (agents) and services to fluctuating network nodes.

## ACKNOWLEDGMENT

We would like to thank Lars Braubach and Alexander Pokahr for their input and support while implementing the simulation prototypes based on the *Jadex* platform, as well as our partners in the *GEWISS* project.

## REFERENCES

[1] T. Logenthiran, D. Srinivasan, and T. Z. Shun, "Demand side management in smart grid using heuristic optimization," *Smart Grid, IEEE Transactions on*, vol. 3, no. 3, pp. 1244–1252, Sept 2012. doi: 10.1109/TSG.2012.2195686

[2] D. Rivola, A. Giusti, M. Salani, A. Rizzoli, R. Rudel, and L. Gambardella, "A decentralized approach to demand side load management: The swiss2grid project," in *Industrial Electronics Society, IECON 2013 - 39th Annual Conference of the IEEE*, Nov 2013. doi: 10.1109/IECON.2013.6699895. ISSN 1553-572X pp. 4704–4709.

[3] T. Preisler, G. Balthasar, T. Dethlefs, and W. Renz, "Servicekomponenten-basierte architektur für mikroskopische und makroskopische simulation der städtischen energieversorgung," in *Proceedings of the German VDE-Congress*, ser. Smart Cities, 2014.

[4] B. Sosinsky, *Cloud computing bible*. John Wiley & Sons, 2010.

[5] L. Braubach, A. Pokahr, and K. Jander, "Jadexcloud - an infrastructure for enterprise cloud applications," in *Multiagent System Technologies*, ser. Lecture Notes in Computer Science, F. Klügl and S. Ossowski, Eds. Springer Berlin Heidelberg, 2011, vol. 6973, pp. 3–15. ISBN 978-3-642-24602-9. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-24603-6_3

[6] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. Katz, S. Shenker, and I. Stoica, "Mesos: A platform for fine-grained resource sharing in the data center," in *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'11. Berkeley, CA, USA: USENIX Association, 2011, pp. 295–308. [Online]. Available: http://dl.acm.org/citation.cfm?id=1972457.1972488

[7] A. Abouzeid, K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, and A. Rasin, "Hadoopdb: An architectural hybrid of mapreduce and dbms technologies for analytical workloads," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 922–933, Aug. 2009. doi: 10.14778/1687627.1687731. [Online]. Available: http://dx.doi.org/10.14778/1687627.1687731

[8] W. Gropp, E. Lusk, and A. Skjellum, *Using MPI: portable parallel programming with the message-passing interface*. MIT press, 1999, vol. 1.

[9] L. Braubach, K. Jander, and A. Pokahr, "A middleware for managing non-functional requirements in cloud paas," in *Cloud and Autonomic Computing (ICCAC), 2014 International Conference on*, Sept 2014. doi: 10.1109/ICCAC.2014.32 pp. 83–92.

[10] S. Guo, F. Bai, and X. Hu, "Simulation software as a service and service-oriented simulation experiment," in *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*, Aug 2011. doi: 10.1109/IRI.2011.6009531 pp. 113–116.

[11] E. Cayirci, "Modeling and simulation as a cloud service: A survey," in *Simulation Conference (WSC), 2013 Winter*, Dec 2013. doi: 10.1109/WSC.2013.6721436 pp. 389–400.

[12] T. Bitterman, P. Calyam, A. Berryman, D. E. Hudak, L. Li, A. Chalker, S. Gordon, D. Zhang, D. Cai, C. Lee *et al.*, "Simulation as a service (smaas): a cloud-based framework to support the educational use of scientific software," *International Journal of Cloud Computing*, vol. 3, no. 2, pp. 177–190, 2014.

[13] S. Schütte, S. Scherfke, and M. Sonnenschein, "Mosaik - smart grid simulation API - toward a semantic based standard for interchanging smart grid simulations," in *SMARTGREENS 2012 - Proceedings of the 1st International Conference on Smart Grids and Green IT Systems, Porto, Portugal, 19 - 20 April, 2012*, B. Donnellan, J. A. P. Lopes, J. F. Martins, and J. Filipe, Eds. SciTePress, 2012. ISBN 978-989-8565-09-9 pp. 14–24.

[14] R. R. van Lon and T. Holvoet, "Rinsim: a simulator for collective adaptive systems in transportation and logistics," in *2012 IEEE Sixth International Conference on Self-Adaptive and Self-Organizing Systems*. IEEE, September 2012. doi: 10.1109/SASO.2012.41 pp. 231–232. [Online]. Available: https://lirias.kuleuven.be/handle/123456789/361419

[15] L. Braubach and A. Pokahr, "The jadex project: Simulation," in *Multiagent Systems and Applications*, ser. Intelligent Systems Reference Library, M. Ganzha and L. C. Jain, Eds. Springer Berlin Heidelberg, 2013, vol. 45, pp. 107–128. ISBN 978-3-642-33322-4. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33323-1_5

[16] T. Preisler, G. Balthasar, T. Dethlefs, and W. Renz, "Scalable integration of 4gl-models and algorithms for massive smart grid simulations and applications," in *28th International Conference on Informatics for Environmental Protection: ICT for Energy Effieciency, EnviroInfo 2014, Oldenburg, Germany, September 10-12, 2014.*, J. M. Gómez, M. Sonnenschein, U. Vogel, A. Winter, B. Rapp, and N. Giesen, Eds. BIS-Verlag, 2014. ISBN 978-3-8142-2317-9 pp. 341–348. [Online]. Available: http://www.enviroinfo2014.org/

[17] T. Preisler and W. Renz, "Scalability and robustness analysis of a multi-agent based self-healing resource-flow system," in *Federated Conference on Computer Science and Information Systems - FedCSIS 2012, Wroclaw, Poland, 9-12 September 2012, Proceedings*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2012. ISBN 978-83-60810-51-4 pp. 1261–1268. [Online]. Available: https://fedcsis.org/proceedings/2012/pliks/194.pdf

[18] S. I. S. Committee *et al.*, "of the ieee computer society. ieee standard for modeling and simulation (m&s) high level architecture (hla)-ieee std 1516-2000, 1516.1-2000, 1516.2-2000. new york: Institute of electrical and electronics engineers," *Inc., New York*, 2000.

[19] C. A. Boer, A. de Bruin, and A. Verbraeck, "Distributed simulation in industry-a survey part 3-the hla standard in industry," in *Simulation Conference, 2008. WSC 2008. Winter*. IEEE, 2008, pp. 1094–1102.

[20] R. Nouvel, C. Schulte, U. Eicker, D. Pietruschka, and V. Coors, "Citygml-based 3d city model for energy diagnostics and urban energy policy support," in *Proceedings of the 13th conference of international Building Performance Simulation Association*, 2013, pp. 218–25.

[21] R. Nouvel, M. Zirak, H. Dastageeri, V. Coors, and U. Eicker, "Urban energy analysis based on 3d city model for national scale applications," in *Presented at the IBPSA Germany Conference*, vol. 8, 2014.

[22] G. Gröger, T. H. Kolbe, A. Czerwinski, and C. Nagel, "Opengis® city geography markup language (citygml) encoding standard. version: 1.0. 0, ogc 08-007r1," 2012.

[23] S. Sicklinger, V. Belsky, B. Engelmann, H. Elmqvist, H. Olsson, R. Wüchner, and K.-U. Bletzinger, "Interface jacobian-based co-simulation," *International Journal for Numerical Methods in Engineering*, vol. 98, no. 6, pp. 418–444, 2014. doi: 10.1002/nme.4637. [Online]. Available: http://dx.doi.org/10.1002/nme.4637

[24] A. Garro and A. Tundis, "On the reliability analysis of systems and sos: The ramsas method and related extensions," *Systems Journal, IEEE*, vol. 9, no. 1, pp. 232–241, March 2015. doi: 10.1109/JSYST.2014.2321617

[25] J. Martin, *Application Development Without Programmers*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1982. ISBN 0130389439

[26] S. Beydeda, M. Bock, and V. Gruhn, *Model-Driven Software Development*. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-25613-7

[27] J. Braunagel, P. Vuthi, W. Renz, H. Schäfers, H. Zarif, and H. Wiechmann, "Determination of load schedules and load shifting potentials of a high number of electrical consumers using mass simulation," in *Electricity Distribution (CIRED 2013), 22nd International Conference and Exhibition on*, June 2013. doi: 10.1049/cp.2013.1240 pp. 1–4.

[28] S. Moss and P. Davidsson, *Multi-Agent-Based Simulation: Second International Workshop, MABS 2000, Boston, MA, USA, July 2000; Revised and Additional Papers*, ser. Lecture Notes in Artificial Intelligence. Springer, 2001. ISBN 9783540415220

Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 1773–1778

DOI: 10.15439/2015F200
ACSIS, Vol. 5

# Energy Expenditure in Multi-Agent Foraging: An Empirical Analysis

Ouarda Zedadra and Hamid Seridi
LabSTIC Laboratory, 8 may 1945 University
P.O.Box 401, 24000 Guelma, Algeria
Department of computer science Badji Mokhtar
Annaba University P.O.Box 12, 23000 Annaba, Algeria
Email: zedadra_nawel1, seridihamid@yahoo.fr

Nicolas Jouandeau
Advanced Computing laboratory
of saint-Denis Paris 8 University
Saint Denis 93526, France
Email: n@ai.univ-paris8.fr

Giancarlo Fortino
DIMES, Universita'
della Calabria
Via P. Bucci, cubo 41c - 87036
- Rende (CS) - Italy
Email:g.fortino@unical.it

***Abstract*—A major challenge in swarm robotics is to minimize energy and time costs. We focus in this paper on multi-agent foraging algorithms that uses ant-like agents with limited energy. By considering energy consumption, we propose a new Energy aware Cooperative Switching Algorithm for Foraging (EC-SAF) that optimizes the whole system search and transport operations needed to collect resources over time. Unnecessary moves are avoided according to the following two premises: (1) Quick search and optimal paths provided by Stigmergic Multi-Ant Search Area (S-MASA) algorithm; (2) Quick homing provided by using the optimal paths created while searching. Results indicate that EC-SAF is promising in reducing swarm energy consumption compared to an energy-aware version of the c-marking algorithm (Ec-marking).**

## I. Introduction

SWARM robotics is concerned with the design of artificial robot swarms based upon the principles of swarm intelligence [1] [2] [3]. Promising solutions are expected by using numerous simple robots [4] [5] [6]. The collection carries out complex tasks based on simple rules, without spending much computational power and much physical energy [7]. Foraging robots are mobile robots that search for objects, and transport them to one or more storage points. It is a benchmark problem used in swarm robotics for several reasons: (1) It integrates several complex sub-tasks such as exploration, navigation, manipulation and transport; (2) It constitutes a canonical problem for the study of robot-robot cooperation; and (3) Many real-world applications are instances of foraging robots (like cleaning, harvesting, searching and rescuing) [8]. Foraging robots perform tasks that consumes energy, and must have a means of obtaining more energy to complete missions successfully. The most common strategies for powering long-lived autonomous robots are: (1) Capture ambient energy directly from the environment, also known as energy scavenging; or (2) Transfer energy from a recharging station [9]. Several options are possible in the last case: (1) Working robots perform their work until their energy falls below a given threshold. At this time, they return to recharging station, to recharge their energy [10] [11]; (2) Working robots can stay at the working site permanently, while special dock robots visit them periodically to provide them with energy [9] [12]; (3) Robots could transfer the energy also between them by comparing their energy's level [8]; and (4) Robots have to decide between search and transport, where transport can be applied to different resources. Even if resources are unknown at the beginning, as robots can be recruited by others, robots have sometimes to choose between carrying resources to the nest, carrying resources to another location in the field or searching for new resources, according to minimize the global energy consumption and maximize the global resources collected (that are equivalent to the accumulated energy) [13].

In this paper, we study the energy expenditure in the Cooperative Switching Algorithm for Foraging (C-SAF) [14] by addressing and analyzing two points:(1) What is the impact of collective exploitation of food provided by recruitment in C-SAF algorithm on energy consumption? (2) Does the division of search space (by using multiple sinks) in C-SAF improves energy efficiency? C-SAF robots are ant-like agents with limited computing and memory capacities. C-SAF provides quick search by using S-MASA algorithm [15] as search strategy, and quick exploitation by recruiting agents. Results are better than the standard reference algorithm and performances are emphasized with cooperation [16]. In this work, we propose to enhance foraging algorithms efficiency by taking energy into account. Individual energy and overall swarm energy are considered during resources location and exploitation.

The remainder of the paper is organized such as follows: in Section II we present the EC-SAF algorithm, its Finite State Machine (FSM) and a description of different states and we present the Ec-marking algorithm, an enhanced version of the c-marking algorithm [16], with which we compared the obtained results. In Section III, we present the performance indices used, describe the scenarios used for simulations and compare obtained results of the two algorithms. We finish with a conclusion in Section IV and some future works.

## II. An Energy-aware Multi-Agent Foraging Algorithm

In this Section, we present the EC-SAF algorithm, the FSM of our foraging agents and a description of different states. We present the Ec-marking algorithm, with which we compare the obtained results.

978-83-60810-66-8/$25.00©2015, IEEE

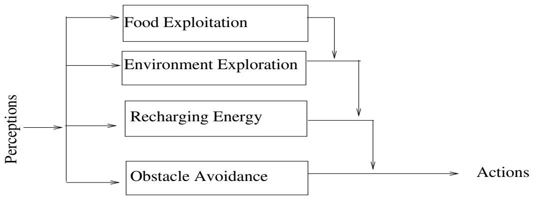


Fig. 1: Subsumption architecture of our foraging agents

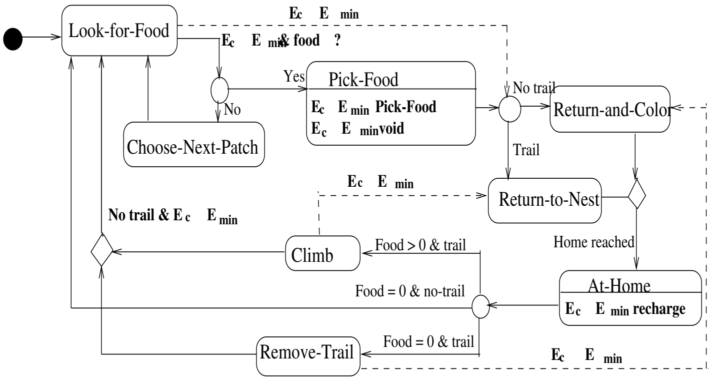


Fig. 2: State transition diagram of a foraging agent. Dotted arrows are the added transitions related to the recharging behavior of agents

### A. EC-SAF Algorithm

Our foraging agents use a four layered subsumption architecture [17] where each layer implements a particular behavior: *Environment exploration* is the lowest priority layer in this architecture. It consists in exploring the environment, therefore, it includes the states *Choose-Next-Patch* and *Look-for-Food*. *Food exploitation* consists in exploiting food when it is found, it envelops the states *Pick-Food*, *Return-to-Nest*, *Return-and-Color*, *At-Home*, *Climb* and *Remove-Trail*. *Recharging energy* consists of the set of states that allow agents to return home to recharge when their energy falls below a threshold, it includes the states *Return-to-Nest*, *Return-and-Color*, *Remove-Trail* and *At-Home*. *Obstacle avoidance* is the higher priority layer, it implements the obstacle avoidance behavior. Higher priority layers are able to subsume lower levels in order to create viable behavior (see Figure 1 for an illustration of the architecture). The behavior of our foraging agents is enhanced from [14] to deal with energy limitation. It is shown by the state transition diagram in Figure 2 where dotted arrows represent the new transitions used when the current energy of an agent ($E_c$) falls below the fixed threshold ($E_{min}$). States are described below and the enhanced algorithm is given by Algorithm **??**:

**Look-for-Food:** If $E_c > E_{min}$ and there exists a food here, agent executes *Pick-Food* state, while if there exists no food it executes *Choose-Next-Patch* state. If its $E_c <= E_{min}$, it turns to *Return-to-Nest* state if there exists a trail, or to *Return-and-Color* state, if there exists no trail.

**Choose-Next-Patch:** If an obstacle is detected, the agent calls the procedure Avoid_Obstacle(). If no obstacle is there, the agent climbs the brown trail to reach the food location if there exists one, it spreads then the information to its left cell. It lays a limited amount of pheromone $\mathbb{P}$ in current cell, adjusts its heading by executing S-MASA Algorithm [15] and moves one step forward. It turns automatically when finished to *Look-for-Food* state.

**Pick-Food:** If $E_c > E_{min}$, agent picks a given amount of food and spreads the information to its left cell. However, If $E_c <= E_{min}$ it does not pick food. It executes in the two cases *Return-to-Nest* state, if there exists a trail or *Return-and-Color* state if there exists no trail.

**Return-to-Nest:** The agent moves to one of colored neighboring cells with the lowest $\mathbb{P}$ value. It remains in this state until home is reached, it turns then to *At-Home* state.

**Return-and-Color:** The agent moves to one of the four neighboring cells with the lowest $\mathbb{P}$ value and marks its trail with yellow and remains in this state until it reaches the home; it turns then to the *At-Home* state.

**At-Home:** The agent unloads food if it carries one. If its current energy *($E_c$)* is below *($E_{min}$)*, it recharges its energy to the maximum amount $E_{max}$. It goes to *Climb* state if there exists one and the amount of food is $> 0$. If amount of food is $= 0$, it executes *Remove-Trail* state, else it turns to *Look-for-Food* state.

**Climb:** Agent moves to one of its four colored neighbors with a $\mathbb{P}$ value greater than the $\mathbb{P}$ value of the current cell. It remains in this state until no colored cell (yellow trail) exists and its $E_c > E_{min}$, it turns then to the *Look-for-Food* state. If its $E_c <= E_{min}$ and since there exists a trail, it executes the *Return-to-Nest* state in order to return home for recharging its energy.

**Remove-Trail:** The agent moves to a colored cell with the greatest $\mathbb{P}$ value and resets its color to the default color (black). It remains in this state until no colored cell is found and $E_c > E_{min}$, it turns then to the *Look-for-Food* state. If $E_c <= E_{min}$, the agent returns to home to recharge and executes the *Return-and-Color* state since it already removed the existing trail and to keep track of the last position from where he will continue removing after it recharges its energy.

The modeling of the components of our multi-agent system are detailed below:

- *Environment Model:* an N X N grid world. It contains a set of agents, food, nest and obstacles.
- *Agent Model:* agents have limited processing power and memory, simple sensors, do not know the position of food nor the map of environment.
- *Pheromone Releasing and Evaporation:* Two types of pheromone have no diffusion and evaporation properties and they are used either to mark a transport or recharging trails or a recruitment trail. The third one is used to mark already visited cells.

### B. Ec-marking Algorithm

The Ec-marking algorithm, which is an enhanced version of the c-marking algorithm [16] to deal with energy limitation,

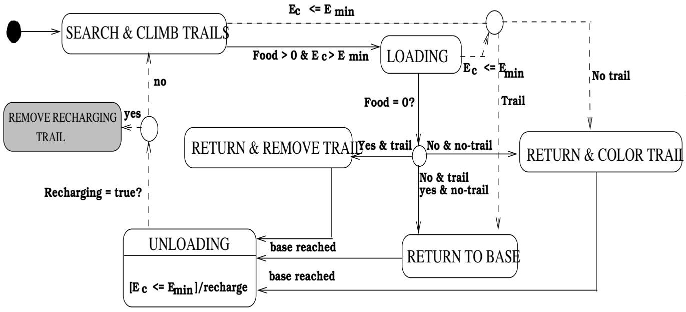


Fig. 3: State transition diagram of an Ec-marking agent, where colored state, dotted transitions and bold guards are related to energy-aware behavior of agents

is given by Figure 3. Agents while exploring the environment build simultaneously paths between food and nest which results in building an ascending Artificial Potential Field (APF) incrementally. An Ec-marking agent is always in one of the states depicted by Figure 3, where the enhancements made represented by filled states, dotted transitions and bold guards. This set of states is described below:

**SEARCH & CLIMB TRAIL:** If food >0 and $E_c > E_{min}$, the agent executes the *LOADING* state. If $E_c <= E_{min}$, it looks for trail, if there exists one it executes *RETURN TO BASE*, else if there exists no trail it executes *RETURN & COLOR TRAIL*, else it moves to food if it is found and if not it moves to a neighboring cell not marked yet.

**LOADING:** The agent picks a given amount of food and food here is exhausted and there exists a trail, it executes *RETURN & REMOVE TRAIL*, if food is not exhausted and there exists no trail, it executes *RETURN & COLOR TRAIL*, else it executes *RETURN TO BASE.*

**RETURN & COLOR TRAIL:** The agent moves to one of the four neighboring cells with the lowest $APF$ value and changes its color to a trail color, it remains in this state until it reaches the home; it turns then to the *UNLOADING* state.

**RETURN & REMOVE TRAIL:** The agent moves to one of the four neighboring cells with the lowest $APF$ value and changes its color to the default one (black), it remains in this state until it reaches the home; it turns then to the *SEARCH & CLIMB TRAIL* state.

**RETURN TO BASE:** The agent moves to a colored neighboring cell with value minimal to its current value, until it reaches the home; it turns then to the *SEARCH & CLIMB TRAIL* state.

**UNLOADING:** The agent unloads the food at home. If $E_c <= E_{min}$, it recharges its energy to $E_{max}$ and if its statue is recharging it changes state to the *REMOVE RECHARGING TRAIL*, else it changes to the *SEARCH & CLIMB TRAIL.*

**REMOVE RECHARGING TRAIL:** The agent moves to a colored neighboring cell with higher $APF$ value and resets its color to the default one, until no colored cell is found it changes then to the *SEARCH & CLIMB TRAIL.*

## III. Simulation Results

Three performance indices are used to compare the algorithms (*Total food returned*, *Energy efficiency* and *Total energy*). Through simulations, we compared the four algorithms (EC-SAF [Algorithm **??**], C-SAF [14], c-marking [16] and Ec-marking [Figure 3]) on the basis of *Total food returned*, to verify if an energy-aware management can improve performances. After that, the two energy-aware algorithms proposed in this paper (EC-SAF and Ec-marking) are compared between each other on the basis of *Energy efficiency* ($E_{eff}$) and *Total energy* performance indices.

- **Total food returned**: is the total amount of food (in units) returned over some elapsed time.
- **Energy efficiency**: it is the energy spent while foraging one food location. It is calculated according to equation 1.

$$E_{eff} = \frac{TotalEnergyOfConsumedFood}{NumberOfReturnedFood} \qquad (1)$$

  Where *Total Energy Of Consumed Food*, is the sum of each agent energy spent in exhausting one food location starting from finding the food until it is exhausted. *Number Of Returned Food* is the quantity of units of food returned;
- **Total energy**: is the total energy spent by all agents to search and exhaust all the food locations.

The energy consumption of an agent at each state is defined on the basis of the power of real equipment (such as motor, sensor and processor) required to achieve that state. It is inspired by the B-swarm model [10]. Agent consumes 1 unit of energy per simulation update for the states that do not need hard work (such as *Climb* and *Return-to-Nest*), while in states that need hard manipulation such as: depositing pheromone (in *Choose-Next-Patch* state), loading food (in *Pick-Food* state), unloading food (in *at-Home* state), pick-up pheromone (in *Remove-Trail* state), and deposit pheromone (in *Return-and-Color* state), the agent consumes 5 units of energy per simulation update. However, for the *Avoid-Obstacle ()*, agent changes its direction only and consumes 3 units of energy per simulation update. For the Ec-marking, the energy consumed is: 5 units of energy per simulation update for states *SEARCH*, *LOADING*, *RETURN & COLOR TRAIL*, *RETURN & REMOVE TRAIL*, *UNLOADING*, *REMOVE RECHARGING TRAIL* and 1 unit of energy per simulation update for states *CLIMB TRAIL*, *RETURN TO BASE* and 3 units of energy per simulation update for the *avoid obstacle()*. Simulation is based on Netlogo [18]. The simulation results are the average of ten simulations. Four kinds of simulations are reported in this paper. In each simulation several related parameters are to be fixed: agent parameters (number and capacity), world parameters (size, complexity and sinks number) and food parameters (density and concentration) where: *Agent's number* is the number of agents that participate at each simulation, *Agent's capacity* is the amount of food (units) that an agent can transport at each time. *World size* is the dimension of the search space, it is a grid of *N X N cells*, the world is obstacle-free or obstacle represent the *world complexity*, *sinks number*

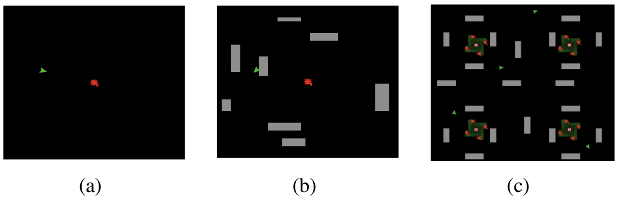


Fig. 4: World setups used in simulations (a) obstacle-free environment (b) obstacle environment (c) environment with 4 sinks, where red arrows are agents, green arrows are food locations, gray blocks are obstacles and pink squares are sinks

TABLE I: Parameters of scenario 1, scenario 2, scenario 3 and scenario 4

| Parameter | | Value |
|---|---|---|
| **Scenario 1** | *Total food returned Analysis* | |
| World size | | 40 X 40 cells |
| Number of agents | | 8 |
| Food density | | 2 sites |
| Food concentration | | 50 units |
| Agent's capacity | | 1 unit |
| Sinks number | | 1 |
| **Scenario 2** | *Energy Efficiency Analysis* | |
| World size | | 40 X 40 cells |
| Number of agents | | 5 – 30 |
| Food density | | 1 site |
| Food concentration | | 20 units |
| Agent's capacity | | 1 unit |
| Sinks number | | 1 |
| **Scenario 3** | *Energy Efficiency Analysis* | |
| World size | | 80 X 80 cells |
| Number of agents | | 48 |
| Food density | | 1 site |
| Food concentration | | 20 units |
| Agent's capacity | | 1 unit |
| Sinks number | | 1 – 16 |
| **Scenario 4** | *Total Energy Analysis* | |
| Number of ticks | | 50 – 300 |
| World size | | 40 X 40 cells |
| Number of agents | | 15 |
| Food density | | 1 site |
| Food concentration | | 20 units |
| Agent's capacity | | 1 unit |
| Sinks number | | 1 sink |

is the number of the home or base station to where agents store food and recharge energy. *Food density* is the number of food locations (sites), distributed randomly in the environment. *Food concentration*, indicates the amount of food that each site contains (we refer to it as unit in the paper). At first stage, we wanted to test if energy-aware can improve efficiency of our C-SAF algorithm or not, therefore, we proposed scenario 1 (see TABLE I), where we calculate the total amount of food returned over some elapsed time. The obtained results with the four algorithms (C-SAF, EC-SAF, c-marking and Ec-marking) are depicted by Figure 5.

From Figure 5, we observe that in C-SAF and c-marking algorithms, the total food returned increases with the increase in ticks (below 300 ticks) because agents still have energy, when their energy is exhausted, agents die and the total food returned does not increase (over 300 ticks). While the total food returned keep increasing in the energy-aware versions of the algorithms (EC-SAF and Ec-marking), because agents return to recharge when their energy falls below the fixed threshold and resumes their tasks. From this experimental results, we conclude that an energy-aware version are needed to improve performances.

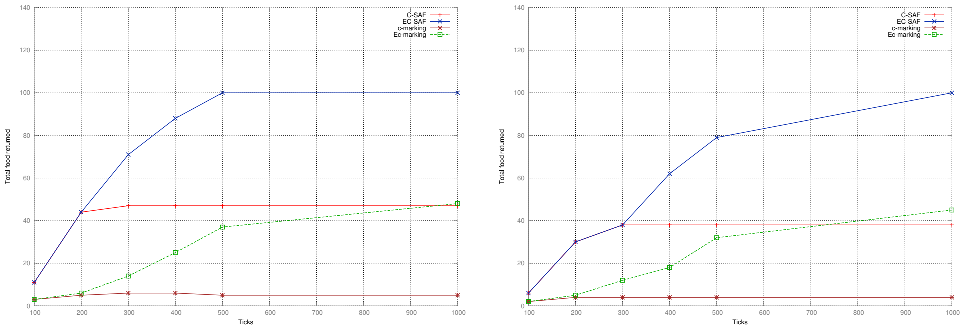
The EC-SAF algorithm is also compared with the Ec-marking one, in order to test if it can improve energy consumption or not. We proposed therefore, Scenario 2, scenario 3 and scenario 4, where the two first ones are used to test the impact of varying agent's number and sink's number (search space division) on energy efficiency. While in scenario 4, we observe the whole energy consumed over some elapsed time, to test whether the algorithm consumes much or less energy when operating (see TABLE I for the description of the three scenarios). The three world setups that are used for simulations including positions of nest, food, obstacles and agents, are reported in Figure 4.

### A. *Results in Scenario 2*

The energy efficiency in EC-SAF does not change when changing the number of agents. In opposite to the number of ticks that is reduced with the increase of agents number, the energy consumed by one agent is the same consumed by multiple agents that participate at food exploitation. However, the energy efficiency in c-marking decreases with each increase in agents number. The energy consumed depends on the length of path that relays the food and the home. In Ec-marking algorithm, the paths are not optimal when number of agents is small thus energy consumption is great, with the increase of agent's number the length of paths will be reduced and the energy consumption decreases. Because of the optimal paths provided by S-MASA algorithm [15], EC-SAF gives better results than the Ec-marking one (see Figure 6(a)). Results in obstacle environment are similar to the ones in obstacle-free environment configuration, with additional steps needed to avoid obstacles (see Figure 6(b)).

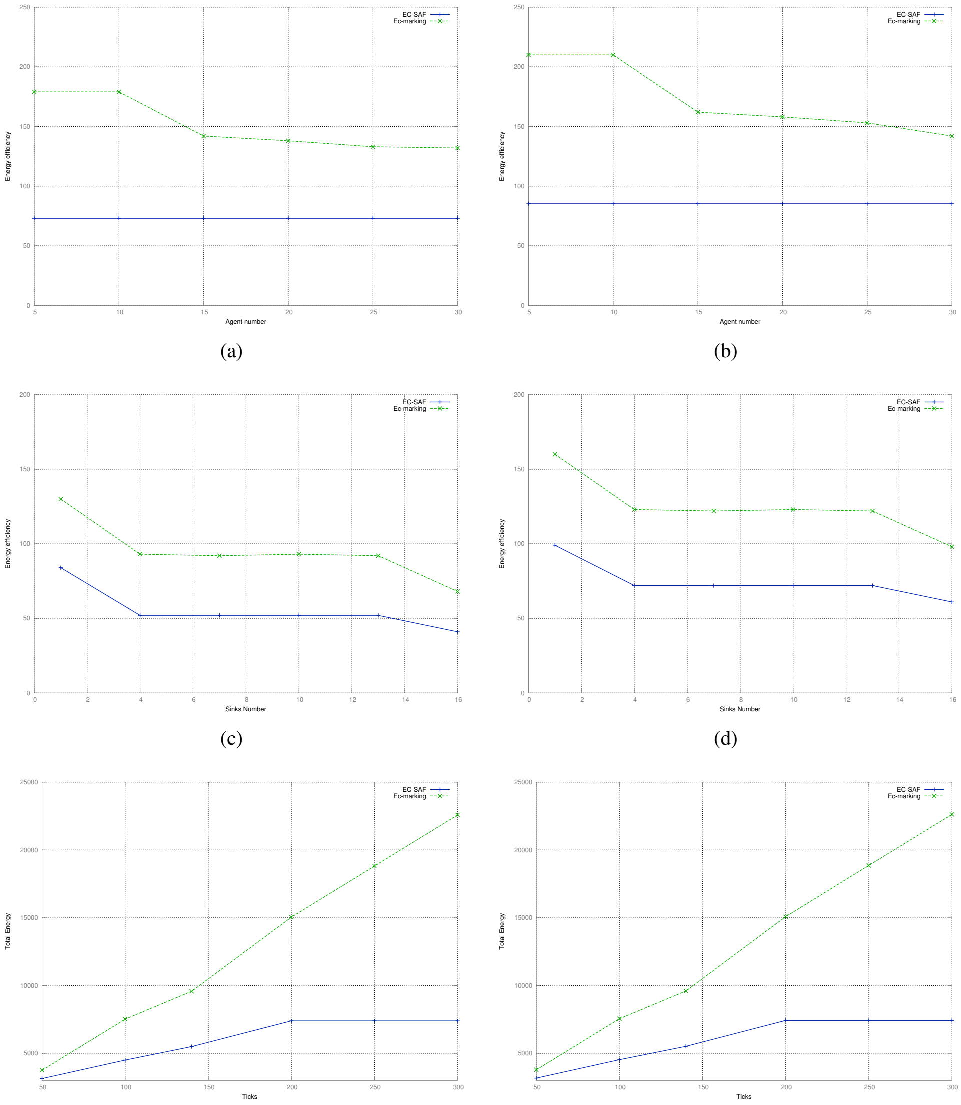
### B. *Results in Scenario 3*

Using multiple sinks divide the whole search-space into sub-search spaces with smaller size. When increasing the the number of sinks the number of sub-spaces is increased too and the size is reduced more. The path length to food is reduced each time we increase the number of sinks (food takes fixed position in all simulations). In EC-SAF less consumption of energy is reached with 16 sinks, where the size of sub-spaces is the smallest and the path length to the food is the smallest (6 cells). The energy consumption is greater with 4, 7, 10 and 13 sinks (the path length to food is 8 cells). However, it is the greatest with 1 sink, because the search-space size is the greatest and the path length to food is the longest (14 cells). In Ec-marking, the energy consumption is great with



(a) (b)

Fig. 5: Simulation results of scenario 1 in : (a) obstacle-free environment, (b) obstacle environment.



(a) (b)

(c) (d)

(e) (f)

Fig. 6: Simulation results in obstacle-free and obstacle environment of: (a), (b) scenario 2. (c), (d) scenario 3. (e), (f) scenario 4

1 sink because the search-space size is the greatest, the path length to food is the longest and paths are not optimal. It is reduced with 16 sinks, since the path length to food is reduced. ES-CAF provide less energy consumption rather than Ec-marking because of the optimal paths induced by the S-MASA algorithm [15]. Figure 6(c) shows the results comparison between the two algorithms. Results in obstacle environment are same as in obstacle-free environment configuration, with additional steps needed to avoid obstacles (see Figure 6(d)).

### C. Results in Scenario 4

The total energy increased with increasing the number of ticks as shown in Figure 6(e). It is stable in EC-SAF above 200 tick because the agents reach the boundaries of the search world (the finish time of the foraging is 140 ticks). However, the finish time of foraging in Ec-marking is 300 ticks and until this time the total energy continue to increase. Also in this scenario EC-SAF provides a less consumption of energy in comparison to Ec-marking one. Also in obstacle environment, the total energy in the two algorithms increased with increasing the number of ticks but it is more in Ec-marking algorithm than in EC-SAF algorithm (see Figure 6(f)).

## IV. Conclusion

We investigated in this paper the energy efficiency and the total energy consumed of the EC-SAF algorithm as changing the number of agents (to test the benefit of collective foraging), changing the sinks number (to test the benefit of dividing search space) and calculating total energy consumed over ticks (to test the impact of the search strategy used). Simulation results show that energy efficiency in EC-SAF, does not change when changing agents number because the set of agents execute the same states as one agent and thus consume the same energy consumed by one agent. It can be reduced when using multiple sinks, and it depends on the path length between food and home, if the path is reduced with search space division the energy efficiency is even reduced (and vice versa). While the total energy increased with the increase in number of ticks, it stops changing and becomes stable when all food is foraged and the search space boundaries are reached. EC-SAF gives better results than the enhanced c-marking one, because of the optimal paths and the quick search provided by S-MASA algorithm [15].

In the future, we intend to explore other environment configurations and examine other possibilities to reduce the energy consumption in EC-SAF.

## References

[1] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, “Swarm robotics: a review from the swarm engineering perspective,” *Swarm Intelligence*, vol. 7, no. 1, pp. 1–41, 2013. [Online]. Available: http://dx.doi.org/10.1007/s11721-012-0075-2

[2] M. Dorigo, M. Birattari, and M. Brambilla, “Swarm robotics,” *Scholarpedia*, vol. 9, no. 1, p. 1463, 2014. [Online]. Available: http://dx.doi.org/10.4249/scholarpedia.1463

[3] J. C. Barca and Y. A. Sekercioglu, “Swarm robotics reviewed,” *Robotica*, vol. 31, no. 03, pp. 345–359, 2013. [Online]. Available: http://dx.doi.org/10.1017/S026357471200032X

[4] S. Konur, C. Dixon, and M. Fisher, “Analysing robot swarm behaviour via probabilistic model checking,” *Robotics and Autonomous Systems*, vol. 60, no. 2, pp. 199–213, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.robot.2011.10.005

[5] A. Saxena, C. Satsangi, and A. Saxena, “Collective collaboration for optimal path formation and goal hunting through swarm robot,” in *5th International Conference on Confluence The Next Generation Information Technology Summit (Confluence)*. IEEE, 2014, pp. 309–312. [Online]. Available: http://dx.doi.org/10.1109/CONFLUENCE.2014.6949364

[6] G. Pini, A. Brutschy, M. Frison, A. Roli, M. Dorigo, and M. Birattari, “Task partitioning in swarms of robots: An adaptive method for strategy selection,” *Swarm Intelligence*, vol. 5, no. 3-4, pp. 283–304, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11721-011-0060-1

[7] D. H. Kim, “Self-organization for multi-agent groups,” *International Journal of Control Automation and Systems*, vol. 2, pp. 333–342, 2004. [Online]. Available: http://dx.doi.org/

[8] A. F. Winfield, S. Kernbach, and T. Schmickl, “Collective foraging: cleaning, energy harvesting and trophallaxis,” *Handbook of Collective Robotics: Fundamentals and Challenges, Pan Stanford Publishing, Singapore*, pp. 257–300, 2011. [Online]. Available: http://dx.doi.org/

[9] A. Couture-Beil and R. T. Vaughan, “Adaptive mobile charging stations for multi-robot systems,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2009, pp. 1363–1368. [Online]. Available: http://dx.doi.org/10.1109/IROS.2009.5354816

[10] L. Pitonakova, R. Crowder, and S. Bullock, “Understanding the role of recruitment in collective robot foraging,” *MIT Press*, 2014. [Online]. Available: http://dx.doi.org/10.7551/978-0-262-32621-6-ch043

[11] J.-H. Lee, C. W. Ahn, and J. An, “A honey bee swarm-inspired cooperation algorithm for foraging swarm robots: An empirical analysis,” in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2013, pp. 489–493. [Online]. Available: http://dx.doi.org/10.1109/AIM.2013.6584139

[12] S. Kernbach and O. Kernbach, “Collective energy homeostasis in a large-scale microrobotic swarm,” *Robotics and Autonomous Systems*, vol. 59, no. 12, pp. 1090–1101, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.robot.2011.08.001

[13] A. Campo and M. Dorigo, “Efficient multi-foraging in swarm robotics,” in *Advances in Artificial Life*. Springer, 2007, pp. 696–705. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-74913-4_70

[14] O. Zedadra, N. Jouandeau, H. Seridi, and G. Fortino, “Design and analysis of cooperative and non-cooperative stigmergy-based models for foraging,” in *Proceedings of the 19th IEEE International Conference on Computer Supported Cooperative Work in Design*, 2015.

[15] O.Zedadra, N. Jouandeau, H. Seridi, and G. Fortino, “S-MASA: A stigmergy based algorithm for multi-target search,” in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 2. IEEE, 2014, pp. 1477–1485. [Online]. Available: http://dx.doi.org/10.15439/2014F395

[16] O. Simonin, F. Charpillet, and E. Thierry, “Revisiting wavefront construction with collective agents: an approach to foraging,” *Swarm Intelligence*, pp. 113–138, 2014. [Online]. Available: http://dx.doi.org/10.1007/s11721-014-0093-3

[17] R. A. Brooks, “A robust layered control system for a mobile robot,” *Journal of Robotics and Automation*, vol. 2, no. 1, pp. 14–23, 1986. [Online]. Available: http://dx.doi.org/10.1109/JRA.1986.1087032

[18] U. Wilensky, “Netlogo. http://ccl.northwestern.edu/netlogo/,,” in *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL*, 1999.

# 4th International Workshop on Smart Energy Networks & Multi-Agent Systems

THE EMERGING smart infrastructure in energy networks represents a major paradigm shift in resource allocation management with the aim to extend the centralised supply management model, towards a decentralised supply-and-demand management that is expected to enable more efficient, reliable and environment-friendly utilisation of primary energy resources.

Together with this vision, there are new and complex tasks to manage, in order to ensure safe, cost-reducing and reliable energy network operations. This includes the integration of various renewable energy systems, like the photovoltaic or the wind energy, which are able to reduce the greenhouse gas emissions but that are working under greater uncertainty; as well as the interaction of transport and storage systems for energy that are envisioned through techniques like 'Power to Gas' and fuel cells, which are using the electrical and the gas transportation network.

Further tasks can be found in the fact that the market participants (e.g. simply households) are becoming more autonomous and intelligent through technologies like smart metering, which requires a coordinated demand side management for millions of producers, consumers or, if this applies, prosumers by means negotiations and agreements.

Information and communication technologies are key enablers of the envisioned efficiencies, both on the demand and the supply sides of the smart energy networks, where the agent-paradigm provides an excellent first modelling approach for the distributed characteristic in energy supply systems. On the demand side they aim at supporting end-users in optimising their individual energy consumption, e.g. through the deployment of smart meters providing real-time usage and cost of the energy and the use of demand-response appliances that can be controlled according to the user preferences, energy cost and carbon footprint. On the supply side they aim at optimising the network load and reliability of the energy provision, e.g. through active monitoring and prediction of the energy usage patterns, and proactive control and management of the reliable energy delivery over the networks. It is also envisaged that they will be able to influence the demand through the dynamic adjustments of the energy price in order to influence the end user behaviour and energy usage patterns throughout and across the energy networks for electricity, gas and heat.

Although a significant effort and investment have been already allocated into the development of smart grids, there are still significant research challenges to be addressed before the promised efficiencies can be realised. This includes distributed, collaborative, autonomous and intelligent software solutions for simulation, monitoring, control and optimization of smart energy networks and interactions between them.

## Topics

The SEN-MAS'15 Workshop aims at providing a forum for presenting and discussing recent advances and experiences in building and using multi-agent systems for modelling, simulation and management of smart energy networks. In particular, it includes (but is not limited to) the following topics of interest:

- Experiences of Smart Grid implementations by using MAS
- Applications of Smart Grid technologies
- Management of distributed generation and storage
- Islands Power Systems, Microgrid Applications
- Real time configurations of energy networks
- Distributed planning process for energy networks by using MAS
- Self-configuring or self-healing energy systems
- Load modelling and control with MAS
- Simulations of Smart Energy Networks
- Software Tools for Smart Energy Networks
- Energy Storage
- Electrical Vehicles
- Interactions and exchange between networks for electricity, gas and heat
- Stability in Energy Networks
- Distributed Optimization in Energy Networks

## Event Chairs

**Derksen, Christian,** University Duisburg-Essen, Germany

**Kowalczyk, Ryszard,** Swinburne University of Technology, Melbourne, Victoria, Australia

## Steering Committee

**Derksen, Christian,** University Duisburg-Essen, Germany

**Lehnhoff, Sebastian,** OFFIS - Institute for Information Technology, Germany

**Kowalczyk, Ryszard,** Swinburne University of Technology, Melbourne, Victoria, Australia

**Nahorski, Zbigniew,** Systems Research Institute - Polish Academy of Science, Poland

## Program Committee

**Andrew, Lachlan**

**Braubach, Lars,** University of Hamburg, Germany

**Guttmann, Christian,** Monash University, Australia

**Klusch, Matthias,** German Research Center for Artificial Intelligence, DFKI, Germany

**Linnenberg, Tobias,** Helmut Schmidt University, Germany

**Moench, Lars,** FernUniversität Hagen, Germany

**Ossowski, Sascha,** University Rey Juan Carlos, Spain

**Sonnenschein, Michael,** Carl von Ossietzky University, Oldenburg, Germany

**Sudeikat, Jan,** Hamburg Energie GmbH, Germany

**Unland, Rainer,** Universität Duisburg-Essen, Germany

**Weber, Christoph**

DOI: 10.15439/2015F310
ACSIS, Vol. 5

# GRENAD, a Modular and Generic Smart-Grid Framework

Sylvain Ductor, Jesus-Javier Gil-Quijano, Nicolas Stefanovitch, Pierrick Roger Mele
CEA-LIST, Département Métrologie Instrumentation et Information (DM2I),
Laboratoire d'Analyse de Données et Intelligence des Systèmes (LADIS),
Digiteo lab - bat 565 - Point courrier 192,
91191 Gif-sur-Yvette Cedex - France Bâtiment 425,
Email: {surname.name}@cea.fr

## I. Introduction

***Abstract*—We present in this paper GRENAD, a Multi-Agent System based framework for the simulation and piloting of power-grids and particularly smart grids. Exploiting a component-based approach, it allows a flexible design of complex smart grid applications by providing a generic canvas where extensible, modular and reusable components, defined on the basis of their functionalities, can be easily combined and connected. Thanks to Multi-Agent approach, a set of such components can naturally be integrated into a coherent economical agent. GRENAD makes no assumption on the energy definition and eases the development of MAS control algorithms for smart grids. The level of details of the energy-related information is controllable. This information is computed either through internal physical models or by interfacing with external simulators. We present here our model, illustrate its features with a rich example which exhibits its genericity, and demonstrate how a coordination protocol can easily be integrated to it.**

ENERGY supply is at the core economy while being also one of the major items of expenditure. Decreasing fossil energy supplies as well as rising concerns about climate made it critical to change in the near future the way energy is produced, distributed and consumed. Smart grids are highly automated power networks that possess fine grained monitoring and control capabilities, from the power plant to the domestic appliance. Easy access to information and autonomous decision making allow smart grids to save energy by quickly reacting to changes in the environment and reorganising demand, production and distribution. Smart grid are also meant to allow the massive integration of distributed renewable energy resources (DER) as well as new equipments such as combined heath and power (CHP) generation and energy storage. As such smart grids are called to play a crucial role in the coming years for the efficient control of power systems.

Hardware for production, monitoring and storage in smart grids already exist. While still not being widely deployed, the penetration of such equipments is constantly progressing and increasingly supported by legislations. However they raise new challenges. Indeed, as opposed to the traditional power grids approches, energy flows may be intermittent and/or bidirectional. Moreover smart grids technologies also foster a radical change in the business of energy supply, by allowing the energy to be islanded, it is to say produced, sold and consumed locally. Lastly, increased interconnection and information exchange between actors enables them to coordinate more closely, providing thus a new opportunity to reach higher efficiency and revenues.

One of the most challenging aspects of smart grids is thus at the engineering and at the logical level: the development of efficient control and coordination algorithms that are able to fully exploit their different potentialities. In order to tackle these challenges the use of a Multi-Agent Systems (MAS) approach is particularly relevant as the structure of MAS closely match the structure and behaviour of smart grids: They are constituted of a set of interconnected distributed autonomous actors, each aiming at maximising their respective goals through coordination. Also, a given smart grid is composed by interconnecting differents components (production, distribution or storage) that may be used, as is, on another smart grid. It is thus relevant to adopt a modular approach for modeling the equipments in a reusable, and thus capitalizable, way. Therefore a key driver for the deployment of smart grids is the conception of a dedicated platform that ease their development by exploiting both oriented-component paradigm for equipment modelisation aspects and oriented-agent paradigm for coordination and optimisation aspects.

In the context of the Resilent FP7 project [20], we have developed such a platform, GRENAD, which stands for "`Gestion des Ressources ENergétiques, Autonome et Distribuée`" (Autonomous and Distributed Management of Energetic Resources). GRENAD presents two main aspects. The first is a feature rich API and Domain Specific Language (DSL) based on JADE [11] which allows to construct component-based JADE agents. As such it benefits of all the features and standard compliance of JADE and can natively be used in conjunction with other JADE-based applications. The second is a generic smart grid model implemented on top of this DSL. Thanks to these two aspects, it is possible to easily capitalize and reuse developed software across different smart grid applications.

While most MAS publications in the smart grid domain focus on the connection with a simulator or the conception of a new control algorithm, our contribution is a platform that offer enough flexibility to allow an easy interfacing with simulators and an easy implementation of control algorithms. Also, each

978-83-60810-66-8/$25.00©2015, IEEE

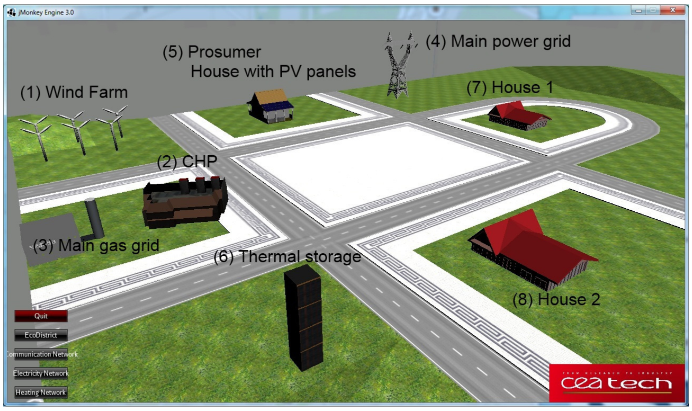


Fig. 1. A rich scenario

component defining the smart grid can natively be monitored, or even controlled, either by an agent, a GUI, or a remote optimisation algorithm.

In Section II we present a smart grid scenario exhibiting all critical features for the applications we are interested in. In Section III we present the related work and how they answer to the objectives highlighted by our scenario. In section IV we present the GRENAD agent and smart grid models and we describe how to implement the scenario with them. In Section V we show how to easily integrate a smart grid optimization algorithms in GRENAD.

## II. A Rich Scenario

In this section we present a district energy system scenario. The richness of this scenario allows us to demonstrate the capabilities of our MAS based tool to model complex energy management systems at the district-level. This scenario exhibits most of the main characteristics that challenge today the design of smart grid energy management systems [19]:

- Interdependence of several energy flows (electricity, heat and gas)
- Distributed generation and storage capabilities
- Interaction with the main energy grids (electricity and gas)
- Local renewable energy generation capabilities
- Prosumers
- Diverse consumption profiles

In the scenario, the considered components (producers, distribution grid, consumers/prosumers and storage systems) are modelled as autonomous agents that interact with the intention of providing real time management at the district level and optimizing the energy production and distribution in the district, in conjunction with the main energy networks.

### A. Business and organizational models

The behaviors of the different components are defined by their individual roles and their internal characteristics (physical, economical, quality of service, *etc.*). Those behaviors are constrained by the business and organizational models of the system as well as by the interconnection mode to the main grid (an energy system can work either in *connected or islanded mode* w.r.t the main energy grids). In our scenario we considered a system connected to the main gas and power grids (represented as components 3 and 4 in Figure 1).

Among the different existing business and organizational models we can list:

- **Free markets**: the components are competitors and interact via a market mechanism [12];
- **Virtual Power Plant (VPP)**: combination of production storage and consumption resources. A VPP [18] is adapted when all the resources belong to or are managed by a single actor;
- **Cooperative Virtual Power Plant (CVPP)**: In a CVPP [3] different actors manage/own the different components, they coordinate in a cooperative way to use shared resources (for instance centralized storage systems, the distribution network, etc.) and provide global services (for instance: grant the energy balance, the security of the network and/or the quality of service).

All these different organizational models can be considered in our approach. In our scenario we model interactions between components as a free market.

### B. Roles

The main roles that we support are *producer, storage, consumer, prosumer and distribution*. The different constraints and characteristics associated to those roles are described below:

*1) Producer components:* The characteristics of the generation of energy provided by producers depend on several aspects:

- **Internal physical constraints**: the nominal power (max power capabilities), start/stop conditions (time to be operational/off operation), their efficiency (it is the rate between the output generated energy and the input primary used energy), and for renewables the variability and the intermittence.
- **Controllability**: The controllability of a given generator depends on the primary energy source that it uses and on internal state variables. Most of renewable-energy sources (e.g. wind, solar) are considered non-controllable, it is, their availability and level of energy cannot be controlled. Nevertheless, some renewable based generators can adapt their internal state (i.e. orientation of wind turbines or solar trackers) in order to secure and optimally generate energy according to the primary-source real time conditions. Generators whose primary energy supply can be controlled provide different control levels, for instance some generators can only provide on/off control while others can provide intermediate levels of functioning.

In our scenario we consider two producers (the components 1 and 2 in Figure 1): a set of wind turbines for power generation and a Combined Heating and Power (CHP) generator that generates both power and heating from combustion of gas. Gas is provided by the main gas grid (component 3 in

Figure 1) and can be stored nearby the CHP. The system is also connected to the main power grid (component 4 in Figure 1). The gas and the power main grids are considered in our model as controllable producers, it is, we consider that the quantity of gas or the power drawn from the main grids can be controlled. In the real system only the gas flow can be directly controlled, the drawn power from the main grid is indirectly controlled and corresponds to balance needed power, it is, the consumed power minus the local producer power.

*2) Storage components:* The behavior of a storage component depends mainly on its nominal capacity, its rate of charge/discharge and its instantaneous state of charge (SoC). Other important variables are the efficiency of the conversion in the charge and discharge phases and the self-discharge rates. The usage of some storage systems, for instance batteries, is constrained by the charge/discharge cycling conditions that optimize their efficiency and increase their lifetime. In our scenario we consider a thermal storage component (see component 6 in Figure 1).

*3) Distribution components:* The distribution of energy generates losses due to the physical characteristics of the transportation system used (i.e. transmission capabilities, dissipation rates, etc.) and the distances of transmission of the energy. In general, at the district level, the electricity losses due to distribution are negligible while the heating and hydraulic (heating network related) losses are not. Some distribution components (as valves in heating networks and some types of transformers) allow the dispatching of input energy among different outputs. While the availability of energy over electrical networks is instantaneous (electricity is transmitted at about the speed of light), the transmission of energy over heating and cooling networks, due to the transport inertia, needs significant time (several minutes per km) to go from the injection to the consumption points. In our scenario we consider local heating and power networks: all the production and consumption components considered in the scenario are connected to at least one of these distribution networks.

*4) Consumer components:* The characteristics of the energy consumed, basically their load curves, depend mainly on the buildings physical characteristics (i.e. thermal inertia, storage capabilities), the behaviors of the inhabitants (presence/absence, used services) and the comfort constraints (e.g. ambient temperature set points). In our scenario we consider two consumers, the components 7 and 8 in Figure 1.

*5) Prosumer components:* Besides the characteristics that define the load curves of the consumer components, prosumers can partially or totally cover their needs in energy and under some conditions produce energy surplus that can be injected into the main or local energy grids. In our scenario we consider one prosumer (component number 5 in figure 1), that locally produces power thanks to the PV solar panels installed on its roof. It uses part of this energy for self-consumption and is able to inject part of this energy into the power distribution network.

### C. Planing and operation

Due to the temporal inertia of most of the components (e.g. time to start or change generator command level; thermal inertia on houses and distribution networks; time to charge/discharge storage systems) and the use of non-controllable sources (that are intermittent and variable), the energy usage needs to be planned before actual generation and delivery. In our approach, we consider *planning strategies* that are combined to *real-time operation strategies*. *Planning* is based on estimation of the consumption loads and generation capabilities and allows to establish negotiated (i.e. agreed by all parties) generation, storage and consumption schedules. For the renewable based generators, the estimation depends on weather and generation capabilities forecasting. For end consumers, the estimation depends mainly on weather and usage forecasting, as well as on thermal inertia evaluation. Those different forecasting based estimations lead to inaccuracies at the planning phase. In order to maintain the balance between consumption and generation and minimize the use of external generated energy, we implement and operate a mechanism that allows the near real-time correction of energy schedules when deviations of generation and load are detected (monitored or forecasted). This mechanism can be seen as a capacity market where global flexibility is provided by the combined individual flexibility capabilities (e.g. deferrable load, dedicated storage capabilities, etc.)

### D. Real-Time payment

Smart-Grid also aims to adapt the consumption to the energy production capacity. This is particularly critical when some producers are not controllable (e.g. weather dependant renewable energy source). By exploiting for instance a building thermal inertia, it is possible to respect a required comfort level while adapting the actual consumption.

A commitment of the consumer on how it will consume energy allows to optimise energy production and distribution during the planning. Several payment approaches [22] propose to impose penalties if such a commitment is not respected during operation phase. The actual payment of a consumer is thus computed by comparing its planned and operational consumption along with some other production-related and/or conventional parameters.

## III. Related Works

Actual testing and deployment of smart grid software require platforms able to directly support the execution of control software developed following as well as the ability simulate or even pilot power systems. Among the few available commercial systems that target specifically smart grids, only a bit of them follow the MAS approach.

The available commercial software that simulate and control traditional grids [15], [7], can to some extent be used to simulate parts of a smart grid, notably for the distribution aspect. However, smart grids possess several specificities that preclude the use of such software for the full chain control. On the other hand, smart grid software have not reached this level of maturity and are mostly experimental or at the level of academic research. Dedicated software solutions for smart grids exist independently, for different aspects presented in this scenario: storage simulation and dimensioning [10], DER integration [5], demand side management [9] and VPP [2]. While none of these commercial software uses a MAS approach, PowerMatcher [17] is an exception. It can perform simulation and piloting and uses auction protocols to interact over a market of flexibilities.

Among non commercial software, GridLab-D [4] is a notable one, it is a low level electrical simulation tool based on a MAS paradigm. It allows to assign different profiles of consumption to the devices on the network, but fails to represent economic actors and is not able to endow agents with advanced smart behaviours that react to change in their environment.

Most academic works in smart grids consider separately the development of a platform, a model and control algorithms. As such they provide very limited reusability and interoperability between the different layers. Platforms in the literature mostly act as a middleware and focus on the coupling of a MAS platform (JADE [11] being the most popular) with an electrical simulator and data exchange formats[1], [23], [21]. The designs provided in papers that present models [16], [8] fail to acknowledge either the simulation or the control aspects and often both of them, limiting their applicability. Finally, while most control algorithms are backed by experimental simulations, they are coded specifically for non smart grid environments and therefore fail to be reused and extended. We believe this shows a clear need for a generic and flexible tool for smart grid algorithms conception, testing and deployment.

In this paper we propose such a tool: a platform and a model that have been thought for reusability in smart grid applications and aims at supporting different smart grid applications at all levels. We demonstrate this by exhibiting the coupling of a state of the art distributed control algorithm. Previously mentioned works lack an abstract representation encompassing both the socio-economical actors and the physical devices, in GRENAD both are explicitly modeled.

The implementation of GRENAD is based on JADE. JADE is a widely popular generic purpose MAS execution platform providing primitives to ease the development of agents and deployment of agents [11]. This platform, by being generic, lacks support for smart grid elements needed to represent, simulate or even control a smart grid. GRENAD extends JADE and enrich its capabilities with additional communication mechanisms and smart grids specific objects and proposes a component based architecture.

## IV. Our Modular Agent-Based Smart-Grid Model

In this section we present and illustrate our contribution: an agent model implemented on top of JADE and a smart-grid model implemented on top of our agent model. The conception has been driven by the Design Pattern approach. For more information about Design Patterns please have a look at [6]. We first present our agent model, then we present our smart-grid model and last we describe how to formalize the scenario thanks to this model.

### A. Agent Model

GRENAD is implemented as an overlay of JADE. It provides a Domain Specific Language (DSL) that exploits a component-based approach in order to describe the agents and enhance JADE with a set of native services (agent building system, ergonomic agent messaging service, generalized support of a publish/subscribe system, ...). The agent building system simplifies the use of JADE for defining agents at compile time and runtime (initialisation step). The first class to understand in the agent building system is `GrenadAgent`. A `GrenadAgent<ID extends GrenadIdentifier>` can be viewed as a pair made of an **identifier** of type Id and a list of **components**. It is an implementation of the *builder pattern* from [6, p. 97]: the identifiers and components hold all the information required to build a JADE `Agent` exposing all the natively proposed features of JADE as well as non natively proposed ones.

A `GrenadIdentifier` is the specification of a JADE `AID` (the JADE *A*gent *ID*entifier): `GrenadIdentifier` allows to obtain the `AID` once the JADE agent is set up by JADE[1]. A component is a stateful and active object that is executed within the environment of the hosting agent. Such an environment provides access to a set of services shared by all the components of the same agent. We distinguish three groups of components:

- **Action components** are components that produce a list of JADE `Behaviour` objects that are to be added in the to-be-constructed JADE `Agent.`
- **Runtime components** allow to modify the way a `GrenadAgent` is handled by JADE. They can modify the `GrenadAgent` or execute side effects that depend on it, at agent set up or take down. They can specify actions to realise while cloning or moving the agent. They can also catch and handle exceptions issued by the agent execution, which is a functionality not natively proposed by JADE.
- **Composite components** are collections of components (as such they follow the *composite pattern* from [6, p. 163]). They allow to manipulate a bag of components as if it was a single one. From the agent point of view, loading a `Composite component` is equivalent to loading each component of the collection it represents.

By definition a component is an independent piece of software: it has its own state (set of constants, variables and methods) and has no access to the other components of the hosting agent (except for certain Runtime Components, as it is their goal). Certain actions allow to specify their state in a separate class and thus easily share it with other components

[1] JADE agent life cycle starts by building the agent, then executes the method `setup` once the agent is alive. It then executes the method `takeDown` if the agent is required to terminate, and finally kills it

of the same agent. The goal is to enhance component re-usability by separating configuration/decision aspects to action/communication ones. For example, different behaviours or protocols may require at some point to have access to the agent launch date, its neighbourhood, its preferences or abilities such as computing an optimal path. Exploiting the *Template Method Pattern* from [6, p. 325] allows to identify those methods that ought to be placed in interfaces implemented by the component state. It is then the responsibility of the agent to globally implement them in states shared by its components, thus ensuring both coherence of its actions and efficient factorisation of its code.

Let us consider, for instance, an auction about a certain type of objects that runs between an auctioneer and several bidders. In this process each bidder is characterized by the value it attributes to the objects. Except for this *decision function*, all other actions executed during this auction process are conventions specified by the auction rules. A correctly factorized implementation would thus define a component that encodes the behaviours resulting from those rules and exploits the evaluation function on the considered bidders, implemented in a separated state. Let us suppose now that one of these same bidders has latter participates in a different kind of auctions about the same type of objects. The expected behaviour for this agent is to evaluate the objects in the same way. Our "correctly factorized" approach does it naturally since the agent will only need to share the previously used state with the component executing this new process.

Please note that we strongly recommend to avoid data mutability in this approach by only sharing constants and methods. If the state has variables, it is up to the programmer to ensure, when building the agent, that the different components that exploit it do not modify them in an incoherent way.

Besides, all the components of a same agent share three services :

- Access to the unique identifier of the agent.
- Access to the agent communication service, which allows to send messages and have access to the mailbox.
- Access to the publish/subscribe service of the agent.

The *Publish/Subscribe Pattern* (also known as the *Observer Pattern*, see [6, p. 293]) is natively proposed by GRENAD, because it provides an efficient way to decouple communication, as opposed to the natural approach of message sending. In the *Publish/Subscribe Pattern*, an agent, the publisher, publishes whenever it decides any information it finds relevant. It does not know *a priori* who is interested by this information. A published information is identified by its type, a name set by the publisher, and the identifier of the publisher. A subscriber agent knows *a priori* which publication of which publisher it is interested in. It thus subscribes to it, and indicates at the same time the action it will execute upon the reception of a new publication (how it will *react*). It may stop this behaviour at any time by unsubscribing.

Simple direct message exchange and the *Publish/Subscribe Pattern* are opposed approaches to communication: in the former, the receiver of the message is decided by the sender, while in the latter the sender of a message is decided by the receiver. Natively proposing both approaches brings a great flexibility in the way of establishing communication neighbourhoods. *Publish/Subscribe Pattern* is also well suited for the event-driven approach that characterizes reactive agents.

The approach proposed so far achieves modularity in the production of multi-agent systems by applying the *Builder Pattern* on a class that encapsulates its own strategy (*i.e.* the list of components), thus allowing to get rid of the "extends" keyword for varying the characteristics of the instantiated agent. Indeed the *Strategy Pattern*, by promoting encapsulation, is more suited than inheritance when it comes to flexibility, re-usability and modularity. Modularity comes from the fact that a component can informally be seen as a full and independent functionality. It can be assembled with other components in order to easily define a complex agent. Component independence is the first root of re-usability. The second root lies in the ability to delegate the responsibility of the configuration or decision methods to the agent rather than to the components. Hence, given a set of already implemented components, it is possible to build heterogeneous agents by varying their preferences, planning methods, *etc.*. Lastly, flexibility is provided by the simultaneous native support of traditional mail-box based message sending and reactive *Publish/Subscribe Pattern*, which avoids imposing constraints upon the initial coupling of neighbours.

## B. *Smart-Grid Model*

On the basis of the agent structure presented in section IV-A, we propose here a model that allows to describe, simulate and pilot complex smart-grids. In order to do so, it exploits physical models of atomic components as well as simulators of groups of components. It allows a fine-grained remote monitoring and control of the energy distribution given information about consumption, production and a "configuration" of each component (*i.e.* an instantiation of its actuators). Also, actuators can be remotely controlled, thus allowing the smart-grid to be controled by a user interface (GUI, service web, ...) or an autonomous decision mechanism. Considered energy information is projected over time and includes generic multi-flow energy demand, generic offer characteristic and whether the offer can meet the demand at each point of the grid. Relevant physical model can simulate the propagation of blackouts or brownouts. Also several simulations can run simultaneously on the same smart-grid, made mutually dependent (*e.g.* planning and operation) and easily compared (*e.g.* real-time payment).

The model proposed here has been implemented in GRENAD as a collection of components (see Section IV-A). Hence it can naturally be integrated with other components and distributed over a network in order to be used in various situations. For instance, it can get its values and set its actuators from and to a physical smart-grid ; it can be used as an isolated simulator ; it can be used in conjunction with a coordination protocol ; encapsulating agents may interact with real clients ; encapsulating agents may negotiate on stock market, ...

In this section we present the different aspects of our smart-grid model: how energy is represented by the different roles, how information about energy is stored and updated, how it is transmitted between the agents and projected over time.

*1) Energy and Roles:* We consider three roles.

- **A consumer** expresses a *demand* as an *amount* of energy on the different flows considered (*e.g.* heat, electricity, ...).
- **A producer** expresses an *offer*, as a function that, given a demand, returns whether it can be produced. If so, it returns also what would be the *Quality of Service* (QoS) (*e.g.* $CO_2$ impact (in $kgCO_2$), base cost (in euros), ...).
- **A distributor** distributes the energy among its neighbours. It associates with each consuming neighbour an offer and each producing neighbour a demand.

A given component has a specific representation of the energy. This representation is structured among two aspects: amount and QoS. No constraints at all are put on these objects: they are only defined w.r.t. the information required by the physical models and the simulators used in the considered application. However, some components require a manipulation function to be provided. For this reason, we consider a dedicated algebra over amounts or QoS. This algebra allows the component to manipulate energy information as a black box, thus ensuring genericity. For instance, let $\mathcal{A}$ be an amount, $\oplus^{\mathcal{A}} : \mathcal{A} \times \mathcal{A} \mapsto \mathcal{A}$ is an additive operator over $\mathcal{A}$. It allows to compute the overall demand of two consumers. $\emptyset^{\mathcal{A}}$ represent a null amount, *i.e.* the neutral element of $\oplus^{\mathcal{A}}$. The same goes for QoS.

A `Demand<`$\mathcal{A}$`>` is a container of several flows of energy of type $\mathcal{A}$. We implements a flow as a class that extends the class $\mathcal{A}$. A `Demand<`$\mathcal{A}$`>` can be manipulated, given an aforementioned algebra over $\mathcal{A}$ by simply applying it independently on each of its flow.

An offer `Offer<`$\mathcal{A}, \mathcal{QoS}$`>` is a function from a `Demand<`$\mathcal{A}$`>` to an `Optional<`$\mathcal{QoS}$`>`. An `Optional<`$\mathcal{QoS}$`>` can either be a value *of* type $\mathcal{QoS}$, if the producer has the capacity to satisfy the demand, or be *empty* if it cannot.

*2) State, Fields and Internal Model:* The state of a smart-grid component is defined as a collection of a dedicated class of fields, `SGField<A>`. An `SGField<A>` allows to handle asynchronicity by outputting four types of values: (1) an object of type `A` if it has been correctly instantiated (in this case, the field is said to be *valued*), (2) `NotReady` if some information is still missing or some computation is still running on, (3) `IncoherentNeighbourhood` if an information received from some neighbour is incoherent and (4) `InvalidInternalState` if the instantiation of the state of the component is invalid w.r.t. the internal model. Note that we do not throw any exception, since the states 2, 3 and 4 can come from a delay of communication due to asynchronicity or distribution. For instance, if two neighbours are required to modify their configuration, the first that receives the request may define an `IncoherentNeighbourhood` while it has not received the updated information of the other component.

We distinguish five types of `SGField`. The first, `Fixed<A>`, has a constant value and, thus, should always be valued. The second and third fields have dynamic values and exploit the Publish/Subscribe pattern: `Output<A>` is an observable field that is observed by a `Sensor<A>` field of a neighbour. Whenever a new value is set in an `Output<A>`, it is published. The neighbour's `Sensor<A>`, upon the reception of the publication, will trigger the occurrence of a new event within the component. This will typically result to the call of the internal model of the component, which may in turn update some outputs, thus propagating the information. The fourth field, `Actuator<A>`, is an `Output<A>` that can be remotely controlled by the fifth field, `Controller<A>`, which is a `Sensor<A>` that may send a request to implement a new value on the actuator it observes, thus modifying the configuration of the smart-grid.

Each class representing a smart-grid role (consumer, producer or distributor) is associated in one hand to a set of fixed, outputs, sensors and actuators (its state) and, in another hand, to an internal model. As stated above, the internal model is triggered upon the reception of a new event, either a new value observed by some sensors, or the modification of some actuators. When called, the internal model will read the value of the fixed fields, the sensors and the actuators and update the outputs. The internal model is an abstract method whose implementation is application specific and may rely on some simple Java models or complex external physical simulators.

*3) Links and Information Propagation:* In our smart-grid model, the neighbourhood relation is defined using outputs and sensors: each component associates an output to each of its neighbours. It also defines a sensor for each output that has been associated to him.

Sensors may transform the observed information. This allows to easily implement link losses or maximal capacity. Sensors may also change the type of the information. Hence, given a function from `A` to `B`, an `Output<A>` can be observed by a `Sensor<B>`. This allows to easily connect heterogeneous components.

Note that each component is associated to a group. Only outputs and sensors of the same group can interact. This allows a same agent to load different instances of the same component, each being uniquely identified by its group. Thanks to this architecture, it is easy to do simultaneously the current *operation* phase and tomorrow *planning* phase on a same agent. The latter may easily compute its real-time payment by comparing the outputs of its *planning* and *operation* components.

The three roles (consumer, producer or distributor) rely on a strict neighbourhood architecture. Figure 2 presents the neighbourhood relation between the roles. Next to the role is represented the type of its output; each output is mirrored by a sensor of the neighbour. For clarity, we always consider in this article homogeneous components: they use the same energy representation, $(\mathcal{A}, \mathcal{QoS})$.

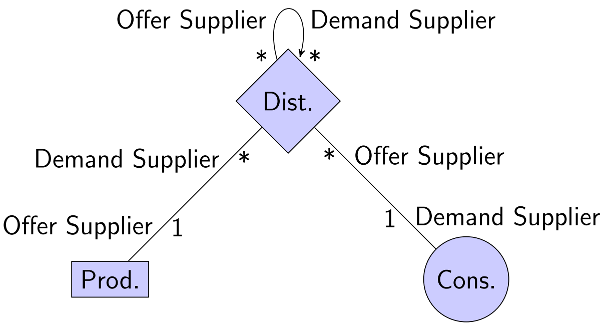


Fig. 2. Relation between roles

Consumers and producers are seen as clients of the distributing network. For this reason, they have exactly one neighbour, which is a distributor. They are responsible for defining either their demand (consumers) or their offer (producers), which is then observed by the distributor. A distributor may have several neighbours, either consumers or producers. The result of the distributed propagation of energy information is that every distributor outputs an offer to each of its neighbour consumer and a demand to each of its neighbouring producer. A distributor may also have several distributor neighbours. For each couple of distributor neighbours, one must be consuming and the other producing. The attribution of the role results from the internal models that may rely, for instance, on the value of actuators. Indeed, the distributor neighbour which consumes outputs a demand and observes an offer, and conversely for the producing one. If the neighbours disagree on their respective roles, they will output an `IncoherentState`. We refer to a distributor neighbour that has the consumer (rep. producer) role as a consumer (resp. producer) neighbour of a distributor. We refer to a distributor neighbour that is either the consumer (rep. producer) or a distributor that is consuming (resp. producing) as a consuming (resp. producing) neighbour of a distributor.

This model has been designed in order to perfectly separate the responsibility of each role. Consumer internal models are only concerned by the definition of the demand. Producer internal models are only concerned by the definition of the offer. Distributor internal models take as input the demand and the offer and are only concerned about their distribution. This abstract model is the basis of our components. Concrete classes require to implement specific internal models. Such models exploit specific actuators or outputs/sensors. Controllers allow a dynamic reconfiguration of actuators, by either a distributed or centralised algorithm, in order to optimize the energy distribution.

*4) Projection over time:* In Section IV-B2, we presented an `SGField<A>` as a container of either `A`, `NotReady`, `IncoherentNeighbourhood` or `IncoherentInternalState`. However, it is a more complex class. Indeed, it is a planning, *i.e.* a function, that, given a date, returns one of those four values or `Undefined` if no value is associated to the date. We propose several implementations of the inner function model, which is set at construction time (see *strategy pattern*). For instance, a *constant* will return the same value for all the dates, a *scatter graph* is only defined on a discrete set of date, a *step function* is defined on a range of dates .... Many other models can be thought of, however, in order to be exploitable they have to implement the following set of methods.

Let us consider `FV<A>` (FV stands for Field Value), an abstract class that can only be instantiated as one of the five types of value that can be returned by an `SGField<A>`. An `SGField<A>` delegates from its inner function model[2]:

- `SGField<B>` `map(Function2<Date,FV<A>,FV<B> >` f), a function that transforms a `SGField<A>` to a `SGField<B>`, given a function that takes a date, a `FV<A>` and output a `FV<B>`.
- `SGField<C>` `zipWith(SGField<B>` that, `Function3<Date,FV<A>,FV<B>,FV<C> >` f), a function that aggregates a `SGField<A>` and a `SGField<B>` into a `SGField<C>`, by applying a function on the couple of values associated to each date.

Both functions aim to provide required generic features, while preserving the fact that the inner model is a black box. `map` allows to modify the content of an `SGField`. For example, you can trigger a “maximum of capacity” for a planning of offers or “losses” on link transmission by mapping the appropriate function into the initial planning. `zipWith` allows to combine two `SGField` objects into one. For example, a planning of demand and a planning of offer can be combined into a planning of `Optional<`$\mathcal{QoS}$`>`, which results, at each date, from the application of the demand to the offer. If the `f` parameter of `zipWith` is a binary operator (*i.e.* `A` equals `B` equals `C`), `zipWith` may be used to fuse a collection of `SGField<A>` into one `SGField<A>`.

Note that the `f` parameters of `zipWith` and `map` take the date as argument. This allows to implement time-dependent operations which is critical for any application that depends on the weather. For example the losses on a heat pipe may be dependent on the external temperature, which can be predicted with weather information.

Note that there exist several ways to simplify the `f` parameters of `zipWith` and `map`. First, if you do not consider a time-dependent application, you can build a function $f : Date \times A \rightarrow B$ from a function $f' : A \rightarrow B$ by simply returning the application of $A$ to $f'$ whatever the date is. Second, you can consider a conventional way of handling `NotReady`, `IncoherentNeighbourhood` and `InvalidInternalState` by order of criticality. If the mapped initial value (or one of the zipped value) is an `InvalidInternalState`, the resulting value is an `InvalidInternalState`. Or if it is an `IncoherentNeighbourhood` the resulting value is an `IncoherentNeighbourhood` since the internal state is valid but not w.r.t a neighbour internal state. If it is `NotReady` the resulting value is a `NotReady` since it means that there is

[2] this approach is inspired from the category pattern[13]

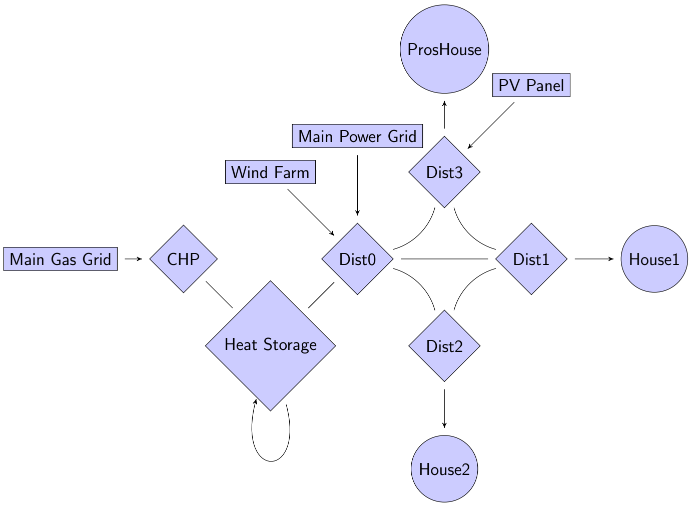


Fig. 3. Formalisation of Figure 1 scenario

*a priori* no error, but a computation is going on and we should wait for its completion. Hence, it is possible, using these rules, to automatically build a `Function<FV<A>,FV<B> >`, from a `Function<Optional<A>,Optional<B> >` where `Optional<A>` is of type `A` if a value is associated to the considered date or of type empty if no value is associated (*i.e.* type `Undefined`). The third and last way of simplifying the `f` parameter is to consider that if the mapped value or one of the zipped value is `Undefined`, the result is `Undefined`. This allows to get rid of the `Optional`.

Hence, given those three simplifying rules, to manipulate an `SGField<A>` one can rely most of the time on `SGField<B> map(Function<A,B> f)` and `SGField<C> zipWith(Function2<A,B,C> f)` for time-independent application and on `SGField<B> map(Function2<Date,A,B> f)` and `SGField<C> zipWith(Function3<Date,A,B,C> f)` for time-dependent applications. The richness of our model allows however more fine-grained control on relevant cases.

### C. Implementing the scenario

Figure 3 is the formalisation of the scenario of Figure 1 described in Section II. Figure 3 uses the same convention as Figure 2: producers are boxes, consumers circles and distributors diamonds. One can verify that Figure 3 is conform to the specification of Figure 2.

Economic actors are modelled using different components, each one characterising one functionality of the actor. Components House1 and House2 define the consumption of the elements 7 and 8 of Figure 1. They are respectively associated to Dist1 and Dist2, which define the quality of production of the energy they are supplied with. CHP and Heat Storage are distributors since they both consume and supply energy.

The links model interactions between actors; arrows are oriented from production to consumption. Both links and components are multiflow. There is a neighbourhood relation in form of a ring, the heating network, that goes along Dist0, then Dist3 then Dist1 then Dist2 to go back to Dist0. Simultaneously, there is a tree shaped network, the electricity network, that connects the root, Dist0 to Dist1, Dist2 and Dist3. For instance, the link between Dist0 and Dist1 only transmits electricity information, the one between Dist3 and Dist1 only heating information and the one between Dist0 and Dist3 convey both. We choose to connect all the main power producers into a single distributor, Dist0, however, Dist0 may hide a complex subnetwork. Indeed, it is possible to fuse several distributors into one distributor. For instance, one could have modelled Figure 1 with only one distributor instead of Dist0, Dist1, Dist2 and Dist3 or even including Heat Storage and CHP. The same operation is feasible with consumers (or producers) connected to a same distributor. Please note that, in this case, one looses direct access to some information. For example, if one fuses every distributor into a single one, GRENAD will provide an easy access to all the information from and to the producers and consumers but not necessarily the information between the distributors. This feature allows to design the application with total control of the level of details. This is particularly useful for delegating the computation of the distribution of energy of certain local parts of the system to an external simulator.

We will now detail the implementation of the components.

*1) Consumers:* Consumers (*a.k.a* House1, House2 and ProsHouse) may be implemented in different ways. Each implementation should provide an output that defines a demand at any time and is aware of the neighbour distributor offer. Also, an enhanced consumer may implement a mechanism similar to the one described in Section V in order to adapt its consumption by coordinating, during the planning phase, with its distributor neighbour. Besides, thanks to the controller of a dedicated actuator, one can provide a total control to a remote end-user during the operation phase.

*2) Common distributor:* The distributors Dist0, Dist1, Dist2 and Dist3 have as sole function to distribute the energy coming from the producers to the consumers. The stability of a grid requires electricity to respects the Kirchoff law: the flow in must equal the flow out[3]. For this reason, we propose a generic distributor class, the `KirchoffStar`. This distributor is characterized by an actuator that indicates the part of the overall demand associated to each of the producing neighbours. Hence, this actuator holds a map that associates, to each producer and distributor neighbour, $n$, a number $p_n \in [0, 100]$. $p_n$ is the percent of the sum of the consumer neighbour demands that the producing neighbour $n$ has to satisfy. The actuator values, at each time point, is valid if it respects the Kirchoff law, that is to say the sum of the $p_n$ equals 100. Note that if a distributor neighbour $n$ is consuming, $p_n$ will be null.

Reciprocally, we associate to each consumer the QoS that corresponds to the part of its consumption w.r.t. the overall consumption. For example, let us suppose that three consumers

[3] Note that as explained in Section IV-B3, losses, or any link related modification of the energy, are handled during the communication between an output and its sensor

with a demand expressed on the considered flow at the considered time point, as 3, 2 and 5 kWh. Let us suppose that there are two producers, the first assuming 30% of the consumption and the second 70%, as such, the first will assume 3 kWh and the second 7. Let us suppose that the first producer emits 4 $kgCO_2$ for producing 3 kWh and the second 6 for producing 7 kWh. Hence, the first consumer demand will be associated to a QoS of 3 $kgCO_2$, the second to a QoS of 2 $kgCO_2$ and the third to a QoS of 5 $kgCO_2$

However, our model requires the distributor to output not a QoS but an offer, that is to say, a function that associates to any demand a QoS, if this demand is satisfiable. The offer output to each consumer is computed under the hypothesis that it is the only one changing its demand. Let $prod_{dist}$ be the offer function associated of a given distributor, $dist$. A given demand $a$ is distributed by $dist$ to each of its producing neighbor, $n$, according to their associated $p_n$. This offer is computed for the demand $a$ as the sum of the QoS returned by each those neighbours for their associated part of the demand. Let $Producing_{dist}$ be the set of the considered producing distributor neighbours,

$$\forall a \in Demand < \mathcal{A} >, prod_{dist}(a) = \begin{cases} empty & \text{if some neighbour can't supply it's part} \\ \sum_{n \in Producing_{dist}} offer_n(a \times p_n) & \text{else} \end{cases} \quad (1)$$

The offer output to a consumer $cons$, $output_{cons}$, considers $prod_{dist}$ as if all other consumers had already consumed. Let $Consuming_{dist}$ be the set of the considered consuming distributor neighbours, and, for any $c \in Consuming_{dist}$, let $a_c$ be its demand,

$$\forall a \in Demand < \mathcal{A} >, output_{cons}(a) = prod_{dist}\Big(\sum_{c \in Consuming_{dist} \backslash cons} c_a \oplus a\Big) \quad (2)$$

Please note that this approach is coherent with the above mentioned examples. It also provides more information and allows to still use a simple model that only considers two types of information, demand and offer.

*3) Prosumer:* The prosumer house with PV panel needs to be separated into 3 components. One defining the offer (PV Panel), the other the demand (ProsHouse) and the third the distribution (Dist3). Please note that, in case the PV panel does not have any actuator and provides energy only to the house, it is possible to remove this component and only use a "gain" on the links, implemented thanks to the sensors (see Section IV-B3). In this case, the sensors of Dist3 and ProsHouse both transform the observed output over time by applying the weather-dependent prediction of the production of the PV Panel. If it is not the case, a separate element is required, either to be able to receive actuator modifications or to be seen by Dist 3 as a source of energy that can be distributed to the network.

*4) Storage:* The Storage unit is a distributor with two neighbours that can be decomposed as three internal components. It holds a consumer, which defines a demand, an internal state, which holds the information about the stored energy, and a producer, which defines an offer. The demand is output to one of the neighbours and the offer to the other.

The decision related to the storage is about how much energy is stored and at which time, this is why the storage demand is held by an actuator. The internal state is a private output that holds the stored energy in the form of a demand, which may be initially not null. It is updated by the storage demand, which adds energy, and the observed consuming neighbour demand, which subtracts energy. Also a special function defines losses by modifying the internal state accordingly.

The maximum capacity offer is defined by the internal state. The QoS of energy offered to the storage consuming neighbour is independent of the consuming neighbours demand. Indeed, the offered energy has already been produced at the time it is stored. Hence, the consuming neighbours are informed of this QoS even if their demands are null. For any time point $t$, let $QoS^t_{stor}$ be the QoS returned by the application of the demand of the storage at time $t$ to the offer of its producing neighbour at time $t$ (assuming this demand is sustainable). Let $stored^t$ be a demand indicating the stored energy at time $t$ (*i.e.*, the internal state). For any demand $a^t$ of the consuming neighbour, the storage can assume $a^t$ if $a^t$ is inferior to $stored^t$, and the offered QoS is always equals to $QoS^t_{stor}$.

Note that, if necessary, you can encode the fact that the storage unit can not simultaneously store and deliver by triggering an `InvalidInternalState` whenever a positive value of the demand actuator matches a positive value of the observed demand of the consuming neighbour.

*5) Main Grids:* Main gas and power grids are typically producers that are out of the scope of control of the considered smart-grids. Hence, they do not provide any actuators. If GRENAD is used as a simulator, they will be implemented with a predefined offer to output. If it is used as a piloting tool, they exploit the *interface pattern* for two purposes. Firstly, to get the information from the external system they represent and construct from it an offer, and secondly, ot transmit the demand computed by GRENAD in the appropriate format.

*6) CHP:* The CHP unit offers energy both on the heat and the electricity flows. It is a distributor since it transforms energy by consuming it from the main gas grid and supplying it to Dist0. The CHP offer to Dist0 is computed from the main gas grid offer, and an internal actuator that indicates how much of the energy consumed is used to produce electricity and how much is used to produce heat. CHP demand to the main gas grid is computed back from Dist0 demand given the internal actuator value.

*7) Wind Farm:* In our scenario, the orientation of the wind farm turbines is controlled by the application. Hence, the wind farm producer provides a collection of actuators, each one defining the orientation of the turbine it is associated with, along the time. Those actuators can either be manually controlled by a remote end-user or dynamically optimised by a coordination mechanism run by GRENAD (see for instance Section V).

In the application we are considering, the objective is typically to optimise the energy production, distribution, and consumption as a whole. The optimisation of the wind farm production is not about maximising independently their production but coordinating the production of the different producers (main grids and CHP) as well as the storage units in order to optimise the overall QoS of the energy supplied to the consumer. One may also wants to include the consumer in order to match the consumption with the production capacities. GRENAD ergonomically supports the implementation of such a coordination protocol by exploiting the *state pattern* coupled with the natively proposed *publish subscribe pattern*, as described in the next section. However, the definition of such a protocol in the complex case of this rich scenario is behind the scope of this article.

## V. Optimisation of energy distribution

In order to present the genericity and flexibility of GRENAD we describe in this section the port in GRENAD of a distributed algorithm that performs an optimal distributed dispatching of electrical power over radial networks. The presented algorithm solves a distributed constraint optimisation problem and is a part of [14], the port to GRENAD introduces some modifications. First, GRENAD generic way of handling energy allows to easily apply the presented algorithm in a more general context than [14] (see Section V-B). Second, the combination of the *State Pattern*, from [6, p. 305], with the natively implemented publish/subscribe pattern allows a simple description and implementation of the distributed algorithm. In particular, no effort have to be put on the communication, one only needs to focus on how each component reacts to the observed state of its neighbourhood.

We first expose an overview of the distributed algorithm, then we describe some hypothesis on the production, and last, we details the two phases of the algorithm.

### A. Overview of the Distributed Algorithm

The application cases of the algorithm are composed of consumers and producers with a static demand and offer. All of them can be directly represented using the proposed model, the nodes being modeled as distributors, which extend the `KirchoffStar` class. We also use the sensor/output algorithm, described in Section IV-B3 to implement maximum capacity of the links. The solution computed by the algorithm defines: for each producer, the demand it will have to deliver, for each consumer, the QoS resulting from its demand, and, for each distributor, the percent of the overall demand of its consuming neighbours it will require from each of its producing neighbours. At the end, if the offer can answer the demand, an optimal distribution will be implemented, otherwise all consumers will be informed that the network is overused. The use of the `KirchoffStar` class either ensures that the demand is satisfied or detects any demand that would violate the flow conservation constraint or the capacity.

The algorithm requires the network to be tree shaped (*e.g.* electric network), and proceeds in two phases. In the first, the *collect phase*, information is going from the leaves to the root: each node acquires and transmits to its parent the overall consumption and production information about its subtree. In the second, the *propagation phase*, information is going from the root to the leaves: each node decides an optimal power dispatching given the one of its parent.

Implementing this algorithm in GRENAD is done in the following way: each distributor possesses an actuator named *proposal* for each of its distributor neighbours, and observes with a sensor the proposals of these respective neighbours. This communication network allows the information required by the *proposal phase* to be transmitted from point to point. Each actuator is updated by the optimisation algorithm. GRENAD allows transparently this optimisation algorithm to be either located at the level agent or in a remote controller. For the sake of clarity, in this section, we refer to the initial state of sensors and outputs of the `KirchoffStar` as *regular* sensors and outputs.

The implementation exploits the *state pattern*: a given behavior is associated to a certain instantiation of the *proposal* and *regular* sensors, output and actuator. We distinguish three states:

1) *Not Ready* distributor: the distributor did not already receive enough information from its neighbours to be able to act. It is therefore waiting.
2) *Collect Ready* distributor: all but one of the distributor neighbours has instantiated their *proposal* actuator, the only neighbour that has not instantiated its proposal is then considered as its parent and the others as its children. When a distributor is in this state, it computes and instantiates its *proposal* actuator.
3) *Propagation Ready* distributor: all distributor neighbours have instantiated their *proposal* actuator. The distributor has then enough information to compute an optimal distribution between itself and its neighbours. The optimal distribution for this component is implemented by instantiating its *regular* outputs. Also, it is propagated by providing instructions to its children by the mean of its *proposal* outputs.

In the next section we detail how a proposal is computed and how the optimal configuration is determined.

### B. Production Characteristics

In order to compute an optimal distribution, certain hypothesis have to be made on producers. We consider a smart-grid where the energy demand is modeled as an amount $\mathcal{A}$ and the quality of production as $\mathcal{QoS}$. We consider the following algebra over $\mathcal{A}$ and $\mathcal{QoS}$, required by the computation of the following sections:

- $\oplus^{\mathcal{A}}$ is a binary operator over $\mathcal{A}$,
- $\oplus^{\mathcal{QoS}}$ is a binary operator over $\mathcal{QoS}$

- $\succ^{\mathcal{A}}$ is a preorder over $\mathcal{A}$.
- $\succ^{\mathcal{QoS}}$ is a preorder over $\mathcal{QoS}$.

In this section, a producer refers to a tuple $(p, maxP_p, \propto_p, offer_p)$, where $p$ is its identifier, $maxP_p$ is the maximum amount of energy it can produce, $\propto_p$ is a comparable object that allows to compare producers with respect to their quality of production and $offer_p$ is their offer function. A higher $\propto$ (given $\succ^{\propto}$) is equivalent to a better quality of production, i.e. :

$$\forall p, q \text{ producers}, \forall a \in \mathcal{A}, a \prec^{\mathcal{A}} maxP_p \wedge a \prec^{\mathcal{A}} maxP_q, \\ \propto_p \succ^{\propto} \propto_q \iff offer_p(a) \succ^{\mathcal{QoS}} offer_q(a) \quad (3)$$

In [14] the author considers a single energy flow whose amount is expressed in $kWh$, and encodes $\propto$ as the slope of a linear function that associates a $CO_2$ impact to the production of a given amount of energy.

In this setting, given a set of producers, an optimal distribution is obtained by saturating the production of producers in decreasing order of $\propto$. This is done by $dist^{opt}$, a function that considers a set of producers and an amount of energy that returns for each producer the amount of energy it must produce in the optimal distribution. $\hat{dist}^{opt}$ is a function that exploits the result of $dist^{opt}$ and returns the part of the consumption attributed to each producer in the optimal distribution, in order to instantiate the actuator of the `KirchoffStar`.

### C. Collect Phase

A proposal is computed by a *Collect Ready Distributor* and transmitted to its parent, which will have to decide the flow of energy between them. The information provided by a *Collect Ready Distributor* in a proposal to its parents is (1) the list of producers of its subtree, composed of the producer neighbours listed in its children proposals and its own producer neighbours, and (2) the overall demand of the consumers of its subtree, computed from the overall demand provided by its children proposals and the demand of its consumer neighbours. Given this proposal, the parent has to decide, during the implementation phase, whether it considers neighbouring distributors to be producing or consuming. It also has to decide of the amount of energy demanded to producers, the remaining energy of the overall consumption being assumed by the producers of the subtree. Please note that the transmitted energy between a distributor and its parent cannot exceed the maximum capacity of their power line. Hence, proposals are modified so that the overall demanded consumption does not exceed this limit. Exceeding energy is deduced from the subtree producers offer.

The proposal of a *Collect Ready Distributor*, $d$, is a tuple ($Prop_d^{Offer}$,$Prop_d^{Demand}$) where $Prop_d^{Offer}$ is a list of producers and $Prop_d^{Demand}$ a demand. They are computed as follow: Let $Cons_d$ be the set of consumer neighbours of $d$, $Prod_d$, the set of its producer neighbours, and $Child_d$, the set of its distributor neighbours that have defined their proposals. To say that $d$ is a *Collect Ready Distributor* is equivalent to say that $Child_d$ refers to all but one of its distributor neighbours, $\dot{d}$, which is its parent. The overall demand of $d$ subtree[4] is computed as the sum of the demands of its consumer neighbours and its children proposals. The list of producers of $d$ subtree is computed as the union of both its producer neighbours and the producers listed in its children proposals.

Let $Excess_d$ be the difference between the line capacity and the overall demand of $d$ subtree. Two cases are possible: (1) $Excess_d$ is negative or null, in which case the parent can answer the demand without any restriction. In this case, the proposal of $d$ is defined as the list of producers and the overall demand of its subtree; (2) $Excess_d$ is positive, in which case the answer of the parent is limited by the transmission line. The exceeding amount has then to be assumed by the subtree. In this case, $d$ proposal demand is defined as the line transmission capacity. The exceeding power is dispatched in an optimal way using $dist^{opt}$ on $d$ subtree list of producers. For each producer $p$ of $d$ subtree, let $local_p$ be the demand attributed to $p$ by the aforementioned call of $dist^{opt}$. $d$ proposal offer is then a modified list of producers of its subtree. First, for any $p$, $local_p$ is subtracted from $maxP_p$. Second, for any $p$, $offer_p$ is modified into $offer'_p$ by assuming that $local_p$ is produced, *i.e.* :

$$\forall a \in Demand < \mathcal{A} >, offer'_p(a) = \\ offer_p(a \oplus^{Amount} local_p) \quad (4)$$

### D. Propagation Phase

A distributor, $\dot{d}$, is in the *Propagation Ready* state if all of its neighbours have either defined their proposal (*i.e.* they are its children) or they outputs (*i.e.* there is exactly one, which is its parent in the network tree or zero if it is the root of the network). Such a distributor has the responsibility to determine a globally optimal distribution to each of its neighbour by instantiating its outputs. To do so it considers the overall demand and the list of producers of its subtree. Those information are computed from the children proposals, the consumer and producer neighbours and the possible parent, which can either be seen as a producing or as a consuming neighbour.

For each producer $p$ of its subtree, let $prod_p^{opt}$ be the demand it must assume in an optimal distribution. $prod_p^{opt}$ is computed by $dist^{opt}$ applied on the overall demand and the list of producers of $\dot{d}$ subtree. To each child $d$ of $\dot{d}$, is attributed the total amount of energy, $Prod_d^{opt}$, the child should supply in a globally optimal solution. $Prod_d^{opt}$ is computed as the sum of $prod_p^{opt}$ for any $p$ belonging to $Prop_d^{Offer}$. Three cases are possible: (1) $Prod_d^{opt} = Prop_d^{Demand}$, which means that in an optimal solution $d$ subtree supports exactly its demand. No energy is transmitted between $d$ and $\dot{d}$. This subnetwork can be isolated; (2) $Prod_d^{opt} > Prop_d^{Demand}$, which means that in an optimal solution $d$ subtree supports its own demand and offers some exceeding energy. In this case $\dot{d}$ is configured so that it is demanding this exceeding

[4] the subtree of the network whom root is $d$

energy from *d;* (3) $Prod_d^{opt} < Prop_d^{Demand}$, which means that in an optimal solution *d* subtree does not support its own demand. In this case, $\dot{d}$ is offering the difference to *d,* $\dot{d}$ declares itself as a producer whose maximum production is the remaining demand. Please note that in this case, even if d does not support its own demand, some of its producers may no be used in the optimal configuration. However, a distributor always consumes first all the energy offered by its parent.

## VI. Conclusion

In this work, we have presented GRENAD, a JADE-based framework, that allows to describe, simulate and pilot smart power grids. The aim of GRENAD is to provide a great flexibility in the design and implementation of smart power grids applications, so that it can be used either as a standard basis to build and test such applications or as an interface to monitor and pilot actual grids. As such, it is compliant with the main business and organizationals models: it proposes the classical roles of production, consumption, prosumption, distribution and storage. GRENAD allows time-dependant monitoring and computation; it supports both internal physical models and connection to external ones; it provides flexibility in manual, automatic or autonomous processing of the computed energy-related information; and finally it eases the integration of distributed smart grid control algorithms. To meet these objectives, GRENAD exploits a combination of Multi-Agent and component-oriented paradigms. Such an approach allows to build complex agents using simple reusable components and, thus, capitalize the development of applications.

We also have proposed a model of smart-grid that handles demand and offer and rely on the most general interpretation of consumption, production and distribution. The information about energy exchanged by the agents is complex. Indeed, it does not impose any assumption on the energy definition, which can be a combination of several flows and it is projected over time. Also, its management handles asynchronicity of computations and incoherence of states. The use of algebras and the category pattern[13] allows an efficient and ergonomic processing of this information. Also, GRENAD does not require to deal with communications, since it implements a transparent event-driven approach for information exchange, supported by the publish-subscribe pattern. Lastly, we demonstrated the ease of implementation of a generic sophisticated optimization distributed algorithm thanks to this architecture and the state pattern.

Future works include the development of more component libraries, of simulator interfaces and of dedicated optimization algorithms. Another interesting line of development would be to improve the reliability of the platform by integrating model checking capabilities.

## References

[1] Kyle Anderson, Jimmy Du, Amit Narayan, and Abbas El Gamal. Gridspice: A distributed simulation platform for the smart grid. In *Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES),* 2013 Workshop on, pages 1–5. IEEE, 2013.

[2] BOSCH. Vppm. https://www.bosch-si.com/solutions/energy/virtual-power-plant/virtual-power-plant.html.

[3] Georgios Chalkiadakis, Valentin Robu, Ramachandra Kota, Alex Rogers, and Nicholas R. Jennings. Cooperatives of distributed energy resources for efficient virtual power plants. In *The 10th International Conference on Autonomous Agents and Multiagent Systems* – Volume 2, AAMAS '11, pages 787–794, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems.

[4] US DOE. Gridlab-d. http://www.gridlabd.org/.

[5] HOMER Energy. Homer. http://www.homerenergy.com/.

[6] Ralph Johnson Erich Gamma, Richard Helm and John Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison-Wesley, 1995.

[7] ETAP. Etap. http://www.etap.com/.

[8] Bou Ghosh, Jingpeng Tang, et al. Agent-oriented designs for a self healing smart grid. In *2010 First IEEE International Conference on Smart Grid Communications,* pages 461–466, 2010.

[9] TRILLIANT Inc. Trilliant. http://www.trilliantinc.com.

[10] Electric Power Research Institute. Esvt. http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId=000000003002000312.

[11] Tellecom Italia. Jade. http://jade.tilab.com/.

[12] J. K. Kok, C. J. Warmer, and I. G. Kamphuis. Powermatcher: Multiagent control in the electricity infrastructure. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems,* AAMAS '05, pages 75–82, New York, NY, USA, 2005. ACM.

[13] Miran Lipovaca. *Learn You a Haskell for Great Good!: A Beginner's Guide.* No Starch Press, San Francisco, CA, USA, 1st edition, 2011.

[14] Sam Miller. *Decentralised coordination of smart distribution networks using message passing.* PhD thesis, University of Southampton, February 2014.

[15] NEPLAN. Neplan. http://www.neplan.ch/.

[16] Manisa Pipattanasomporn, Hassan Feroze, and S Rahman. Multi-agent systems in a distributed smart grid: Design and implementation. *In Power Systems Conference and Exposition,* 2009. PSCE'09. IEEE/PES, pages 1–8. IEEE, 2009.

[17] Flexible Power. Powermatcher. http://flexiblepower.github.io.

[18] D Pudjianto, C Ramsay, and G Strbac. *Virtual power plant and system integration of distributed energy resources.* IET RENEWABLE POWER GENERATION, 1:10–16, 2007.

[19] Sarvapali D. Ramchurn, Perukrishnen Vytelingum, Alex Rogers, and Nicholas R. Jennings. Putting the 'smarts' into the smart grid: A grand challenge for artificial intelligence. *Commun. ACM,* 55(4):86–97, April 2012.

[20] Resilient. Resilient project. http://www.resilient-project.eu/.

[21] S Schutte, Stefan Scherfke, and M Troschel. Mosaik: A framework for modular simulation of active components in smart grids. In *Smart Grid Modeling and Simulation (SGMS), 2011 IEEE First International Workshop on,* pages 55–60. IEEE, 2011.

[22] G. M. Team. *Electricity and gas supply market report.* Technical Report 176/11, The Office of Gas and Electricity Markets (Ofgem), December 2011.

[23] Chia-han Yang, Gulnara Zhabelova, Chen-Wei Yang, and Valeriy Vyatkin. Cosimulation environment for event-driven distributed controls of smart grid. *Industrial Informatics, IEEE Transactions on,* 9(3):1423–1435, 2013.

# Author Index